Mathematics for Artificial Intelligence

Christophe Giraud

Université Paris Saclay, 2023

université
PARIS-SACLAY

FACULTÉ
DES SCIENCES
D'ORSAY

Foreword

These lecture notes cover the program of the course “ Mathematics for Artificial Intelligence ” from the first year (M1) of the master program in Mathematics of the Université Paris Saclay. https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/MathIA2.html

This course aims at presenting some mathematical theory for the analysis and the understanding of machine learning algorithms. The primary focus is on theory, presented all along with central algorithms in data analysis. Some exercises and some numerical illustrations are provided at the end of each chapter, with source code available online.

The first part presents some important optimization theory in the context of sequential learning. It is an introduction to Stochastic Gradient optimization, Learning with Expert advices, and Bandits problems. These three topics are very active in machine learning, with a wide range of applications in science and in the daily life.

The second part covers some theory and some important applications around a central tool of linear algebra: the Singular Value Decomposition. Some perturbation and concentration bounds (Weyl inequalities, Davis-Kahan perturbation bound, Hanson-Wright inequalities) play an important role in the theoretical understanding of some popular tools in data analysis: Principal Component Analysis and Spectral Clustering. These algorithms and the surrounding theory are presented in the Chapters 5–8.

Any comments or corrections are welcome :)
christophe.giraud@universite-paris-saclay.fr

Enjoy your reading!

Orsay, August 2023

Christophe

Contents

Chapter 1

Sub-Gaussian random variables

1.1 Refresher on Gaussian random variables

A real random variable X follows a $\mathcal{N}(0,\sigma^2)$ Gaussian distribution, if its distribution has the probability density function (p.d.f.) with respect to the Lebesgue measure on $\mathbb{R}$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-x^2/(2\sigma^2)}.$$

In particular, if Z follows a $\mathcal{N}(0,1)$ Gaussian distribution then $X=\sigma Z$ follows a $\mathcal{N}(0,\sigma^2)$ Gaussian distribution.
The moment generating function, or Laplace transform, of a random variable X with $\mathcal{N}(0,\sigma^2)$ Gaussian distribution is given by

$$\mathbb{E}[e^{sX}] = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} e^{sx}e^{-x^2/2\sigma^2}dx = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\sigma^2s^2/2}\int_{\mathbb{R}} e^{-(x-s)^2/2\sigma^2}dx = e^{\sigma^2s^2/2}.$$

1.2 SubGaussian random variables

1.2.1 Definition and examples

Definition. A random variable X follows a SubGaussian distribution with variance proxy σ^2, denoted by $X\in SubG(\sigma^2)$ if it is centered $\mathbb{E}[X]=0$ and

$$\mathbb{E}[e^{sX}] \leq e^{\sigma^2s^2/2} \quad \text{for all} \quad s\in\mathbb{R}.$$

In addition, we write $X\in subG(\mu,\sigma^2)$ if $X-\mu\in subG(\sigma^2)$.

Remark. If $X\in SubG(\sigma^2)$ then $-X\in SubG(\sigma^2)$. So any property holding for X, also holds for $-X$.

Remark. If $X\in SubG(1)$, then $\sigma X\in SubG(\sigma^2)$.

Exercise. By investigating the behavior of $\mathbb{E}[e^{sX}]-1$ for s vanishing to 0, check that $\operatorname{var}(X)\leq\sigma^2$.

Exercise. With the convexity of $x\to e^x$, prove that if $X\in SubG(\sigma_x^2)$ and $Y\in SubG(\sigma_y^2)$ then $X+Y\in SubG\left((\sigma_x+\sigma_y)^2\right)$. This inequality can be thought as a “triangular inequality” on σ.

A classical example of SubGaussian random variables are bounded variables.

Lemma 1.1 Bounded random variable.
If X is a random variables taking values in $[a,b]$, then $X-\mathbb{E}[X]\in subG((b-a)^2/4)$

Proof of Lemma 1.1. Replacing X by $X-\mathbb{E}[X]$ and $[a,b]$ by $[a-\mathbb{E}[X],b-\mathbb{E}[X]]$, we can assume with no loss of generality that $\mathbb{E}[X]=0$. Since X is bounded, the log-Laplace tranform

$\psi(s) = \log \mathbb{E}[e^{sX}]$ exists. We can also compute differentials of ψ by switching expectation and derivation

$$\psi'(s) = \frac{\mathbb{E}\left[Xe^{sX}\right]}{\mathbb{E}\left[e^{sX}\right]} \quad \text{and} \quad \psi''(s) = \frac{\mathbb{E}\left[X^2e^{sX}\right]}{\mathbb{E}\left[e^{sX}\right]} - \left(\frac{\mathbb{E}\left[Xe^{sX}\right]}{\mathbb{E}\left[e^{sX}\right]}\right)^2.$$

Setting

$$d\mathbb{P}_s(\omega) = \frac{e^{sX(\omega)}}{\mathbb{E}\left[e^{sX}\right]}\, d\mathbb{P}(\omega),$$

we observe that

$$\psi''(s) = \mathbb{E}_s[X^2] - \mathbb{E}_s[X]^2 = \mathbb{E}_s[(X - \mathbb{E}_s[X])^2] \le \mathbb{E}_s\left[\left(X - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4}.$$

Hence, integrating twice gives

$$\psi(s) = \psi(0) + s\psi'(0) + \int_{u=0}^{s}\int_{x=0}^{u} \psi''(x)\, dx\, du \le 0 + 0 + \frac{s^2(b-a)^2}{8},$$

so, $X \in subG((b-a)^2/4)$. □

Remark: for bounded variables, the variance proxy $(b-a)^2/4$ can be much larger than the variance itself. As an illustration, let us consider the case of a Bernoulli variable X with parameter p. As X takes values 0 or 1, the random variable X is subGaussian with variance proxy $1/4$. In comparison, the variance of X is $\text{var}(X) = p(1-p)$, which can be much smaller than $1/4$ when p is close to 0 or 1.

Next lemma can be thought as a "Pythagorean (in)equality" on σ.

Lemma 1.2 Sum of subGaussian random variables.
If $X_1, \ldots, X_n \in subG(1)$ are n independent random variables, then

$$\sum_{i=1}^{n} a_i X_i \in subG(\|a\|^2).$$

Proof of Lemma 1.2. By linearity of the expectation, we have $\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = 0$. In addition, by independence

$$\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n} a_i X_i\right)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(s a_i X_i\right)\right] \le \prod_{i=1}^{n} e^{(a_i s)^2/2} = e^{\|a\|^2 s^2/2}.$$

Hence $\sum_{i=1}^n a_i X_i \in subG(\|a\|^2)$. □

1.2.2 Tails of subGaussian random variables and Hoefdding inequality

An important property of subGaussian random variables are the fast decreasing of their tails.

Lemma 1.3 Tails of a subGaussian random variable.
The tail of a random variable $X \in subG(\sigma^2)$ fulfills for any $t \ge 0$

$$\mathbb{P}[X \ge t] \le e^{-t^2/(2\sigma^2)}.$$

Proof of Lemma 1.3. The method of the proof is called the Chernoff method. This method is as important as the result.

For any $s \geq 0$, Markov inequality gives

$$\mathbb{P}[X \geq t] \leq e^{-st}\mathbb{E}[e^{sX}] \leq \exp\left(-st + \sigma^2 s^2/2\right).$$

The right-hand side is minimum pour $s = t/\sigma^2$, which gives the result. □

Remark. It is worth to notice that if $\mathbb{E}[X] = 0$ and $\mathbb{P}[|X| \geq t] \leq 2e^{-t^2/(2\sigma^2)}$, then $X \in subG(11\sigma^2)$. Indeed, we have

$$\mathbb{E}[e^{sX}] \leq \mathbb{E}[e^{|sX|}] = \int_1^\infty \mathbb{P}[e^{|sX|} \geq t]\, dt.$$

The change of variable $t = e^{|s|u}$ gives

$$\mathbb{E}[e^{sX}] \leq \int_0^\infty \mathbb{P}[|X| \geq u]|s|e^{|s|u}\, du \leq \int_0^\infty 2e^{-u^2/(2\sigma^2)}|s|e^{|s|u}\, du \leq \sigma|s|\sqrt{8\pi}e^{\sigma^2 s^2/2} \leq e^{11\sigma^2 s^2/2},$$

where the last inequality follows from $xe^{1/2} \leq e^{x^2/2}$ and hence $\sigma|s|\sqrt{8\pi} \leq \exp\left(\frac{s^2 8\pi\sigma^2}{2e}\right) \leq e^{10\sigma^2 s^2/2}$.

Corollary 1.4 Tail for sums of subGaussian random variables.
If $X_1, \ldots, X_n \in subG(\sigma^2)$ are n independent random variables, then for any $t \geq 0$

$$\mathbb{P}\left[\sum_{i=1}^n a_i X_i \geq t\right] \leq \exp\left(\frac{-t^2}{2\sigma^2\|a\|^2}\right).$$

As a consequence, if $X_1, \ldots, X_n$ are n independent random variables, with $\mathbb{E}[X_i] = \mu$ and $X_i - \mu \in subG(\sigma^2)$, then for any $L \geq 0$

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \sqrt{\frac{2L\sigma^2}{n}}\right] \leq e^{-L}. \tag{1.1}$$

Corollary 1.5 Hoeffding concentration inequality.
Let $X_1, \ldots, X_n$ be n independent random variables with $X_i \in [a_i, b_i]$ for $i = 1, \ldots, n$, then for any $t \geq 0$

$$\mathbb{P}\left[\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right] \geq t\right] \leq \exp\left(-\frac{2n^2t^2}{\|b-a\|^2}\right).$$

With a union bound, we get deviation bounds for suprema of subGaussian random variables.

Lemma 1.6 Supremum of subGaussian random variables.
Let $X_1, \ldots, X_p \in subG(\sigma^2)$ be p random variables. Then, for any $L \geq 0$, we have

$$\mathbb{P}\left[\max_{i=1,\ldots p} X_i \geq \sigma\sqrt{2(\log(p) + L)}\right] \leq e^{-L} \quad \textit{and} \quad \mathbb{E}\left[\max_{i=1,\ldots p} X_i\right] \leq \sigma\sqrt{2\log(p)}.$$

Proof of Lemma 1.6. The first inequality directly follows from Lemma 1.3 and a union bound. For the second inequality, as $\log(x)$ is concave, we apply Jensen inequality to get for any $s > 0$

$$\begin{aligned}
\mathbb{E}\left[\max_{i=1,\ldots p} X_i\right] &= \mathbb{E}\left[\frac{1}{s}\log\left(e^{s\max_{i=1,\ldots p} X_i}\right)\right] \\
&\leq \frac{1}{s}\log\left(\mathbb{E}\left[e^{s\max_{i=1,\ldots p} X_i}\right]\right) \\
&\leq \frac{1}{s}\log\left(\sum_{i=1}^p \mathbb{E}\left[e^{sX_i}\right]\right) \leq \frac{1}{s}\log\left(\sum_{i=1}^p e^{s^2\sigma^2/2}\right) = \frac{\log(p)}{s} + \frac{\sigma^2 s}{2}.
\end{aligned}$$

Lemma 1.6 follows by setting $s = \sqrt{2\log(p)/\sigma^2}$. □

1.2.3 Moments of subGaussian random variables

The bound on the Moment Generating function induces some bounds on the moments of a subGaussian random variables.

Lemma 1.7 Moments of subGaussian random variables.
Let $X \in subG(\sigma^2)$. Then, we have

$$\mathbb{E}\left[X^{2k}\right] \le 2^{k+1}k!\sigma^{2k} \qquad \textit{for } k \ge 1,$$
$$\mathbb{E}\left[e^{s(X^2-\mathbb{E}[X^2])}\right] \le 1 + 64(s\sigma^2)^2 \le e^{64(s\sigma^2)^2} \qquad \textit{for } (s\sigma^2)^2 \le 1/32.$$

Proof of Lemma 1.7. Replacing X by X/σ, we can assume with no loss of generality that $X \in subG(1)$. Let j be a positive integer. With a change of variable $t = (2u)^{j/2}$ we get

$$\begin{aligned}
\mathbb{E}\left[|X|^j\right] &= \int_0^\infty \mathbb{P}\left[|X|^j \ge t\right]\, dt \\
&= 2^{j/2-1} j \int_0^\infty \mathbb{P}\left[|X| \ge \sqrt{2u}\right] u^{j/2-1}\, du \\
&\le 2^{j/2} j \int_0^\infty e^{-u} u^{j/2-1}\, du = 2^{j/2} j\, \Gamma(j/2) = 2^{j/2+1}\Gamma(j/2+1).
\end{aligned}$$

The first bound follows by setting $j = 2k$.

For the second bound, as $x \longrightarrow e^{sx}$ and $x \longrightarrow e^{-sx}$ are convex, Jensen inequality gives

$$\begin{aligned}
\mathbb{E}\left[e^{s(X^2-\mathbb{E}[X^2])}\right] &\le \frac{1}{2}\mathbb{E}\left[e^{2sX^2} + e^{-2s\mathbb{E}[X^2]}\right] \\
&\le \frac{1}{2}\mathbb{E}\left[e^{2sX^2} + e^{-2sX^2}\right] = \mathbb{E}\left[\mathrm{ch}(2sX^2)\right] = 1 + \sum_{k\ge 1}\frac{(2s)^{2k}}{(2k)!}\mathbb{E}[X^{4k}].
\end{aligned}$$

We can now use the bound $\mathbb{E}[X^{4k}] \le 2^{2k+1}(2k)!$ on the moments of X, to get for $|s| < 1/4$

$$\mathbb{E}\left[e^{s(X^2-\mathbb{E}[X^2])}\right] \le 1 + 2\sum_{k\ge 1}(4s)^{2k} = 1 + \frac{2(4s)^2}{1-(4s)^2}.$$

To conclude, we notice that the right-hand side is smaller than $1 + 4(4s)^2$ for $s^2 \le 1/32$. □

As a corollary of Lemma 1.7, we have the following concentration inequality for square subGaussian random variables.

Corollary 1.8 Concentration for sum of squares.
Let $X_1, \ldots, X_n$ be n independent random variables in $subG(1)$. Then, for any $t \ge 0$

$$\mathbb{P}\left[\sum_{i=1}^n a_i\left(X_i^2 - \mathbb{E}[X_i^2]\right) \ge t\right] \le \exp\left(-\left(\frac{t}{16\|a\|}\right)^2 \wedge \left(\frac{t}{12|a|_\infty}\right)\right).$$

Proof of Corollary 1.8. According to Lemma 1.7, for $32|a|_\infty^2 s^2 \le 1$, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^n a_i\left(X_i^2 - \mathbb{E}[X_i^2]\right) \ge t\right] &\le e^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^n a_i\left(X_i^2 - \mathbb{E}[X_i^2]\right)\right)\right] \\
&\le e^{-st}\prod_{i=1}^n e^{64(a_i s)^2} = \exp(-st + 64s^2\|a\|^2).
\end{aligned}$$

Let us compute the minimum of $\phi(s) = -st + 64s^2\|a\|^2$ over the s fulfilling $32|a|_\infty^2 s^2 \leq 1$. We observe that the unconstrained minimum of ϕ is achieved for $s_* = t/(128\|a\|^2)$. Hence, we consider appart the two cases $32|a|_\infty^2 s_*^2 \leq 1$ and $32|a|_\infty^2 s_*^2 \geq 1$.

Case $32|a|_\infty^2 s_*^2 \leq 1$: then

$$\min_{32|a|_\infty^2 s^2 \leq 1} \phi(s) = \phi(s^*) = \frac{-t^2}{128\|a\|^2} + \frac{64t^2\|a\|^2}{128^2\|a\|^4} = \frac{-t^2}{256\|a\|^2}.$$

Case $32|a|_\infty^2 s_*^2 \geq 1$: then

$$\min_{32|a|_\infty^2 s^2 \leq 1} \phi(s) = \phi\left(\frac{1}{|a|_\infty\sqrt{32}}\right) = \frac{-t}{|a|_\infty\sqrt{32}} + \frac{64\|a\|^2}{32|a|_\infty^2} \leq \frac{-t}{|a|_\infty\sqrt{32}} + \frac{t}{2|a|_\infty\sqrt{32}} = \frac{-t}{|a|_\infty\sqrt{128}},$$

where the inequality follows from $32|a|_\infty^2 s_*^2 \geq 1$ with $s_* = t/(128\|a\|^2)$.

We then have proved that for any $t \geq 0$

$$\mathbb{P}\left[\sum_{i=1}^n a_i\left(X_i^2 - \mathbb{E}[X_i^2]\right) \geq t\right] \leq \exp\left(-\left(\frac{t}{16\|a\|}\right)^2 \wedge \left(\frac{t}{\sqrt{128}|a|_\infty}\right)\right).$$

Since $\sqrt{128} \leq 12$ the Corollary 1.8 follows. □

Example: specifying Corollary 1.8 with $a_i = \sigma^2/n$ and $t = 16\sigma^2\left(\sqrt{\frac{L}{n}} \vee \frac{L}{n}\right)$, we get for any $X_1, \ldots, X_n$ independent with $subG(\sigma^2)$ distribution and $L \geq 0$

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \left(X_i^2 - \mathbb{E}[X_i^2]\right) \geq 16\sigma^2\left(\sqrt{\frac{L}{n}} \vee \frac{L}{n}\right)\right] \leq e^{-L}. \tag{1.2}$$

1.3 Problems

1.3.1 Median of Means

A) Median of Means estimator

Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2$. The Central Limit Theorem ensures that for any $L \geq 0$

$$\mathbb{P}\left[\bar{X}_n - \mu \geq \sigma\sqrt{\frac{2L}{n}}\right] \overset{n\to\infty}{\to} \mathbb{P}\left[\mathcal{N}(0,1) \geq \sqrt{2L}\right] \leq e^{-L}. \tag{1.3}$$

This bound holds in the asymptotic where $n \to \infty$. Can we have a bound similar to (1.3) for any $n \geq 1$ and $L \geq 0$?
As for $n = 1$, the bound (1.3) enforces that $X \in subG(\mu, 11\sigma^2)$, the answer is ”no” without additional assumptions on X. The answer becomes ”yes” for subGaussian random variables, as the Bound (1.1) gives a non-asymptotic version of (1.3) for any $X \in subG(\mu, \sigma^2)$.

Let us consider the case where we only have $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) \leq \sigma^2$. We will investigate a slightly different question: Without additional assumptions on X, can we find an estimator $\widehat{\mu}$ of μ fulfilling a non-asymptotic version of the concentration bound (1.3)?

As discussed above, such a bound is not achievable in general for the empirical mean $\bar{X}_n$. But it is (under some restrictions) for a robust version of it, called “median-of-means”, as we will see in this exercise.

Let $K \leq n/2$ and assume that n can be divided by K. Then, we can split $\{1, \ldots, n\}$ into K disjoint blocs $B_1, \ldots, B_K$ of size $m = n/K$.

1. Check that
$$\mathbb{P}\left[\bar{X}_{B_j} - \mu \geq \frac{2\sigma}{\sqrt{m}}\right] \leq \frac{1}{4}.$$
2. Check that $\widehat{\mu}_K = \text{median}(\bar{X}_{B_1}, \ldots, \bar{X}_{B_K})$ fulfills
$$\mathbb{P}\left[\widehat{\mu}_K - \mu \geq \frac{2\sigma}{\sqrt{m}}\right] \leq \mathbb{P}\left[\text{Binomial}(K, 1/4) \geq K/2\right].$$
3. Conclude that
$$\mathbb{P}\left[\widehat{\mu}_K - \mu \geq 2\sigma\sqrt{\frac{K}{n}}\right] \leq e^{-K/8}.$$

This last deviation bound is similar (up to constants) to (1.1) with $L = K/8$. Notice yet the two important features:

- the estimator $\widehat{\mu}_K$ depends on the confidence level K;
- $K \leq n/2$ by construction, so we do not have subGaussian deviation bounds for values of L larger than $n/16$.

B) Illustration

In this section, we illustrate the behavior of the MOM estimator. The numerics have been performed with the R software `https://cran.r-project.org`. You can reproduce them by downloading the R-code at `https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/MathIA/MOM.R`

We fix the sample size to $n = 200$ and compare the estimation errors of the empirical mean $\bar{X}_n$ and the MOM estimator $\widehat{\mu}_k$ with $K = 10$. In order to mimic the distribution of the errors, we repeat the experiment $N = 10000$ and for each estimator, we store the N errors $err_1, \ldots, err_N$. The better the estimator, the closer to zero are the errors. In order to visualize the distribution of the errors, we plot some boxplots of the absolute values of the errors $\{|err_1|, \ldots, |err_N|\}$.

Boxplots are a popular way to sketch and visualize the spread of the distribution of a set $Z = \{Z_1, \ldots, Z_N\}$ of values. Let us denote by Q_k the k-th quartile of Z. We also define $Q_0 = \min Z$ the smallest value in Z and $Q_4 = \max Z$ the largest value in Z. In a boxplot, a box is drawn representing the first quartile Q_1, the median Q_2 and the third quartile Q_3 of Z, see Figure 1.1. In addition to the box, an interval is drawn, with left value $Q_0 \vee (Q_1 - 1.5(Q_3 - Q_1))$ and right value $Q_4 \wedge (Q_3 + 1.5(Q_3 - Q_1))$. Finally, if some values fall outside the interval, they are represented as dots. We often refer to these values as "outliers".

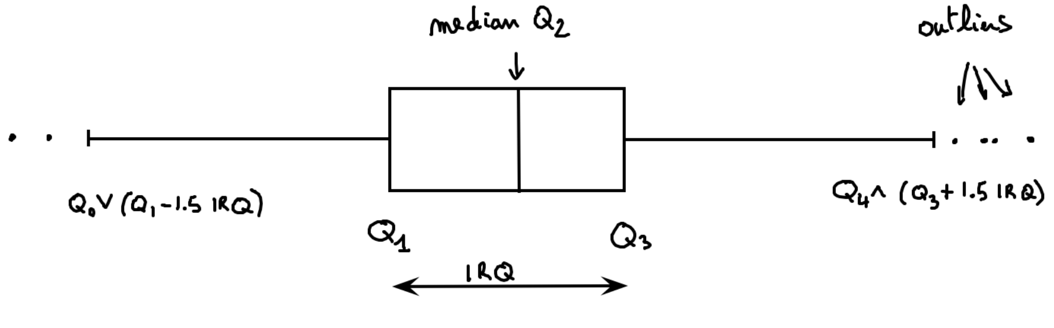

median Q_2
outliers
$Q_0 \vee (Q_1 - 1.5\ \text{IRQ})$
$Q_4 \wedge (Q_3 + 1.5\ \text{IRQ})$
Q_1
Q_3
IRQ

Figure 1.1: Description of the boxplot representation of a set of values.

We compare the behaviors of the empirical mean and the MOM estimator for three different distributions. We start with the Gaussian distribution. In this case, the empirical mean estimator behaves very well. Actually, $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, so for any value of n, the fluctuations of $\bar{X}_n - \mu$ exactly matches the asymptotic fluctuations given by the Central Limit Theorem. We cannot expect the MOM estimator to be as good as the empirical mean in this case. We observe yet in Figure 1.2 that the spread of the absolute errors of the MOM and empirical mean are similar.

The concentration bound (1.1) ensures that the empirical mean works well with subGaussian distribution. Let us consider now a distribution with heavy tails, much heavier than the tails of Lemma 1.3. As a first example, we take the Student(3) distribution. It has heavy tails, since the Student(3) distribution has an infinite moment of order 3. We observe in this case that the absolute errors of the empirical mean and the MOM estimators have similar median and first / third quartile, but the empirical mean as much more outliers, see Figure 1.2. This illustrates the facts that the empirical mean is not robust to heavy tails and that the MOM estimator is much more robust than the empirical mean.

Let us now consider a last example. We emphasize that the MOM estimator $\widehat{\mu}_K$ is kind of an interpolation between the empirical mean estimator (which is MOM with $K = 1$) and the empirical median estimator (which is MOM with $K = n$). So, MOM estimator is favored by a situation, where the median and the mean are the same, as for the Student(3) distribution. We illustrate the case where the mean and the median are different by taking a Gamma distribution with parameters $(0.01, 10)$. We observe in this case, two interesting features. First, we observe that the median absolute error of the MOM estimator is much higher than the median absolute error of the empirical mean. This reflect the fact that the MOM estimator is biased towards the median. Yet, we observe that the MOM estimator has much less ”outliers” than the empirical mean. This reflects the robustness of the MOM estimator compared to the empirical mean.

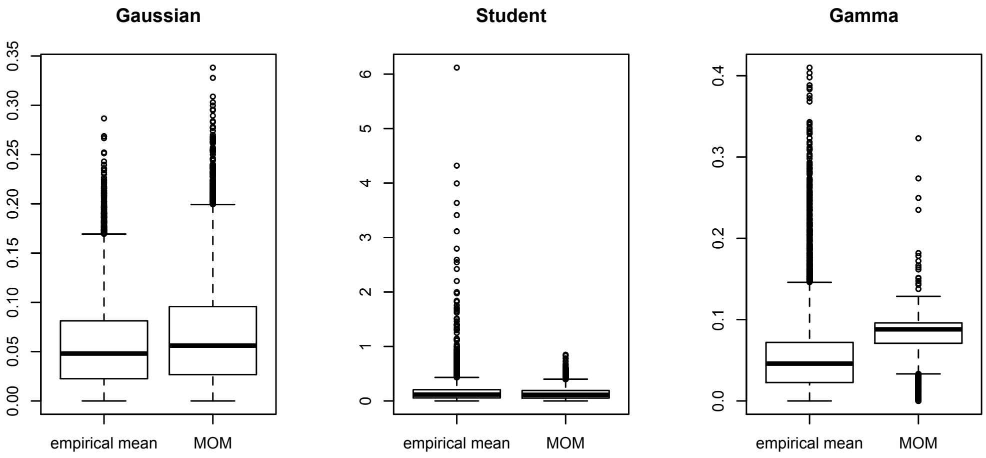


Figure 1.2: Boxplots of the absolute errors of the empirical mean and the MOM estimators. The closer the absolute values to zero, the better the estimator. Left: Gaussian distribution. Center: Student(3) distribution. Right: Gamma(0.01,10) distribution.

1.3.2 Bennett and Bernstein concentration inequalities

Hoeffding concentration inequality (Corollary 1.5) provides the following concentration bound for averages of independent bounded random variables. When the X_i are independent and fulfill $X_i \in$

$[a_i, a_i + B]$ almost surely, then

$$\mathbb{P}\left[\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right] \geq \sqrt{\frac{B^2 L}{2n}}\right] \leq e^{-L}, \quad \text{for any } L \geq 0. \tag{1.4}$$

Assume that the variables X_i are i.i.d. with variance σ^2. Then, the Central Limit Theorem ensures that

$$\lim_{n\to\infty} \mathbb{P}\left[\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right] \geq \sqrt{\frac{2\sigma^2 L}{n}}\right] = \mathbb{P}\left[\mathcal{N}(0,1) \geq \sqrt{2L}\right] \leq e^{-L}, \quad \text{for any } L \geq 0. \tag{1.5}$$

If we compare (1.4) and (1.5), we observe that the variance σ^2 in (1.5) is replaced by the upper-bound $\sigma^2 \leq B^2/4$ in (1.4). This discrepancy between the two bounds can be important as discussed below.

Let $p \in (0,1)$ and assume that $X_1, \ldots, X_n$ are i.i.d. with Bernoulli distribution $\mathcal{B}(p)$. Then, the variance is given by $\sigma^2 = p(1-p)$, while $B = 1$. When $p = 1/2$, we have $\sigma^2 = B^2/4$ so (1.4) and (1.5) are equivalent, but when p is very close to 0 or 1, the ratio $4\sigma^2/B^2 = 4p(1-p)$ is very close to 0 and there is a large discrepancy between (1.4) and (1.5). Can we improve upon Hoeffding concentration inequality in order to (partially) close this gap?

Let us discuss further the Bernoulli setting in order to figure out what we can hope to prove. First, let us explain why we cannot hope to get a non-asymptotic version of (1.5). Indeed, when $p = \lambda/n$ with $\lambda > 0$, the random variable $n\bar{X}_n$ follows a binomial distribution with parameter $(n, \lambda/n)$ and converges in distribution towards a Poisson distribution of parameter λ when n goes to infinity. We simultaneously have

$$n\sqrt{\frac{2(\lambda/n)(1-\lambda/n)L}{n}} \to \sqrt{2\lambda L}.$$

It can be proved that, for some constant $c > 0$,

$$\mathbb{P}\left[\text{Poisson}(\lambda) - \lambda \geq \sqrt{2\lambda L}\right] \geq \exp(-c\sqrt{L}\log(L)), \quad \text{when } L \to \infty,$$

so the upper bound (1.5) cannot hold non-asymptotically. Below, we will prove two bounds which behave like (1.5) when $p \gg 1/n$ and like Poisson deviation bounds when $p = O(1/n)$.

Bennett concentration bound

In this part, we will prove Bennett concentration inequality: for any independent real random variables $X_1, \ldots, X_n$ fulfilling

$$\mathbb{E}\left[X_i\right] = 0, \quad \sigma_i^2 := \mathbb{E}\left[X_i^2\right] < +\infty, \quad \text{and} \quad X_i \leq 1 \text{ a.s.} \tag{1.6}$$

we have

$$\mathbb{P}\left[\bar{X}_n \geq t\right] \leq \exp\left(-n\sigma^2 h(t/\sigma^2)\right), \text{ with } h(u) = (1+u)\log(1+u) - u \quad \text{and } \sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2. \tag{1.7}$$

1. Let $\psi(x) = e^x - x - 1 = \sum_{k\geq 2} x^k/k!$. With Taylor inequality and the power series expansion, prove the bounds

$$\begin{aligned}
&\psi(x) \leq x^2/2, && \text{for } x \leq 0,\\
&\psi(x) \geq x^2/2, && \text{for } x \geq 0,\\
&\psi(sx) \leq x^2\psi(s), && \text{for } x \in [0,1], \text{ and } s \geq 0.
\end{aligned}$$

2. We define $x_+ = x \vee 0$ and $x_- = x \wedge 0$. For $s \geq 0$, prove the bounds

$$\mathbb{E}\left[e^{sX_i}\right] \leq 1 + \psi(s)\mathbb{E}\left[(X_i)_+^2\right] + \frac{s^2}{2}\mathbb{E}\left[(X_i)_-^2\right] \leq 1 + \psi(s)\sigma_i^2 \leq e^{\psi(s)\sigma_i^2}.$$

3. Derive from the previous question, the upper bound

$$\mathbb{P}\left[X_1 + \ldots + X_n \geq t\right] \leq \exp\left(\inf_{s>0}\left\{n\sigma^2\psi(s) - st\right\}\right).$$

4. Conclude the proof of (1.7).

Bernstein concentration bound

Bernstein concentration inequality is directly derived from Bennett inequality by using the lower-bound $h(u) \geq u^2/(2 + 2u/3)$. The statement is weaker, but more handy.

Bernstein concentration inequality states that for n independent random variables fulfilling

$$X_i - \mathbb{E}\left[X_i\right] \leq B \ \text{ a.s.} \quad \text{and} \quad \sigma_i^2 := \mathrm{var}\left(X_i\right) < +\infty\,, \tag{1.8}$$

we have, for any $L \geq 0$

$$\mathbb{P}\left[\bar{X}_n - \mathbb{E}\left[\bar{X}_n\right] \geq \sqrt{\frac{2\sigma^2 L}{n}} + \frac{2BL}{3n}\right] \leq e^{-L}, \quad \text{with} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2. \tag{1.9}$$

Before proving this bound, it is worth to comment it. You can observe that it looks like a non-asymptotic version of (1.5), except that an additional term $2BL/(3n)$ has appeared. When n goes to infinity, with σ^2 and B fixed, we observe that this additional term is negligible compared to $\sqrt{2\sigma^2 L/n}$, so we recover the Central Limit Theorem bound (1.5). On the other hand, when $B = 1$ and $\sigma^2 \sim \lambda/n$ as in our Bernoulli example, the second term becomes the dominant one when L goes to infinity.

1. Check that it is enough to prove (1.9) for variables fulfilling (1.6).
2. By comparing the seconde derivative on both side, check that $(1 + u/3)h(u) \geq u^2/2$ for any $u \geq 0$.
3. Prove the first Bernstein inequality: For variables fulfilling (1.6) we have for any $t \geq 0$

$$\mathbb{P}\left[\bar{X}_n \geq t\right] \leq \exp\left(\frac{-nt^2}{2(\sigma^2 + t/3)}\right).$$

4. Check that for $x \geq 0$ and $t = \sqrt{2\sigma^2 x} + 2x/3$, we have

$$\frac{t^2}{2(\sigma^2 + t/3)} \geq x.$$

5. Conclude the proof of (1.9).

Part I

Sequential learning

Chapter 2

Statistical learning and optimisation

2.1 Batch statistical learning problem

Supervised learning problem. A central problem in machine learning is to predict an outcome $Y \in \mathcal{Y}$ from some covariates or "features" $X \in \mathcal{X}$. The prediction is done with a predictor $h : \mathcal{X} \to \mathcal{Y}$ built by the data scientist.

Loss and risk. In statistical learning, we assume that the couple (X, Y) is a random variable with distribution $\mathbb{P}^{(X,Y)}$. For a specified measurable function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, usually called "loss function", and a measurable predictor $h : \mathcal{X} \to \mathcal{Y}$, we define the so-called risk of the predictor h by

$$r(h) = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(X,Y)}}\left[\ell(Y, h(X))\right].$$

The best predictor in terms of the risk r, is then the predictor $h^* : \mathcal{X} \to \mathcal{Y}$ minimizing $r(h)$ over all the possible measurable maps $h : \mathcal{X} \to \mathcal{Y}$.

Examples of loss functions:

- L^p loss: $\ell(y, y') = (y - y')^p$, for $p \geq 1$,
- Hard loss: $\ell(y, y') = \mathbf{1}_{y \neq y'}$,
- Logistic loss: $\ell(y, y') = \log(1 + e^{-yy'})$,
- Hinge loss: $\ell(y, y') = [1 - yy']_+$, where $[x]_+ = \max(x, 0)$.

For example, when $\ell(y, y') = \mathbf{1}_{y \neq y'}$, then the risk $r(h) = \mathbb{P}(Y \neq h(X))$ is the probability of misprediction.

Learning from data. In practice, the distribution $\mathbb{P}^{(X,Y)}$ is unknown, so we can neither compute $r(h)$ nor h^*. Instead, we have access to a sample $Z = (X_i, Y_i)_{i=1,\dots,n} \in (\mathcal{X} \times \mathcal{Y})^n$ gathering n observations i.i.d. with distribution $\mathbb{P}^{(X,Y)}$. We then build a predictor based on these data. We choose a mapping $H : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{Y}$ and predict Y with $H(Z, X)$, resulting with a risk $\mathbb{E}_{(X,Y)\sim\mathbb{P}^{(X,Y)}}\left[\ell(Y, H(Z, X))\right]$. The central question is then, how to choose and compute the mapping H?

Classically, the statistical learner specifies a family of predictors $h : \Theta \times \mathcal{X} \to \mathcal{Y}$, for example $h(\theta, x) = \langle \theta, x \rangle$ for data in $\mathbb{R}^d$, and she/he considers predictors of the form $H(Z, X) = h(g(Z), X)$, with $g : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ measurable.

Examples of family of predictors:

- linear predictors $h(\theta, x) = \langle \theta, x \rangle$,
- logit predictors $h(\theta, x) = \exp(\langle \theta, x \rangle)/(1 + \exp(\langle \theta, x \rangle))$,
- kernel predictors $h(\theta, x) = \sum_{i=1}^{d} \theta_i k(z_i, x)$,
- neural networks, etc.

Goal in statistical learning. Setting

$$f(\theta) := r(h(\theta, \cdot)) = \mathbb{E}_{(X,Y)\sim\mathbb{P}^{(X,Y)}}\left[\ell(Y, h(\theta, X)\right], \quad \text{for } \theta \in \Theta, \tag{2.1}$$

the goal is to design some $\sigma(Z)$-measurable variable $\widehat{\theta} = g(Z)$, such that $f(\widehat{\theta})$ is as small as possible in expectation or with high probability with respect to the randomness of Z.

Typical results in statistical learning theory provide

- some upper bounds on the so-called "excess risk"

$$f(\widehat{\theta}) - \min_{\theta \in \Theta} f(\theta) = r(h(\theta, \cdot)) - \min_{\theta \in \Theta} r(h(\theta, \cdot)),$$

  either in expectation, or with high probability with respect to the randomness of Z.
- some lower-bounds on the best possible excess risk over a class of problems.

Remark: In practice, the data set that we observe corresponds to a realization $Z(\omega)$ of the random variable Z. The statistical learner then computes the parameter $\widehat{\theta}(\omega) = g(Z(\omega))$ and he/she uses the function $h(\widehat{\theta}(\omega), \cdot) : \mathcal{X} \to \mathcal{Y}$ for prediction. Yet, when we investigate the statistical properties of the predictor, we consider Z, and hence $\widehat{\theta}$, as a random variable. The goal is to understand the distribution of the risk $f(\widehat{\theta})$ or the distribution of the parameter $\widehat{\theta}$ with respect to the randomness of Z.

2.2 Learning with gradient descent

Let us assume that $\mathbb{P}^{(X,Y)}$ is unknown, but we observe some i.i.d. data $Z = (X_i, Y_i)_{i=1,\ldots,n}$ distributed according to $\mathbb{P}^{(X,Y)}$. In order to find a $\widehat{\theta}$ such that $f(\widehat{\theta})$ is as small as possible, a classical strategy is

1. to replace $f(\theta)$ by some empirical version $\widehat{f_Z}(\theta)$ of it

$$\widehat{f_Z}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, h(\theta, X_i)) + \Omega(\theta),$$

   where $\Omega(\theta)$ is a regularization convex penalty, for example $\Omega(\theta) = \lambda \|\theta\|^2$ with $\lambda \geq 0$;
2. to estimate θ with the so-called Penalized Empirical Risk Minimizer (PERM):

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \Theta} \widehat{f_Z}(\theta). \tag{2.2}$$

To implement this method, how can we perform the minimization above?

a) Gradient Flow

Let us assume for simplicity that $\Theta = \mathbb{R}^d$. We want to solve an optimisation problem of the form

$$\theta^* \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} f(\theta), \tag{2.3}$$

for some function $f : \mathbb{R}^d \to \mathbb{R}$, for example $f(\theta) = \widehat{f_Z}(\theta)$ for the PERM. In most cases, there exists no explicit formula for θ^*, and we must evaluate it numerically. When f is differentiable, a strategy is to start from some $\theta_0^{\mathrm{GF}} \in \mathbb{R}^d$, and then follow the steepest descent, called Gradient Flow (GF),

$$\frac{d\theta_t^{\mathrm{GF}}}{dt} = -\nabla f(\theta_t^{\mathrm{GF}}). \tag{2.4}$$

Analyzing GF: The GF can be analyzed very simply. Assume that f is convex and differentiable. Then, according to Lemma B.1, we have

$$\begin{aligned} f(\theta_t^{\mathrm{GF}}) - f(\theta^*) &\overset{\text{Lemma } B.1}{\leq} \langle \nabla f(\theta_t^{\mathrm{GF}}), \theta_t^{\mathrm{GF}} - \theta^* \rangle \\ &= -\langle \frac{d\theta_t^{\mathrm{GF}}}{dt}, \theta_t^{\mathrm{GF}} - \theta^* \rangle = -\frac{1}{2} \frac{d}{dt} \|\theta_t^{\mathrm{GF}} - \theta^*\|^2. \end{aligned} \tag{2.5}$$

Let us consider now the average of the Gradient Flow trajectory over $[0, T]$ for some $T > 0$

$$\bar{\theta}_T^{\mathrm{GF}} := \frac{1}{T}\int_0^T \theta_t \, dt.$$

According to the convexity of f, we get from Jensen Inequality and (2.5)

$$\begin{aligned} f(\bar{\theta}_T^{\mathrm{GF}}) - f(\theta^*) &\leq \frac{1}{T}\int_0^T (f(\theta_t^{\mathrm{GF}}) - f(\theta^*))\, dt \\ &\overset{(2.5)}{\leq} -\frac{1}{2T}\int_0^T \frac{d}{dt}\|\theta_t^{\mathrm{GF}} - \theta^*\|^2 \, dt \\ &= \frac{1}{2T}\left(\|\theta_0^{\mathrm{GF}} - \theta^*\|^2 - \|\theta_T^{\mathrm{GF}} - \theta^*\|^2\right) \leq \frac{\|\theta_0^{\mathrm{GF}} - \theta^*\|^2}{2T}. \end{aligned}$$

This inequality ensures that when taking the average $\bar{\theta}_T^{\mathrm{GF}}$ of the Gradient Flow, we minimise f up to an $O(1/T)$-optimisation error.

Remark: the mapping $t \longrightarrow f(\theta_t^{\mathrm{GF}})$ is non-increasing since

$$\frac{d}{dt} f(\theta_t^{\mathrm{GF}}) = \langle \nabla f(\theta_t^{\mathrm{GF}}), \frac{d\theta_t^{\mathrm{GF}}}{dt} = -\|\nabla f(\theta_t^{\mathrm{GF}})\|^2 \leq 0,$$

so $f(\theta_T^{\mathrm{GF}}) \leq f(\bar{\theta}_T^{\mathrm{GF}})$ and therefore θ_T^{GF} also minimises f up to an $O(1/T)$-optimisation error.

b) Gradient Descent

Computing continuous time dynamics as the Gradient Flow (2.4) is computationally very intensive in general. Gradient Decent (GD) is a simple discretization of the Gradient Flow, which is easy to implement, and which is widely used to seek for solutions to (2.3).

Gradient Descent (GD)

```
Input: $\eta > 0$, and $\theta_1 \in \mathbb{R}^d$.
Iterate: for $t = 1, \dots, T-1$,
    $\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$
```

Next theorem shows that the average $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^T \theta_t$ of the sequence of the GD minimises f up to an $O(1/\sqrt{T})$-optimisation error, provided that the gradient step η is chosen proportional to $T^{-1/2}$.

Theorem 2.1 GD: Rate for Lipschitz convex function.
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable and convex function. Define the sequence $(\theta_t)_{t\geq 1}$ by induction,

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t), \quad t \geq 1, \quad \textit{with } \eta = \frac{R}{L\sqrt{T}},$$

where $R \geq \|\theta_1 - \theta^*\|$ and where $L \geq \|\nabla f(\theta_t)\|$ for all $t = 1, \dots, T$.
Then, the average $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^T \theta_t$ of the sequence of the GD fulfills

$$f(\bar{\theta}_T) - \min_{\theta \in \mathbb{R}^d} f(\theta) \leq \frac{LR}{\sqrt{T}}.$$

Remarks.

- Compared to Gradient Flow, the optimisation error for Gradient Descent is $O(1/\sqrt{T})$. This slower rates is due to the discretization of the Gradient Flow dynamic.

- The choice $\eta = R/(L\sqrt{T})$ can be infeasible in practice, as the constant R and L are usually unknown. If we set $\eta = \alpha/\sqrt{T}$ for some $\alpha > 0$, then according to (2.9), the upper bound becomes

$$f(\bar{\theta}_T) - \min_{\theta \in \mathbb{R}^d} f(\theta) \leq \frac{\alpha^{-1}R^2 + \alpha L^2}{2\sqrt{T}}.$$

  Choosing $\eta = \alpha/\sqrt{T}$ has to drawback: first it requires to know in advance the time horizon T, second an arbitrary choice of α, say $\alpha = 1$, may be suboptimal if L and R have very different sizes. These two issues can be simply overcome by some adaptive choice of the step size η. This topic is yet beyond the scope of these lecture notes.

Proof of Theorem 2.1. The proof of Theorem 2.1 is a reminiscence of the analysis of the Gradient Flow page 14. As for the Gradient Flow, we get from the convexity of f and Lemma B.1

$$f(\theta_t) - f(\theta^*) \leq \langle \nabla f(\theta_t), \theta_t - \theta^* \rangle. \tag{2.6}$$

From the polarisation formula $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|b-a\|^2$, we get

$$\begin{aligned} 2\langle \eta \nabla f(\theta_t), \theta_t - \theta^* \rangle &= \|\theta_t - \theta^*\|^2 + \eta^2 \|\nabla f(\theta_t)\|^2 - \|\theta_t - \theta^* - \eta \nabla f(\theta_t)\|^2 \\ &= \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2 + \eta^2 \|\nabla f(\theta_t)\|^2. \end{aligned} \tag{2.7}$$

Summing (2.7) over t generates a telescopic sum, leading to

$$\begin{aligned} \sum_{t=1}^{T} (f(\theta_t) - f(\theta^*)) &\overset{(2.6)}{\leq} \sum_{t=1}^{T} \langle \nabla f(\theta_t), \theta_t - \theta^* \rangle \\ &\overset{(2.7)}{=} \frac{\|\theta_1 - \theta^*\|^2 - \|\theta_{T+1} - \theta^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f(\theta_t)\|^2 && (2.8) \\ &\leq \frac{R^2}{2\eta} + \frac{\eta}{2} L^2 T. && (2.9) \end{aligned}$$

The convexity of f finally ensures that

$$f(\bar{\theta}_T) \leq \frac{1}{T} \sum_{t=1}^{T} f(\theta_t),$$

so the result of Theorem 2.1 follows by plugging the value $\eta = R/(L\sqrt{T})$ in (2.9). □

c) Stochastic Gradient Descent

In practice, Stochastic Gradient Descent (SGD) is widely used for minimizing the (penalized) empirical risk $\widehat{f}_Z$ defined in (2.2). It was already implemented in the early 60's to train a linear regression on the first generations of computers.

The recipe of SGD is to replace at each step the gradient of $\widehat{f}_Z$ by a gradient based on a random subsample of the data points $(X_i, Y_i)_{i \in I}$. More precisely, for some $B \in \mathbb{N}^*$, let $I_1, I_2, \ldots \subset \{1, \ldots, n\}$ be a random sequence of subsets of $\{1, \ldots, n\}$ of size B, with uniform distribution, and independent of the data Z. We define $F_i(\theta) = \ell(Y_i, h(\theta, X_i)) + \Omega(\theta)$ and

$$\bar{F}_{I_t} = \frac{1}{B} \sum_{i \in I_t} F_i.$$

Batch-SGD amount to iterate

$$\theta_{t+1} = \theta_t - \eta \nabla \bar{F}_{I_t}(\theta_t), \quad \text{for some } \eta > 0.$$

An important advantage of Batch-SGD compared to vanilla gradient descent

$$\theta_{t+1} = \theta_t - \eta \nabla \widehat{f}_Z(\theta_t), \quad t \geq 1,$$

is that at each time step, we only need to compute $\nabla \bar{F}_{I_t}(\theta_t)$ instead of $\nabla \widehat{f}_Z(\theta_t) = \frac{1}{n}\sum_{i=1}^n \nabla F_i(\theta_t)$, reducing the computational complexity by a factor n/B. When the datasets are large, with millions of samples, the speed-up is substantial.

The rationale for using $\nabla \bar{F}_{I_t}(\theta_t)$ instead of $\nabla \widehat{f}_Z(\theta_t)$ is that

$$\nabla \widehat{f}_Z(\theta_t) = \mathbb{E}\left[\nabla \bar{F}_{I_t}(\theta_t)|\mathcal{F}_{t-1}\right], \quad \text{where} \quad \mathcal{F}_{t-1} = \sigma(Z, I_1, \ldots, I_{t-1}), \tag{2.10}$$

so the stochastic gradient $\nabla \bar{F}_{I_t}(\theta_t)$ can be viewed as a "noisy" version of $\nabla \widehat{f}_Z(\theta_t)$. Other choices of stochastic gradient g_t can be done, and the next theorem provides a convergence analysis for minimizing (2.3) with any stochastic gradient g_t fulfilling $\mathbb{E}\left[g_t|\theta_1, \ldots, \theta_t\right] = \nabla f(\theta_t)$.

Theorem 2.2 SGD: Rate for Lipschitz convex function.
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable and convex function and $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Define the sequence $(\theta_t)_{t\geq 1}$ by induction,

$$\theta_{t+1} = \theta_t - \eta g_t, \quad t \geq 1, \quad \text{with } \eta = \frac{R}{L\sqrt{T}},$$

where $R \geq \|\theta_1 - \theta^*\|$ and where g_t is any $\mathcal{F}_t$-measurable random variable fulfilling

$$\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f(\theta_t), \quad \text{and} \quad \mathbb{E}[\|g_t\|^2] \leq L^2, \quad \text{for } t = 1, \ldots, T.$$

Then, the average $\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^T \theta_t$ of the sequence of the SGD fulfills

$$\mathbb{E}\left[f(\bar{\theta}_T) - \min_{\theta\in\mathbb{R}^d} f(\theta)\right] \leq \frac{LR}{\sqrt{T}}.$$

Remark. The bound is similar to the one in Theorem 2.1, except that it holds only on average over the randomness of the stochastic gradients.

Proof of Theorem 2.2. The proof follows the same lines as the proof of Theorem 2.1. We start again from (2.6) and since $\mathbb{E}\left[g_t|\mathcal{F}_{t-1}\right] = \nabla f(\theta_t)$, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T (f(\theta_t) - f(\theta^*))\right] &\leq \mathbb{E}\left[\sum_{t=1}^T \langle \nabla f(\theta_t), \theta_t - \theta^*\rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}\left[\langle g_t, \theta_t - \theta^*\rangle|\mathcal{F}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \langle g_t, \theta_t - \theta^*\rangle\right].\end{aligned} \tag{2.11}$$

Following exactly the same computations as for proving (2.8), with $\nabla f(\theta_t)$ replaced by g_t, we get

$$\sum_{t=1}^T \langle g_t, \theta_t - \theta^*\rangle = \frac{\|\theta_1 - \theta^*\|^2 - \|\theta_{T+1} - \theta^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|g_t\|^2. \tag{2.12}$$

Plugging this bound into (2.11) leads to

$$\mathbb{E}\left[\sum_{t=1}^{T}(f(\theta_t)-f(\theta^*))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\langle g_t, \theta_t-\theta^*\rangle\right]$$
$$\leq \frac{R^2}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|g_t\|^2\right] \leq \frac{R^2}{2\eta}+\frac{\eta}{2}L^2T.$$

We then conclude as in the proof of Theorem 2.1. □

d) One-pass Stochastic Gradient Descent

We observe that Batch-SGD presented above aims at minimizing the penalized empirical risk minimizer $\widehat{f}_Z(\theta)$. Our ultimate goal is yet to get the risk $f(\theta) = \mathbb{E}_{(X,Y)}\left[\ell(Y, h(\theta, X)\right]$ as small as possible.
Minimizing f by GD would require to compute the gradient

$$\nabla f(\theta_t) = \nabla_\theta\, \mathbb{E}_{(X,Y)}\left[\ell(Y, h(\theta_t, X)\right].$$

This quantity is unknown, but at each step t, we can compute

$$g_t = \nabla_\theta \ell(Y_t, h(\theta_t, X_t)).$$

Under the assumption that $(X_t, Y_t)_{t=1,\dots,n}$ are i.i.d., and some dominance assumptions ensuring that we can invert gradient and expectation

$$\mathbb{E}\left[\nabla_\theta \ell(Y_t, h(\theta, X_t))\right] = \nabla_\theta \mathbb{E}\left[\ell(Y_t, h(\theta, X_t))\right], \quad \text{for all } \theta \in \mathbb{R}^d, \tag{2.13}$$

the gradient g_t fulfills

$$\mathbb{E}\left[g_t|\mathcal{F}_{t-1}\right] = \nabla f(\theta_t), \quad \text{where} \quad \mathcal{F}_{t-1} = \sigma\left((X_i, Y_i) : i = 1, \dots, t-1\right).$$

In light of Theorem 2.2 this suggests to apply a one-pass SGD on the data.

One-Pass Stochastic Gradient Descent (One-pass SGD)

```
Input: $\eta > 0$, and $\theta_1 \in \mathbb{R}^d$.
Iterate: $\theta_{t+1} = \theta_t - \eta\nabla_\theta \ell(Y_t, h(\theta_t, X_t))$,    for $t = 1, \dots, n-1$
Output: $\widehat{\theta} = \frac{1}{n}\sum_{t=1}^{n}\theta_t$
```

Under the assumptions of Theorem 2.2 and assumptions ensuring that we can invert gradient and expectation as in (2.13), we get from Theorem 2.2 that $\mathbb{E}_Z\left[f(\widehat{\theta}) - f(\theta^*)\right] \leq LR/\sqrt{n}$, i.e.

$$\mathbb{E}_Z\left[\mathbb{E}_{(X,Y)}\left[\ell(Y, h(\widehat{\theta}, X)\right]\right] - \min_{\theta\in\mathbb{R}^d}\mathbb{E}_{(X,Y)}\left[\ell(Y, h(\theta, X)\right] \leq \frac{LR}{\sqrt{n}}.$$

Remark 1. At first sight, the application of Batch-SGD with $B = 1$ for empirical risk minimization looks similar to the application of One-pass SGD. We emphasize yet two important differences:

- For empirical risk minimisation, a data point $Z_i = (X_i, Y_i)$ can be involved in multiple steps in Batch-SGD. The algorithm does "multiple passes" on the data. A contrario, in one-pass SGD, each data point Z_i is used only once. The algorithm does a "single pass" on the data.

- The goal in empirical risk minimization with Batch-SGD and one-pass SGD are different. In the first case, we seek to minimize the (penalized) empirical risk

$$\widehat{f}_Z(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, h(\theta, X_i)) + \Omega(\theta),$$

  while in the second case, we seek to get the prediction risk

$$f(\widehat{\theta}) = \mathbb{E}_{(X,Y)}\left[\ell(Y, h(\widehat{\theta}, X))\right],$$

  as small as possible, which is our ultimate goal.

It is important to notice that while we can compute the function $\widehat{f}_Z(\theta)$, the function $f(\theta)$ is unknown to the data scientist, since $P^{(X,Y)}$ is unknown.

Remark 2. Another virtue of one-pass SGD is that we do not need to have all the data $Z = (X_i, Y_i)_{i=1,\dots,n}$ from the very beginning. One-pass SGD can handle data that arrives sequentially in time, and it outputs a stream of predictor as time passes. This setting, where data is observed sequentially in time, is precisely the topic of theses lecture notes.

2.3 Sequential statistical learning

In batch statistical learning, we have access from the start to a whole data set $Z = (X_i, Y_i)_{i=1,\dots,n}$. While this situation arises when analyzing data produced by some experiments, it does not fit many practical situations where:

- the data $Z_t = (X_t, Y_t)$ is collected sequentially as time passes;
- a decision or prediction must be performed at each time step (based on the data available at this time);
- possibly, the learner can choose at each time step the covariates X_t, with a choice based on the past observations.

Our focus on the first part of the lecture notes, will be on such problems. We will consider a set of problems that can be summarized, at a high-level, as follows. At each time $t = 1, 2, \dots$, we will have to choose an "action" $\widehat{\theta}_t$ from available data at time t. Each action $\widehat{\theta}_t$ has a risk or "cost" $f_t(\widehat{\theta}_t)$ and our goal will be to find strategies $\widehat{\theta}$ such that the cumulated cost

$$\sum_{t=1}^{T} f_t(\widehat{\theta}_t),$$

is as small as possible.

Let us sketch succinctly the three problems that we will consider. These problems will be described in full details in the next sections and chapters.

1) Online learning.
In this case, at each time t, we seek to predict a random outcome $Y_t \in \mathbb{R}$ from random covariates $X_t \in \mathcal{X}$. The prediction is assumed to be of the form $h(\widehat{\theta}_t, X_t)$, where $h(\theta, x)$ is a prescribed family of predictors. For a given loss function $\ell : \mathbb{R}^2 \to \mathbb{R}_+$, the "cost" associated to the choice of a parameter θ is the integrated loss

$$f_t(\theta) = \mathbb{E}_{(X_t,Y_t)}\left[\ell(Y_t, h(\theta, X_t))\right].$$

2) Sequential prediction with expert advices.
A variant of the previous learning problem, is when at each time t, we have access to some predictions $h_1(t), \dots, h_d(t)$ of Y_t from a set of d experts. The question is then how to aggregate this

predictions in order to get a loss as small as possible? For a given loss function ℓ, we will consider convex combinations of the predictions $\sum_{k=1}^{d} \widehat{\theta}_{k,t} h_k(t)$ and we will seek to minimise the cumulated loss

$$\sum_{t=1}^{T} \ell \left(Y_t, \sum_{k=1}^{d} \widehat{\theta}_{k,t} h_k(t) \right).$$

3) Multi-armed bandit problems.

In the Multi-armed bandit (MAB) problems, the player chooses at each time t an action $\widehat{\theta}_t \in \Theta$ and receives a pay-off Z_t with conditional mean $\mathbb{E}\left[Z_t|\theta\right] = \mu(\theta)$ for any $\theta \in \Theta$. The function $\mu : \Theta \longrightarrow \mathbb{R}$ is unknown and the only information available at time t are the past outcomes $Z_1, \dots, Z_{t-1}$. We will seek for strategies $\widehat{\theta}$ maximizing the predictable pay-off

$$\sum_{t=1}^{T} \mathbb{E}\left[Z_t|\widehat{\theta}_t\right] = \sum_{t=1}^{T} \mu(\widehat{\theta}_t),$$

or equivalently minimizing the regret

$$\sum_{t=1}^{T} \left(\max_{\theta\in\Theta} \mu(\theta) - \mu(\widehat{\theta}_t) \right).$$

2.4 Online learning with stochastic gradient descent

2.4.1 Stochastic sequential prediction

Let us consider the sequential prediction problem, where at time t we want to predict a real valued outcome Y_t from covariates X_t via a predictor $h(\widehat{\theta}_t, X_t)$. More precisely, we consider $\Theta \subset \mathbb{R}^d$, a set $\mathcal{X}$ and a prescribed regression function $h : \Theta \times \mathcal{X} \to \mathbb{R}$, differentiable in the first variable θ.

For a parameter θ and a differentiable loss function $\ell : \mathbb{R}^2 \to \mathbb{R}_+$, we consider the integrated loss $f_t(\theta) = \mathbb{E}[\ell(Y_t, h(\theta, X_t)]$, that we also assume to be differentiable. We have in mind a situation where the functions f_t do not vary too much accros time t, and hence we will compare ourself to the best "constant" strategy

$$\theta^* \in \operatorname*{argmin}_{\theta\in\Theta} \sum_{t=1}^{T} f_t(\theta), \tag{2.14}$$

which we assume to exist.

We cannot use the parameter θ^*, as we do not have access to the functions f_t. Actually, at each time t, we have to choose the parameter $\widehat{\theta}_t$ according to the only information $\mathcal{F}_{t-1} = \sigma((X_1, Y_1), \dots, (X_{t-1}, Y_{t-1}))$ available after the step $t-1$.

2.4.2 Sequential Stochastic gradient descent

a) Adapting One-pass SGD.

Let us assume for simplicity that $\Theta = \mathbb{R}^d$. We wish to adapt (S)GD to our setting. The main difficulty is that $\widehat{\theta}_t$ can only be computed based on information available at time t. In particular, at time t, we do not have access to $\sum_{t=1}^{T} f_t(\theta)$, nor to an empirical version of it. We may yet adapt the one-pass SGD idea to our setting, by computing sequentially

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \eta g_t, \quad \text{with} \quad g_t = \nabla_\theta \ell(Y_t, h(\widehat{\theta}_t, X_t)). \tag{2.15}$$

Under the assumption that (X_t, Y_t) is independent of $\mathcal{F}_{t-1}$ and some dominance assumptions ensuring that we can invert gradient and expectation

$$\mathbb{E}\left[\nabla_\theta \ell(Y_t, h(\theta, X_t))\right] = \nabla_\theta \mathbb{E}\left[\ell(Y_t, h(\theta, X_t))\right], \quad \text{for all } \theta \in \mathbb{R}^d,$$

the gradient g_t fulfills

$$\mathbb{E}\left[g_t|\mathcal{F}_{t-1}\right] = \nabla f_t(\widehat{\theta}_t).$$

Hence, as in one-pass SGD, we may consider using the algorithm (2.15).

b) Sequential Stochastic Gradient Descent.

Motivated by the above example, we consider below the generic problem where, for a sequence of differentiable convex functions $\{f_t : t = 1, \dots, T\}$, and a filtration $(\mathcal{F}_t)_{t\geq 0}$, we seek to minimize the regret

$$\sum_{t=1}^{T}(f_t(\widehat{\theta}_t) - f_t(\theta^*)),$$

under the constraint that $\widehat{\theta}_{t+1}$ only has access to the past value $\widehat{\theta}_t$ which is $\mathcal{F}_{t-1}$-measurable and to some $\mathcal{F}_t$-measurable random vector g_t fulfilling

$$\mathbb{E}\left[g_t|\mathcal{F}_{t-1}\right] = \nabla f_t(\widehat{\theta}_t).$$

Any random vector g_t fulfilling the above property is called a stochastic gradient of f_t at $\widehat{\theta}_t$.

```
Sequential Stochastic Gradient Descent (Seq-SGD)
Input: $\eta > 0$, and $\theta_1$.
Iterate: for $t = 1, \dots, T-1$,
    $\widehat{\theta}_{t+1} = \widehat{\theta}_t - \eta g_t$
```

We have the following upper bound on the regret.

Theorem 2.3 Seq-SGD: Rate for Lipschitz convex function.
Let $(f_t : \mathbb{R}^d \to \mathbb{R})_{t\geq 1}$ be a sequence of differentiable and convex functions and $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Define the sequence $(\widehat{\theta}_t)_{t\geq 1}$ by induction,

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - \eta g_t, \quad t \geq 1, \quad \textit{with } \eta = \frac{R}{L\sqrt{T}},$$

where $R \geq \|\widehat{\theta}_1 - \theta^*\|$ and where g_t is any $\mathcal{F}_t$-measurable random variable fulfilling

$$\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f_t(\widehat{\theta}_t), \quad \textit{and} \quad \mathbb{E}[\|g_t\|^2] \leq L^2, \quad \textit{for } t = 1, \dots, T.$$

Then, the mean regret of the sequence of the SGD is upper-bounded by

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_t(\widehat{\theta}_t) - \min_{\theta\in\mathbb{R}^d}\frac{1}{T}\sum_{t=1}^{T} f_t(\theta)\right] \leq \frac{LR}{\sqrt{T}}.$$

Remark. Theorem 2.3 ensures that the average cost $\frac{1}{T}\sum_{t=1}^{T} f_t(\widehat{\theta}_t)$ of the sequence $(\widehat{\theta}_t)_{t\geq 1}$ induced by Seq-SGD is almost as small as the average cost $\frac{1}{T}\sum_{t=1}^{T} f_t(\theta^*)$ of the best constant choice θ^*, up to a $O(T^{-1/2})$-term.

Proof of Theorem 2.3. The proof is exactly the same as for Theorem 2.2, simply replacing everywhere f by f_t. We give yet the main lines for convenience of the reader.
From (2.6) and $\mathbb{E}\left[g_t|\mathcal{F}_{t-1}\right] = \nabla f_t(\widehat{\theta}_t)$, we get as for Theorem 2.2

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}(f_t(\widehat{\theta}_t) - f_t(\theta^*))\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla f_t(\widehat{\theta}_t), \widehat{\theta}_t - \theta^*\rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\langle g_t, \widehat{\theta}_t - \theta^*\rangle\right]. \qquad (2.16)\end{aligned}$$

Plugging the Equality (2.12), leads to

$$\mathbb{E}\left[\sum_{t=1}^{T}(f_t(\widehat{\theta}_t) - f_t(\theta^*))\right] \leq \frac{R^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|g_t\|^2\right] \leq \frac{R^2}{2\eta} + \frac{\eta}{2}L^2 T.$$

The result of Theorem 2.3 follows by taking the value $\eta = R/(L\sqrt{T})$. □

2.5 Problem: Sequential Projected Gradient Descent

For a sequence $(f_t)_{t\geq 1}$ of convex functions $f_t : \mathcal{D} \to \mathbb{R}$ and C included in the interior of $\mathcal{D}$, we want to sequentially approximate the minimum

$$\min_{\theta\in C}\frac{1}{T}\sum_{t=1}^{T} f_t(\theta). \tag{2.17}$$

In the previous sections, we have made the simplifying assumption that $\mathcal{D} = C = \mathbb{R}^d$. Yet, in many instances, we need to minimize (2.17) on some specific set C, and not on the whole $\mathbb{R}^d$. In this section, we will adapt (S)GD by adding a projection step in order to handle (2.17).
We consider a (non-empty) compact convex set $C \subset \mathbb{R}^d$, included in the interior of $\mathcal{D}$. We assume below that there exists a unique $\theta^* \in C$ solution to the minimisation problem (2.17).

2.5.1 Projection onto a convex set

As C is compact, the set of minimizers $\mathrm{argmin}_{u\in C}\|z-u\|^2$ is non-empty. We denote by

$$\pi_C z \in \underset{u\in C}{\mathrm{argmin}}\,\|z-u\|^2$$

one of these minimizers.

1. Let us fix $u \in C$ and $0 < t < 1$. Why do we have $\|z - (tu + (1-t)\pi_C z)\|^2 \geq \|z - \pi_C z\|^2$?
2. Investigating this inequality for t vanishing to 0, prove that

$$\langle u - \pi_C z, z - \pi_C z\rangle \leq 0 \quad \text{and} \quad \|\pi_C z - z\|^2 + \|u - \pi_C z\|^2 \leq \|u - z\|^2. \tag{2.18}$$

3. Prove that the projection $\pi_C z$ of z onto the convex set C is uniquely defined.

2.5.2 Rate for Lipschitz convex functions

We assume in the following that each function f_t is differentiable on the interior of $\mathcal{D}$. To solve (2.17), we can apply the Sequential Projected Gradient Descent algorithm (with $\eta > 0$):

Sequential Projected Gradient Descent (SeqPGD)
Input: $\eta > 0$, and $\theta_1 \in C$.
Iterate: $\theta_{t+1} = \pi_C(\theta_t - \eta\nabla f_t(\theta_t))$, for $t = 1, \ldots, T-1$
Output: $\bar{\theta}_T$ or θ_T

We will prove the following theorem.

Theorem 2.4 Sequential PGD.
Assume that

$$\max_{t=1,\ldots,T}\max_{\theta\in C}\|\nabla f_t(\theta)\| \leq L, \quad \textit{and} \quad \max_{\theta\in C}\|\theta - \theta_1\| \leq R.$$

Then, for $\eta = R/(L\sqrt{T})$

$$\frac{1}{T}\sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta\in C}\frac{1}{T}\sum_{t=1}^{T} f_t(\theta) \leq \frac{LR}{\sqrt{T}}.$$

Remark. We get a bound similar to the bound in Theorem 2.3. Compared to Theorem 2.3, we notice that the choice $\eta = R/(L\sqrt{T})$ can be implemented in practice, as R and L can be computed from C and f_t.

Proof of Theorem 2.4.
For the analysis of the algorithm, we introduce the notation $\theta_t^+ = \theta_t - \eta\nabla f_t(\theta_t)$.

1. Similarly as in the unconstraint case, prove that

$$f_t(\theta_t) - f_t(\theta^*) \leq \frac{1}{\eta}\langle \theta_t - \theta_t^+, \theta_t - \theta^* \rangle = \frac{\eta}{2}\|\nabla f_t(\theta_t)\|^2 + \frac{1}{2\eta}\left(\|\theta_t - \theta^*\|^2 - \|\theta_t^+ - \theta^*\|^2\right).$$

2. With (2.18), prove that

$$\frac{1}{T}\sum_{t=1}^{T}(f_t(\theta_t) - f_t(\theta^*)) \leq \frac{\eta L^2}{2} + \frac{\|\theta_1 - \theta_*\|^2}{2\eta T}.$$

3. Conclude.

The proof is complete. □

Exercise. Extend the analysis above to the case where, as in Theorem 2.3, we only have access to a stochastic gradient g_t instead of $\nabla f_t(\theta_t)$.
Remark. When all the f_t are identically equal to f, we have the next corollary of Theorem 2.4.

Corollary 2.5 Rate for Lipschitz convex function.
Assume that $f_t = f$ *for all* $t \geq 1$. *Assume also that* $\max_{x\in C}\|\nabla f(x)\| \leq L$ *and that* $\max_{x,y\in C}\|x - y\| \leq R$.
Then, for $\eta = R/(L\sqrt{T})$

$$f(\bar{\theta}_T) - \min_{\theta\in C} f(\theta) \leq \frac{LR}{\sqrt{T}}.$$

Chapter 3

Prediction with experts

3.1 Introduction

3.1.1 The learning problem

We consider the problem were we want to predict a sequence $(y(t))_{t\geq 1}$ of real valued outcomes, based on some expert predictions. More precisely, at each time $t \geq 1$, we have access to d predictions of experts $h(t) = (h_1(t), \dots, h_d(t)) \in \mathbb{R}^d$ and our goal is to predict $y(t)$ based on these expert predictions $h(t)$. We will predict $y(t)$ by taking a convex combination $\langle \theta_t, h(t)\rangle$ of the expert predictions, usually referred to as ”convex aggregation” of the expert predictions.

The information available at time t for the prediction is

$$I_t = \big(y(1), \dots, y(t-1), h(1), \dots, h(t-1)\big) \in \mathbb{R}^{(d+1)(t-1)}.$$

An aggregation strategy $\widehat{\theta}$ is a sequence of mapping $\{\widehat{\theta}_t : t \geq 1\}$ with $\widehat{\theta}_t : \mathbb{R}^{(d+1)(t-1)} \to S_d$, where S_d is the simplex $S_d = \left\{x \in [0,1]^d : |x|_1 = 1\right\}$.
For a given strategy $\widehat{\theta}$, the outcome $y(t)$ is predicted by the convex aggregation

$$\langle \widehat{\theta}_t(I_t), h(t)\rangle = \sum_{j=1}^{d} [\widehat{\theta}_t(I_t)]_j \, h_j(t).$$

To avoid cluttered notations, we will use the simple notation $\theta_t = \widehat{\theta}_t(I_t)$ in the following.

3.1.2 The regret of a strategy

Let us consider a loss function $\ell : \mathbb{R}^2 \to \mathbb{R}$, convex in the first variable. Our goal is to find a strategy $\widehat{\theta}$ such that the cumulated loss

$$\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) \tag{3.1}$$

is as small as possible.

We will compare a strategy to a best constant aggregation strategy θ^*

$$\theta^* \in \operatorname*{argmin}_{\theta \in S_d} \sum_{t=1}^{T} \ell\left(\langle \theta, h(t)\rangle, y(t)\right).$$

We observe that θ^* always exists as the objective function $\theta \to \sum_{t=1}^{T} \ell\left(\langle \theta, h(t)\rangle, y(t)\right)$ is convex on the compact convex set S_d.

The regret of a strategy $\widehat{\theta}$ is defined as

$$\mathcal{R}(\widehat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) - \frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta^*, h(t)\rangle, y(t)\right).$$

In most of this chapter, we will derive upper-bounds on $\mathcal{R}(\widehat{\theta})$, which will be valid for *any* sequences $(y(t))_{t\geq 1}$ and $(h(t))_{t\geq 1}$. The case where $(y(t))_{t\geq 1}$ is generated according to some random mechanism, will be discussed briefly in the last section.

3.2 Warm-up: aggregation with SeqPGD

3.2.1 SeqPGD for expert aggregation

Setting $f_t(\theta) = \ell\left(\langle\theta, h(t)\rangle, y(t)\right)$, we are exactly in the setting investigated in Chapter 2, Section 2.5, where we want to get $\sum_{t=1}^T f_t(\theta_t)$ as small as possible. Hence, if the loss ℓ is differentiable in the first variable, we can differentiate f_t

$$\nabla f_t(\theta) = \partial_1\ell\left(\langle\theta, h(t)\rangle, y(t)\right) h(t),$$

and use as an aggregation strategy the sequence $(\theta_t)_{t\geq 1}$ produced by the SeqPGD

$$\theta_{t+1} = \pi_{S_d}(\theta_t - \eta\nabla f_t(\theta_t)), \quad t = 1, 2, \dots, \tag{3.2}$$

where π_{S_d} is the projection onto the simplex S_d as defined in Section 2.5.1.

According to Theorem 2.4, the regret can then be upper-bounded in terms of $R = \max_{\theta,\theta'\in S_d} \|\theta - \theta'\|$ and

$$L = \max_{t=1,\dots,T}\max_{\theta\in S_d} \|\nabla f_t(\theta)\| = \max_{t=1,\dots,T}\max_{\theta\in S_d} \|\partial_1\ell\left(\langle\theta, h(t)\rangle, y(t)\right) h(t)\|.$$

We observe that $|\theta_j - \theta'_j| \leq 1$ for any $\theta, \theta' \in S_d$, so

$$R^2 \leq \max_{\theta,\theta'\in S_d} |\theta - \theta'|_1 \leq 2.$$

As for L, let us assume that both the outcomes $y(t)$ and the expert predictions $h_j(t)$ take values in $[-M, M]$. Then, if $\partial_1\ell$ is continuous, the derivative $\partial_1\ell\left(\langle\theta, h(t)\rangle, y(t)\right)$ is bounded in absolute value by some constant C for any $\theta \in S_d$ and $t \geq 1$ and hence

$$\max_{t=1,\dots,T}\max_{\theta\in S_d} |\nabla f_t(\theta)|_\infty = \max_{t=1,\dots,T}\max_{\theta\in S_d} |\partial_1\ell\left(\langle\theta, h(t)\rangle, y(t)\right) h(t)|_\infty \leq CM.$$

Therefore we have $L \leq CM\sqrt{d}$ and the upper-bound of Theorem 2.4, Chapter 2, gives

$$\mathcal{R}(\widehat{\theta}^{\text{SeqPGD}}) \leq CM\sqrt{\frac{2d}{T}}. \tag{3.3}$$

Remark. Let us comment on the nature of this result.

1. The regret (3.3) for SeqPGD holds for *any* sequences $(y(t))_{t\geq 1}$ and $(h_j(t))_{t\geq 1}$, for $j = 1, \dots, d$, with values in $[-M, M]$. It means that, whatever these sequences, without any additional knowledge, we can be almost as good as the best combination of the experts, with the regret (3.3) tending to 0 at rate $1/\sqrt{T}$. This can sound as magic, but we emphasize that:
   - We only compare to the best combination of experts ”on average”, and this best combination of experts ”on average” may give very bad prediction at some epoch t;
   - Even ”on average”, if all experts are very poor in terms of prediction, the aggregated prediction will also be very poor.
2. In the above bound, we notice that the upper-bound on the regret grows like $\sqrt{d}$ with the number d of experts. As we may wish to combine the predictions of many different experts, it is important to understand if we can have a better dependence on the number of experts. As we will see, the $\sqrt{d}$ can be reduced to a $\sqrt{\log(d)}$ for some more suitable aggregation strategies.

3.2.2 Linearized problem

To get a better intuition on the problem, let us consider a linearized version of our problem. Let us set

$$\ell_t = [\ell_{j,t}]_{j=1,\dots,d} = [\ell(h_j(t), y(t))]_{j=1,\dots,d}.$$

Since ℓ is convex on the first variable and since $\theta_t \in S_d$, Jensen inequality ensures the upper-bound

$$\ell\left(\langle\theta_t, h(t)\rangle, y(t)\right) = \ell\left(\sum_{j=1}^{d} \theta_{j,t} h_j(t), y(t)\right) \leq \sum_{j=1}^{d} \theta_{j,t} \ell(h_j(t), y(t)) = \langle\theta_t, \ell_t\rangle. \tag{3.4}$$

As a warm-up, we may investigate first the simpler problem with linear objective functions $f_t(\theta) = \langle\theta, \ell_t\rangle$. For a strategy $\widehat{\theta}$, we wish to control the linearized regret

$$\mathcal{R}_{lin}(\widehat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\langle\theta_t, \ell_t\rangle - \min_{\theta\in S_d}\frac{1}{T}\sum_{t=1}^{T}\langle\theta, \ell_t\rangle = \frac{1}{T}\sum_{t=1}^{T}\langle\theta_t, \ell_t\rangle - \min_{j=1,\dots,d}\frac{1}{T}\sum_{t=1}^{T}\ell_{j,t}, \tag{3.5}$$

where the last equality follows from the fact that the map $\theta \to \sum_{t=1}^{T}\langle\theta, \ell_t\rangle$ is linear, so it is maximized at one of the extremal points of S_d, namely, at one of the vectors of the canonical basis in $\mathbb{R}^d$.

As before, we can aggregate the predictions $h(t)$ according to a SeqPGD. As $\nabla\langle\theta, \ell_t\rangle = \ell_t$, the updates are then $\theta_{t+1} = \pi_{S_d}(\theta_t - \eta\ell_t)$. If we assume that the losses $\ell_{j,t} = \ell(h_j(t), y(t))$ are uniformly bounded by some B, then $\|\ell_t\| \leq \sqrt{d}|\ell_t|_\infty \leq \sqrt{d}B$ and the upper-bound of of Theorem 2.4 gives

$$\mathcal{R}_{lin}(\widehat{\theta}^{\text{SeqPGD}}) \leq B\sqrt{\frac{2d}{T}}. \tag{3.6}$$

3.3 Aggregation with exponential updates

3.3.1 Exponential updates in the linearized problem

In the linearized problem with $f_t(\theta) = \langle\theta, \ell_t\rangle$, the SeqPGD iterate at time $t+1$ amounts to update θ_t into $\theta_t - \eta\ell_t$ and then to project the result on S_d according to π_{S_d}. A glimpse at (3.5) shows that the optimal θ is concentrated on a single expert, so we may wish to discard more strongly the weights $\theta_{j,t}$ corresponding to experts with strong loss $\ell_{j,t}$.

Following this direction, we can replace the linear discount $\theta_t - \eta\ell_t$, by an exponential discount $\theta_t e^{-\eta\ell_t}$. As for the projection step, we may simply renormalized the update by its ℓ^1-norm. This lead us to the exponential weights strategy

$$\theta_1 = \frac{\mathbf{1}}{d}, \quad \theta_{t+1} = \frac{\theta_t e^{-\eta\ell_t}}{|\theta_t e^{-\eta\ell_t}|_1}, \quad t = 1, 2, \dots, \tag{3.7}$$

where $\mathbf{1}$ stands for the d-dimensional vector with all coordinates equal to 1, and $\theta_t e^{-\eta\ell_t}$ stands for the vector with coordinates $\theta_{j,t}e^{-\eta\ell_{j,t}}$, for $j = 1, \dots, d$.

By a simple induction, we get the following closed-form formula for the weights θ_t of the exponential weights strategy

$$\theta_{j,t} = \frac{\exp\left(-\eta\sum_{s=1}^{t-1}\ell_{j,s}\right)}{\sum_{k=1}^{d}\exp\left(-\eta\sum_{s=1}^{t-1}\ell_{k,s}\right)}, \quad j = 1, \dots, d, \quad t = 1, 2, \dots. \tag{3.8}$$

In plain words, the exponential weights strategy consider the cumulated loss $L_{j,t-1} = \sum_{s=1}^{t-1}\ell_{j,s}$

of each expert j up to time t, and then gives a weight to the prediction $h_j(t)$ proportional to $\exp(-\eta L_{j,t-1})$. Hence, the smaller the cumulated loss $L_{j,t-1}$, the larger the weight $\theta_{j,t}$.

Next lemma bounds the regret (3.5) for the sequence $(\theta_t)_{t\geq 1}$ defined by (3.7) and for *any* bounded sequence $(\ell_t)_{t\geq 1}$.

Lemma 3.1 Bound for exponential aggregation.
For any sequence $(\ell_t)_{t\geq 1}$, with $\ell_t \in [a,b]^d$, the sequence $(\theta_t)_{t\geq 1}$ defined by (3.7) with $\eta = \sqrt{\frac{8\log(d)}{(b-a)^2 T}}$ fulfills

$$\sum_{t=1}^{T} \langle \theta_t, \ell_t \rangle - \min_{j=1,\dots,d} \sum_{t=1}^{T} \ell_{j,t} \leq (b-a)\sqrt{\frac{T\log(d)}{2}}.$$

As a direct corollary of Lemma 3.1, we have the following theorem.

Theorem 3.2 Bound on the regret.
Let $\mathcal{Y}$ be an interval in $\mathbb{R}$. Assume that $\ell : \mathcal{Y}\times\mathcal{Y} \to [0,B]$ is convex in the first variable. Then for any sequences $(y(t))_{t\geq 1}$ and $(h_j(t))_{t\geq 1}$, for $j = 1,\dots,d$, with values in $\mathcal{Y}$, the exponential weight strategy (3.7) with $\eta = \sqrt{\frac{8\log(d)}{B^2 T}}$ fulfills

$$\frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) - \min_{j=1,\dots,d} \frac{1}{T}\sum_{t=1}^{T} \ell\left(h_j(t), y(t)\right) \leq B\sqrt{\frac{\log(d)}{2T}}.$$

Remark. Before proving Lemma 3.1 and Theorem 3.2, let us comment on theses last results. Comparing the above regrets with the regret (3.6) of the SeqPGD algorithm, we observe that:

- We have the same scaling $1/\sqrt{T}$ with respect to T;
- The $\sqrt{d}$ for SeqPGD has been replaced by $\sqrt{\log(d)}$ for exponential weighting. So the price of adding many experts is much lower in the exponential weight strategy than in SeqPGD.
- We compare the mean loss $\frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right)$ to the mean loss of the best expert $\min_{j=1,\dots,d} \frac{1}{T}\sum_{t=1}^{T} \ell\left(h_j(t), y(t)\right)$, and not to the mean loss of the best convex combination of the experts $\min_{\theta\in S_d} \frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta, h(t)\rangle, y(t)\right)$ as in (3.3).

Proof of Theorem 3.2.
According to (3.4) we have

$$\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) - \min_{j=1,\dots,d} \sum_{t=1}^{T} \ell\left(h_j(t), y(t)\right) \leq \sum_{t=1}^{T} \langle \theta_t, \ell_t\rangle - \min_{j=1,\dots,d} \sum_{t=1}^{T} \ell_{j,t}. \tag{3.9}$$

The proof of Theorem 3.2 then simply follows from Lemma 3.1 with $a = 0$ and $b = B$. □

Proof of Lemma 3.1.
For any $t = 1,\dots,T$, let $Z_t : \Omega \to [a,b]$ be a random variable with distribution

$$\mathbb{P}[Z_t = \ell_{j,t}] = \theta_{j,t}, \quad j = 1,\dots,d.$$

According to Lemma 1.1 Chapter 1, we have

$$\log \mathbb{E}\left[e^{-\eta(Z_t - \mathbb{E}[Z_t])}\right] \leq \frac{(b-a)^2\eta^2}{8},$$

from which follows

$$\mathbb{E}\left[Z_t\right] \leq \frac{(b-a)^2\eta}{8} - \frac{1}{\eta}\log\left(\mathbb{E}\left[e^{-\eta Z_t}\right]\right). \tag{3.10}$$

We observe first that $\mathbb{E}[Z_t] = \langle \theta_t, \ell_t \rangle$. Second, setting $L_t = \sum_{s=1}^t \ell_s$, with the convention that $L_0 = 0 \in \mathbb{R}^d$, we have from (3.8)

$$\theta_t = \frac{e^{-\eta L_{t-1}}}{|e^{-\eta L_{t-1}}|_1}, \quad \text{for any } t = 1, \dots, T,$$

and hence

$$\mathbb{E}\left[e^{-\eta Z_t}\right] = \frac{\sum_{j=1}^d e^{-\eta L_{j,t-1}} e^{-\eta \ell_{j,t}}}{|e^{-\eta L_{t-1}}|_1} = \frac{|e^{-\eta L_t}|_1}{|e^{-\eta L_{t-1}}|_1}, \quad \text{for any } t = 1, \dots, T.$$

Summing the Inequality (3.10) over t then gives

$$\begin{aligned}\sum_{t=1}^T \langle \theta_t, \ell_t \rangle &\leq \frac{(b-a)^2 \eta T}{8} - \frac{1}{\eta} \sum_{t=1}^T \left(\log(|e^{-\eta L_t}|_1) - \log(|e^{-\eta L_{t-1}}|_1) \right) \\ &= \frac{(b-a)^2 \eta T}{8} - \frac{1}{\eta} \left(\log(|e^{-\eta L_T}|_1) - \log(|e^{-\eta L_0}|_1) \right).\end{aligned}$$

To conclude, we notice that $|e^{-\eta L_0}|_1 = d$ and $|e^{-\eta L_T}|_1 \geq \max_{j=1,\dots,d} e^{-\eta L_{j,T}}$, so that

$$\sum_{t=1}^T \langle \theta_t, \ell_t \rangle \leq \frac{(b-a)^2 \eta T}{8} + \frac{\log(d)}{\eta} + \min_{j=1,\dots,d} L_{j,T}.$$

For $\eta = \sqrt{\frac{8 \log(d)}{(b-a)^2 T}}$, the claim of Lemma 3.1 follows. □

Remark: Do you notice some similarities between the proof of Lemma 3.1 and the proof of Theorem 2.2?

3.3.2 Faster aggregation rates for square loss

We may wonder whether the learning rate $\sqrt{\log(d)/T}$ appearing in Theorem 3.2 is optimal. It can be shown that the Lemma 3.1 cannot be (significantly) improved, in the following sense. There exist some sequences $(\ell_t)_{t \geq 1}$ for which, for any sequence $(\theta_t)_{t \geq 1}$ with θ_t depending only on the past losses $\ell_1, \dots, \ell_{t-1}$, we have the minoration

$$\liminf_{d \to \infty} \liminf_{T \to \infty} \sqrt{\frac{2}{(b-a)^2 T \log(d)}} \left(\sum_{t=1}^T \langle \theta_t, \ell_t \rangle - \min_{j=1,\dots,d} \sum_{t=1}^T \ell_{j,t} \right) \geq 1.$$

Such sequences can be generated by sampling the $\ell_{j,t}$ i.i.d. with Bernoulli$(1/2)$ distribution.

Yet, the proof of Theorem 3.2 starts with Jensen inequality (3.9), and there is a room for improvement for strongly convex losses. In this section, we exhibit this phenomenon for the square loss $\ell(\langle \theta, h(t) \rangle, y(t)) = (y(t) - \langle \theta, h(t) \rangle)^2$.

We assume that

$$\sup_{t \geq 1} |y(t)| \leq \sqrt{B}/2, \quad \sup_{t \geq 1} |h(t)|_\infty \leq \sqrt{B}/2, \tag{3.11}$$

so that the losses $\ell_{j,t} = (y(t) - h_j(t))^2$ belong to $[0, B]$ as in the Theorem 3.2.

Theorem 3.3 Bound on the regret for the quadratic loss.
For any sequences $(y(t))_{t \geq 1}$ and $(h_j(t))_{t \geq 1}$, for $j = 1, \dots, d$ fulfilling (3.11), the exponential weight strategy (3.7) with $\eta = 1/(2B)$ fulfills

$$\frac{1}{T} \sum_{t=1}^T (y(t) - \langle \theta_t, h(t) \rangle)^2 - \min_{j=1,\dots,d} \frac{1}{T} \sum_{t=1}^T \left(y(t) - h_j(t) \right)^2 \leq \frac{2B \log(d)}{T}.$$

Remarks. In the setting where $T \geq 2\log(d)$, which typically holds when time grows, we notice that:

1. We can choose $\eta = 1/(2B)$ which is possibly much larger than the choice $\eta = \sqrt{\frac{8\log(d)}{B^2T}}$ of Theorem 3.2;
2. The regret then scales as $\log(d)/T$ instead of $\sqrt{\log(d)/T}$ as in Theorem 3.2;
3. This improvement is linked to this larger choice for η. Indeed, inspecting the proof below, we observe that for the choice $\eta = \sqrt{\frac{8\log(d)}{B^2T}}$ as in Theorem 3.2, we would only a get the bound $B\sqrt{\frac{\log(d)}{8T}}$ on the regret.

Proof of Theorem 3.3.
According to (3.11), we have $|y(t) - h_j(t)| \leq \sqrt{B}$, for all $j = 1, \ldots, d$ and $t = 1, \ldots, T$.
Let us define the random variable $Z_t : \Omega \longrightarrow [-\sqrt{B}, \sqrt{B}]$ by

$$\mathbb{P}\left[Z_t = y(t) - h_j(t)\right] = \theta_{j,t}, \quad \text{for } j = 1, \ldots, d.$$

The map $x \longrightarrow e^{-\eta x^2}$ is concave on $[-(2\eta)^{-1/2}, (2\eta)^{-1/2}]$, and for $\eta \leq 1/(2B)$, we have $[-\sqrt{B}, \sqrt{B}] \subset [-(2\eta)^{-1/2}, (2\eta)^{-1/2}]$. So, Jensen inequality

$$\mathbb{E}\left[\exp\left(-\eta Z_t^2\right)\right] \leq \exp\left(-\eta \mathbb{E}\left[Z_t\right]^2\right)$$

ensures that

$$\begin{aligned} |\theta_t e^{-\eta \ell_t}|_1 &= \sum_{j=1}^{d} \theta_{j,t} e^{-\eta(y(t)-h_j(t))^2} \\ &\leq \exp\left(-\eta\left(\sum_{j=1}^{d}\theta_{j,t}(y(t) - h_j(t))\right)^2\right) = \exp\left(-\eta\big(y(t) - \langle\theta_t, h(t)\rangle\big)^2\right). \end{aligned} \tag{3.12}$$

Setting $L_t = \sum_{s=1}^{t} \ell_s$, with the convention that $L_0 = 0 \in \mathbb{R}^d$, we have from (3.8)

$$\theta_t = \frac{e^{-\eta L_{t-1}}}{|e^{-\eta L_{t-1}}|_1}, \quad \text{for any } t = 1, \ldots, T.$$

So inequality (3.12) gives for any $t = 1, \ldots, T$

$$(y(t) - \langle\theta_t, h(t)\rangle)^2 \leq -\frac{1}{\eta}\log|\theta_t e^{-\eta\ell_t}|_1 = -\frac{1}{\eta}\log\left(\frac{|e^{-\eta L_t}|_1}{|e^{-\eta L_{t-1}}|_1}\right).$$

Summing these inequalities over t, we conclude the proof as in Lemma 3.1

$$\sum_{t=1}^{T}(y(t) - \langle\theta_t, h(t)\rangle)^2 \leq -\frac{1}{\eta}\left(\log(|e^{-\eta L_T}|_1) - \log(|e^{-\eta L_0}|_1)\right) \leq \frac{\log(d)}{\eta} + \min_{j=1,\ldots,d} L_{j,T}.$$

As $\eta = 1/(2B)$, the proof is complete. □

3.3.3 Exponential updates for the original problem

In Section 3.3.1, we have derived the strategy (3.7) from the linearized problem, by replacing linear discounts, by exponential discounts. Let us come back to the original problem from Section 3.2.1, and try to adapt this strategy directly to the original problem.

Let us set $g_t := \nabla f_t(\theta_t) = \partial_1 \ell(\langle \theta_t, h(t)\rangle, y(t))h(t)$. We can replace the linear discount $\theta_t - \eta g_t$ in the SPGD (3.2), by an exponential discount $\theta_t e^{-\eta g_t}$. Hence, we consider the exponential weight strategy relative to the gradients

$$\theta_1 = \frac{\mathbf{1}}{d}, \quad \theta_{t+1} = \frac{\theta_t e^{-\eta g_t}}{|\theta_t e^{-\eta g_t}|_1}, \quad t = 1, 2, \ldots. \tag{3.13}$$

As before, we have the closed form formula

$$\theta_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} g_{j,s}\right)}{\sum_{k=1}^{d} \exp\left(-\eta \sum_{s=1}^{t-1} g_{k,s}\right)}. \tag{3.14}$$

A simple adaptation of the proof of Theorem 3.2 gives the following bound on the regret.

Theorem 3.4 Bound on the regret for (3.13).
Assume that $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is convex and differentiable in the first variable. Then for any sequences $(y(t))_{t \geq 1}$ with values in $\mathbb{R}$ and $(h(t))_{t \geq 1}$ with values in $\mathbb{R}^d$ such that,

$$\max_{t=1,\ldots,T} \max_{\theta \in S_d} |\partial_1 \ell(\langle \theta, h(t)\rangle, y(t))h(t)|_\infty \leq L, \tag{3.15}$$

the exponential weight strategy (3.13) with $\eta = \sqrt{\frac{2\log(d)}{L^2 T}}$ fulfills

$$\frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) - \min_{\theta \in S_d} \frac{1}{T}\sum_{t=1}^{T} \ell\left(\langle \theta, h(t)\rangle, y(t)\right) \leq L\sqrt{\frac{2\log(d)}{T}}.$$

Remarks. Let us compare this result to the result (3.3) for SeqPGD and to Theorem 3.2.

1. Similarly as for Theorem 3.2, we observe that compared to SeqPGD, the scaling of the regret bound with respect to the number d of experts is $\sqrt{\log(d)}$ instead of $\sqrt{d}$, which is much better.
2. The bound of Theorem 3.4 may seem very similar to the one of Theorem 3.2. There is a major difference though. We compare the performance of the aggregation strategy (3.13) to the performance to the best aggregation strategy θ^*, instead of the performance of the best expert. The performance of the best aggregated predictor can be much better than the performance of the best of the experts. In practice, it is common to observe that the predictions obtained from the aggregation (3.13) of the expert advices outperform the predictions of the best of the experts.

Proof of Theorem 3.4.
As the function $\theta \to \ell\left(\langle \theta, h(t)\rangle, y(t)\right)$ is convex and $g_t = \nabla_\theta \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right)$, we have

$$\begin{aligned}\sum_{t=1}^{T} \ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) - \min_{\theta \in S_d}\sum_{t=1}^{T} \ell\left(\langle \theta, h(t)\rangle, y(t)\right) &= \max_{\theta \in S_d}\sum_{t=1}^{T}\left(\ell\left(\langle \theta_t, h(t)\rangle, y(t)\right) - \ell\left(\langle \theta, h(t)\rangle, y(t)\right)\right)\\ &\leq \max_{\theta \in S_d}\sum_{t=1}^{T}\langle g_t, \theta_t - \theta\rangle\\ &= \sum_{t=1}^{T}\langle g_t, \theta_t\rangle - \min_{j=1,\ldots,d}\sum_{t=1}^{T} g_{j,t},\end{aligned}$$

where the last inequality follows from the fact that the minimum of a linear function on the simplex is achieved at the extremal points of the simplex.

As $g_t \in [-L, L]^d$, we can apply Lemma 3.1 with $a = -L$ and $b = L$ to get

$$\sum_{t=1}^{T} \langle g_t, \theta_t \rangle - \min_{j=1,\dots,d} \sum_{t=1}^{T} g_{j,t} \leq L\sqrt{2\log(d)T}.$$

The proof of Theorem 3.4 is complete. □

3.4 Mirror descent

3.4.1 Changing the geometry in gradient descent

Let us recall the recipe behind gradient descent. If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, then Taylor expansion ensures that for any $\theta, \theta_t \in \mathbb{R}^d$,

$$f(\theta) = f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + o(\|\theta - \theta_t\|).$$

When we do not have a closed-form formula for $\min_{\theta \in \mathbb{R}^d} f(\theta)$, in order to minimize f over $\mathbb{R}^d$, we may wish to replace f by a proxy more easily amenable to computations. If we replace f by the linear part in the Taylor expansion $\theta \longrightarrow f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle$, then the minimum is achieved for some θ with diverging norm and the difference between f and the linear proxy becomes large. Hence, we must constrain the minimizer not to be too far away from θ_t. A simple recipe is then to add a quadratic term $\|\theta - \theta_t\|^2$ which prevents the minimizer from being far away from θ_t. Hence, we can replace the minimisation problem $\min_{\theta \in \mathbb{R}^d} f(\theta)$ by

$$\min_{\theta \in \mathbb{R}^d} \left\{ f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta}\|\theta - \theta_t\|^2 \right\}. \tag{3.16}$$

The solution θ_{t+1} to the minimisation problem (3.16) is given by the closed-form formula

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t),$$

which corresponds to a step of gradient descent.

The penalization $\frac{1}{2\eta}\|\theta - \theta_t\|^2$ in (3.16) may not be the most suited one for the minimisation problem. It can be suboptimal in some cases, as in Section 3.3.1, where the minimum is achieved in some specific directions. It is then worth to replace the Euclidean norm $\frac{1}{2}\|\theta - \theta_t\|^2$ by some more suited divergence $D(\theta, \theta_t)$ allowing some larger steps in directions of interest.

Which divergence $D(\theta, \theta_t)$ shall we choose? If we come back to our minimisation problem $\min_{\theta \in \mathbb{R}^d} f(\theta)$, the ideal divergence is

$$D(\theta, \theta_t) = f(\theta) - f(\theta_t) - \langle \nabla f(\theta_t), \theta - \theta_t \rangle,$$

as then $f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + D(\theta, \theta_t) = f(\theta)$. Of course, this makes no sense in terms of optimization algorithm, as we do not have closed-form updates for minimizing f. Instead, we may use a proxy ϕ, which is amenable to closed-form updates and which induces a geometry suited to the minimisation problem. This motivates the definition of the Bregman divergence.

Bregman divergence

Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The Bregman divergence associated to ϕ is

$$D_\phi(\theta, \omega) = \phi(\theta) - \phi(\omega) - \langle \nabla \phi(\omega), \theta - \omega \rangle, \tag{3.17}$$

for $\theta, \omega \in \mathbb{R}^d$.

As ϕ is convex, we observe that the Bregman divergence takes non-negative values.

Replacing in (3.16) the Euclidean norm penalization $\frac{1}{2}\|\theta - \theta_t\|^2$ by the Bregman divergence $D_\phi(\theta, \theta_t)$, we get the update

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{\eta} D_\phi(\theta, \theta_t) \right\}. \tag{3.18}$$

As $\nabla_\theta D_\phi(\theta, \theta_t) = \nabla\phi(\theta) - \nabla\phi(\theta_t)$, differentiating the objective function, we get that θ_{t+1} is solution to

$$\nabla\phi(\theta_{t+1}) = \nabla\phi(\theta_t) - \eta\nabla f(\theta_t).$$

The algorithm iterating these updates is called Mirror Descent (MD).

Exercise. check that for $\phi(\theta) = \|\theta\|^2/2$, we have $D_\phi(\theta, \omega) = \|\theta - \omega\|^2/2$, and hence gradient descent is a special case of mirror descent for this choice of ϕ.

More generally, when we have a sequence f_t of objective functions, we may consider the sequential mirror descent.

Sequential Mirror Descent (SeqMD)
Input: $\theta_1 \in \mathbb{R}^d$, $\eta > 0$.
Iterate: For $t = 1, \dots, T-1$,

$$\nabla\phi(\theta_{t+1}) = \nabla\phi(\theta_t) - \eta\nabla f_t(\theta_t)$$

When the minimisation is constrained to occur in a compact convex set C, we will constrain the update (3.18) to occur in C

$$\begin{aligned}
\theta_{t+1} &\in \operatorname*{argmin}_{\theta \in C} \left\{ f_t(\theta_t) + \langle \nabla f_t(\theta_t), \theta - \theta_t \rangle + \frac{1}{\eta} D_\phi(\theta, \theta_t) \right\} \\
&= \operatorname*{argmin}_{\theta \in C} \left\{ \langle \nabla f_t(\theta_t), \theta \rangle + \frac{1}{\eta} \left(\phi(\theta) - \langle \nabla\phi(\theta_t), \theta \rangle \right) \right\} \\
&= \operatorname*{argmin}_{\theta \in C} \left\{ \phi(\theta) - \langle \nabla\phi(\theta_t) - \eta\nabla f_t(\theta_t), \theta \rangle \right\} \\
&= \operatorname*{argmin}_{\theta \in C} \underbrace{\left\{ \phi(\theta) - \phi(\omega_{t+1}) - \langle \nabla\phi(\omega_{t+1}), \theta \rangle \right\}}_{= D_\phi(\theta, \omega_{t+1}) - \langle \nabla\phi(\omega_{t+1}), \omega_{t+1} \rangle}, \quad \text{with } \nabla\phi(\omega_{t+1}) := \nabla\phi(\theta_t) - \eta\nabla f_t(\theta_t).
\end{aligned}$$

Let us denote by π_C^ϕ the projection relative to the Bregman divergence

$$\pi_C^\phi(\omega) \in \operatorname*{argmin}_{\theta \in C} D_\phi(\theta, \omega), \tag{3.19}$$

which is shown to be well defined in Section 3.4.2. The Projected Mirror Descent is then defined as follows.

Projected Mirror Descent (PMD)
Input: $\theta_1 \in \mathbb{R}^d$, $\eta > 0$.
Iterate: For $t = 1, \dots, T-1$,

- $\nabla\phi(\omega_{t+1}) = \nabla\phi(\theta_t) - \eta\nabla f_t(\theta_t)$
- $\theta_{t+1} = \pi_C^\phi(\omega_{t+1})$

As discussed above, the mirror descent principle allows to change the geometry of the updates of the gradient descent by replacing the Euclidean norm control on the step sizes by a control based on another metric, the Bregman divergence. For example, in the case of expert aggregation we want to favor large steps when far away from the extremal point of the simplex. It turns out that the exponential weight aggregation of experts described in Section 3.3.3 corresponds to a projected mirror descent with the geometry induced by the negative entropy function. This connection is established and illustrated in Section 3.4.3, after the general analysis of PMD in Section 3.4.2.

3.4.2 Regret bound for Projected Mirror Descent

In this part, we provide an upper-bound on the regret of PMD, in the spirit of the results for SeqPGD established in Section 2.5.

On Bregman projection

In the remaining of this section, we consider a slightly more general version of the setting considered in Section 3.4.1. Let $\mathcal{D}$ be a convex set of $\mathbb{R}^d$ and $\phi : \mathcal{D} \to \mathbb{R}$ a convex function differentiable on the interior $\mathring{\mathcal{D}}$ of its domain. The Bregman divergence $D_\phi : \mathcal{D} \times \mathring{\mathcal{D}} \to \mathbb{R}_+$ is then defined on $\mathcal{D} \times \mathring{\mathcal{D}}$ by (3.17).

Let $C \subset \mathcal{D}$ be a compact convex set. For any $\omega \in \mathring{\mathcal{D}}$, the function $\theta \to D_\phi(\theta, \omega)$ is convex on the compact set C, so the set of minimizers $\operatorname{argmin}_{\theta \in C} D_\phi(\theta, \omega)$ is not empty. In the following, as we want to be able to iterate the updates of the PMD, we assume that the following property holds

$$\text{for any } \omega \in \mathring{\mathcal{D}}, \text{ we have} \quad \mathring{\mathcal{D}} \cap \operatorname*{argmin}_{\theta \in C} D_\phi(\theta, \omega) \neq \emptyset. \tag{3.20}$$

Remark. The above property may not hold in some cases. As a counterexample, take $C = S_2$, the simplex in dimension 2, $\mathcal{D} = [0, +\infty)^2$, and $\phi(x) = \|x\|^2$.

The next lemma generalizes the results of Section 2.5.1 on the projection operator π_C.

Lemma 3.5 *Assume that $\phi : \mathcal{D} \to \mathbb{R}$ is strictly convex on C, so that $\mathcal{D}_\phi(x, y) > 0$ for any $x \in C$, $y \in C \cap \mathring{\mathcal{D}}$, with $x \neq y$. Assume also that (3.20) holds. Then, for any $y \in \mathring{\mathcal{D}}$ and any $z \in C$, we have*

1. *the projection $\pi_C^\phi(y)$ is uniquely defined and belongs to $C \cap \mathring{\mathcal{D}}$,*
2. *$\langle \nabla\phi(\pi_C^\phi(y)) - \nabla\phi(y), \pi_C^\phi(y) - z \rangle \leq 0$,*
3. *$D_\phi(z, \pi_C^\phi(y)) + D_\phi(\pi_C^\phi(y), y) \leq D_\phi(z, y)$.*

Remark. A simple induction shows that the assumptions of Lemma 3.5 ensure that if $\theta_1 \in C \cap \mathring{\mathcal{D}}$, then the subsequent updates $\theta_2, \theta_3, \ldots$ of the Projected Mirror Descent are also in $C \cap \mathring{\mathcal{D}}$. Hence, the PMD algorithm can be run indefinitely.

Proof of Lemma 3.5. Let πy be any element in $\mathring{\mathcal{D}} \cap \operatorname{argmin}_{\theta \in C} D_\phi(\theta, y)$, which is a non-empty set according to Assumption (3.20).
2- The function $H(s) = D_\phi(\pi y + s(z - \pi y), y)$ is defined on [0,1]. By definition of πy, it reaches its minimum at $s = 0$. Hence, the right derivative $H'(0) = \langle \nabla_1 D_\phi(\pi y, y), z - \pi y \rangle$ is non negative. As $\nabla_1 D_\phi(\pi y, y) = \nabla\phi(\pi y) - \nabla\phi(y)$, we get the second claim

$$\langle \nabla\phi(\pi y) - \nabla\phi(y), \pi y - z \rangle \leq 0.$$

3- For the third claim, we apply an analog of the polarisation formula

$$\langle \nabla\phi(a) - \nabla\phi(b), a - c \rangle = D_\phi(a, b) + D_\phi(c, a) - D_\phi(c, b), \quad \text{for any } a, b \in \mathring{\mathcal{D}},\ c \in \mathcal{D}, \tag{3.21}$$

with $a = \pi y$, $b = y$ and $c = z$ and we get

$$D_\phi(\pi y, y) + D_\phi(z, \pi y) - D_\phi(z, y) = \langle \nabla\phi(\pi y) - \nabla\phi(y), \pi y - z \rangle \leq 0. \tag{3.22}$$

The result follows.

1- It remains to prove the first claim. Let $z \in \operatorname{argmin}_{\theta \in C} D_\phi(\theta, y)$. As $D_\phi(\pi y, y) = D_\phi(z, y)$, the Inequality (3.22) gives $D_\phi(z, \pi y) = 0$ and hence $z = \pi y$ according to the strict convexity of ϕ. So, the projection $\pi_C^\phi y$ is uniquely defined and belongs to $C \cap \mathring{\mathcal{D}}$. □

Regret bound for PMD

In this section, we generalize the analysis of Section 2.5.2.

Let $|\cdot|$ be a norm on $\mathbb{R}^d$. We assume henceforth that ϕ is α-strongly convex with respect to this norm:

- α-strong convexity: $D_\phi(x,y) \geq \frac{\alpha}{2}|y-x|^2$ for all $x, y \in \mathring{\mathcal{D}} \cap C$. ($\alpha$-Cvx)

We notice that the α-strong convexity property (α-Cvx) implies the strict convexity of ϕ on C.

The objective functions $f_t : \mathcal{D} \to \mathbb{R}$ are assumed to be convex on $\mathcal{D}$ and differentiable on $\mathring{\mathcal{D}}$. Let us denote by $|\cdot|_*$ the dual norm of $|\cdot|$ on $\mathbb{R}^d$ with respect to the Euclidean scalar product

$$|y|_* = \sup_{|x|\leq 1} \langle x, y\rangle.$$

We assume below that the functions f_t are uniformly Lipschitz with respect to $|\cdot|$:

- Lipschitz condition: $|\nabla f_t(\theta)|_* \leq L$ for all $\theta \in \mathring{\mathcal{D}} \cap C$. (Lip)

Theorem 3.6 Regret bound for PMD

Under the Assumptions (3.20), (Lip), (α-Cvx), and $D_\phi(\theta^*, \theta_1) \leq R^2$, *we have for* $\eta = \sqrt{\frac{2\alpha R^2}{L^2 T}}$

$$\frac{1}{T}\sum_{t=1}^{T}(f_t(\theta_t) - f_t(\theta^*)) \leq RL\sqrt{\frac{2}{\alpha T}}.$$

Discussion. Before proving Theorem 3.6, let us discuss it. We observe that we obtain for PMD a result of the same nature as for SPGD, with the two following differences :

- R controls the divergence D_ϕ between θ^* and θ_1,
- L controls the dual norm of the gradients of f_t.

Hence, a good choice of ϕ for the PMD, is a choice fulfilling

- the updates of PMD can be easily computed,
- the maximum divergence $R^2 = \max_{\theta\in C} D_\phi(\theta, \theta_1)$ is as small as possible,
- the Lipschitz constant $L = \max_{\theta\in\mathring{\mathcal{D}}\cap C}\max_{t=1,\dots,T}|\nabla f_t(\theta)|_*$ in the dual norm related to D_ϕ is as small as possible.

Exercise. For $\phi(\theta) = \|\theta\|^2/2$, recover the result proved for SPGD in Chapter 2.

Proof of Theorem 3.6. In the analysis of SPGD, the starting point was the polarisation formula $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, applied with $a = \eta\nabla f_t(\theta_t)$ and $b = \theta_t - \theta^*$. We follow the same argument, but with the polarisation formula (3.21)

$$\begin{aligned}\eta(f_t(\theta_t) - f_t(\theta^*)) &\overset{convex}{\leq} \langle \eta\nabla f_t(\theta_t), \theta_t - \theta^*\rangle \\ &= \langle \nabla\phi(\theta_t) - \nabla\phi(\omega_{t+1}), \theta_t - \theta^*\rangle \\ &\overset{polar.}{=} D_\phi(\theta_t, \omega_{t+1}) + D_\phi(\theta^*, \theta_t) - D_\phi(\theta^*, \omega_{t+1}) \\ &\overset{Lem.3.5}{\leq} D_\phi(\theta_t, \omega_{t+1}) - D_\phi(\theta_{t+1}, \omega_{t+1}) + D_\phi(\theta^*, \theta_t) - D_\phi(\theta^*, \theta_{t+1}). \quad (3.23)\end{aligned}$$

Let us upper-bound the first difference in terms of the dual norm of the gradient. According to the assumption (α-Cvx) and according to the definition of the dual norm, we have

$$\begin{aligned}D_\phi(\theta_t,\omega_{t+1})-D_\phi(\theta_{t+1},\omega_{t+1}) &= \phi(\theta_t)-\phi(\theta_{t+1})-\langle\nabla\phi(\omega_{t+1}),\theta_t-\theta_{t+1}\rangle\\ &= \langle\underbrace{\nabla\phi(\theta_t)-\nabla\phi(\omega_{t+1})}_{=\eta\nabla f_t(\theta_t)},\theta_t-\theta_{t+1}\rangle - D_\phi(\theta_{t+1},\theta_t)\\ &\le \eta|\nabla f_t(\theta_t)|_*|\theta_t-\theta_{t+1}|-\frac{\alpha}{2}|\theta_{t+1}-\theta_t|^2\\ &\le \frac{\eta^2}{2\alpha}|\nabla f_t(\theta_t)|_*^2\,, \end{aligned}\tag{3.24}$$

where, for the last inequality, we used $2ab-b^2\le a^2$. Combining (3.23) and (3.24) and summing over t, we get

$$\begin{aligned}\sum_{t=1}^T(f_t(\theta_t)-f_t(\theta^*)) &\le \frac{\eta}{2\alpha}\sum_{t=1}^T|\nabla f_t(\theta_t)|_*^2+\frac{1}{\eta}(D_\phi(\theta^*,\theta_1)-D_\phi(\theta^*,\theta_{T+1}))\\ &\overset{\text{(Lip)}}{\le} \frac{\eta}{2\alpha}L^2T+\frac{R^2}{\eta}.\end{aligned}$$

Setting $\eta=\sqrt{\frac{2\alpha R^2}{L^2T}}$, we get the result. □

3.4.3 Problem: Projected Mirror Descent for expert aggregation

Exponential weights as PMD

Let us come back to our original problem, where, as discussed in Section 3.2.1, we seek for a strategy $\widehat{\theta}$ which minimizes the regret

$$\frac{1}{T}\sum_{t=1}^T f_t(\theta_t)-\min_{\theta\in S_d}\frac{1}{T}\sum_{t=1}^T f_t(\theta),$$

with $f_t(\theta)=\ell\left(\langle\theta,h(t)\rangle,y(t)\right)$. We denote by $g_t=\partial_1\ell\left(\langle\theta,h(t)\rangle,y(t)\right)h(t)$, the gradient of $f_t(\theta)$ and the constraint set is $C=S_d$.

Let us choose a function ϕ. As the gradients g_t fulfill (3.15), and as we wish to control the dual norm $|g_t|_*$ of the gradients, we wish that the dual norm $|\cdot|_*$ coincides with the sup-norm ℓ^∞. Hence, we want to choose a function ϕ which is strongly convex with respect to the ℓ^1-norm. At the same time, we wish to have a maximum divergence $\max_{\theta\in S_d}D_\phi(\theta,\theta_1)$ which is as small as possible. Let $\mathcal{D}=[0,+\infty)^d$ and let us define $\phi:\mathcal{D}\to\mathbb{R}$ by

$$\phi(\theta)=\sum_{j=1}^d\theta_j\log(\theta_j),\quad\text{for }\theta\in\mathcal{D},\tag{3.25}$$

with the convention $0\log(0)=0$. The gradient of ϕ is given by

$$\nabla\phi(\theta)=\left[1+\log(\theta_j)\right]_{j=1,\dots,d},\quad\text{for }\theta\in\mathring{\mathcal{D}},$$

and the Bregman divergence is given by

$$D_\phi(\theta,\omega)=\sum_{j=1}^d\left(\theta_j\log\left(\frac{\theta_j}{\omega_j}\right)+\omega_j-\theta_j\right),\quad\text{for }\theta\in\mathcal{D},\ \omega\in\mathring{\mathcal{D}}.\tag{3.26}$$

For $\theta\in S_d$, and $\omega\in S_d\cap\mathring{\mathcal{D}}$, this divergence is commonly called the Kullback-Leibler divergence, which will be investigated into more details in Chapter 4, Section 4.3.1.

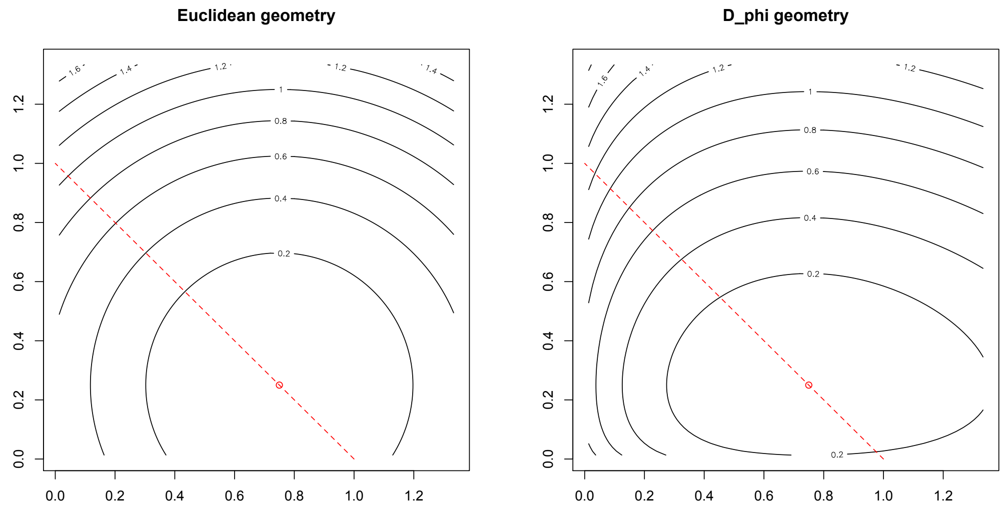


Figure 3.1: The Euclidean geometry $\theta \longrightarrow \|\theta - \omega\|^2$ (left) and the geometry $\theta \longrightarrow D_\phi(\theta, \omega)$ (right), as seen from $\omega = (3/4, 1/4)$ (red dot). The red dashed line represents the simplex S_2.

Let us compute the update of the PMD for this choice of ϕ. We have for $\theta_t > 0$

$$\log([\omega_{t+1}]_j) = \log([\theta_t]_j) - \eta g_{j,t}$$

so $\omega_{t+1} = \theta_t e^{-\eta g_t}$. Let us compute now the projection operator $\pi^\phi_{S_d}$ related to ϕ.

Lemma 3.7 *Let $\phi : \mathcal{D} \longrightarrow \mathbb{R}_+$ be defined by (3.25). Then for any $\omega \in \mathring{\mathcal{D}}$, we have*

$$\pi^\phi_{S_d}(\omega) = \frac{\omega}{|\omega|_1} \in \mathring{\mathcal{D}}.$$

As a consequence, we have the immediate corollary.

Corollary 3.8 *If we set $\theta_1 = \mathbf{1}/d$, the sequence produced by the projected mirror descent with ϕ given by (3.25) exactly corresponds to the exponential weight strategy (3.13).*

Proof of Lemma 3.7. You will prove the lemma by solving the next three questions.
Let us consider the Lagrangian associated to the constrained convex minimisation problem (3.19) with $C = S_d$

$$L(\theta, \mu) = D_\phi(\theta, \omega) + \mu\left(\sum_{j=1}^{d} \theta_j - 1\right), \quad \text{for } \theta \in \mathcal{D},\ \mu \in \mathbb{R}.$$

1. Compute the partial derivative with respect to θ_j of $L(\theta, \mu)$.
2. Check that the solution θ_μ to $\nabla_\theta L(\theta_\mu, \mu) = 0$ is $\theta_\mu = e^{-\mu}\omega$.
3. Conclude the proof of Lemma 3.7.

Upper bound on the regret

Let us translate the bound of Theorem 3.6 in the expert prediction setting, under the hypotheses of Theorem 3.4. The first step is to prove that D_ϕ is strongly convex with respect to the ℓ^1 norm on the simplex S_d.

Lemma 3.9 Pinsker inequality
For the Bregman divergence (3.26), we have

$$D_\phi(\theta, \omega) \geq \frac{1}{2}|\theta - \omega|_1^2, \quad \text{for any } \theta \in S_d, \text{ and } \omega \in S_d \cap \mathring{\mathcal{D}}.$$

Proof of Pinsker inequality. You will prove the Pinsker inequality by solving the next five questions.
Let us set $r_j = \theta_j/\omega_j - 1$ for $j = 1, \ldots, d$ and $\psi(t) = (1+t)\log(1+t) - t$ for $t \geq -1$, with the convention that $0\log(0) = 0$.

1. Check that g defined by $g(t) := (1 + t/3)\psi(t)$ for $t > -1$, fulfills $g'(0) = 0$ and
$$g''(t) = 1 + \frac{2\psi(t)}{3(1+t)}, \quad \text{for } t > -1.$$
2. Prove that
$$\psi(t) \geq \frac{t^2}{2(1+t/3)}, \quad \text{for all } t \geq -1.$$
3. Check that
$$D_\phi(\theta, \omega) = \sum_{j=1}^d \omega_j \psi(r_j) \geq \frac{1}{2}\sum_{j=1}^d \omega_j \frac{r_j^2}{1 + r_j/3}.$$
4. Prove the inequalities
$$|\theta - \omega|_1^2 = \left(\sum_{j=1}^d \omega_j |r_j|\right)^2 \leq \sum_{j=1}^d \omega_j \frac{r_j^2}{1+r_j/3} \times \sum_{k=1}^d \omega_k (1 + r_k/3)$$
$$\leq 2D_\phi(\theta, \omega).$$

The proof of Pinsker inequality is complete. □

Exercise. Adapt the proof of Lemma 3.9 to prove the general version of Pinsker inequality: For any probability distributions $\mathbb{P}, \mathbb{Q}$, with $\mathbb{P} \ll \mathbb{Q}$, we have

$$\frac{1}{2}\mathbb{E}_{\mathbb{Q}}\left[\left|\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right|\right]^2 \leq \mathbb{E}_{\mathbb{P}}\left[\log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right].$$

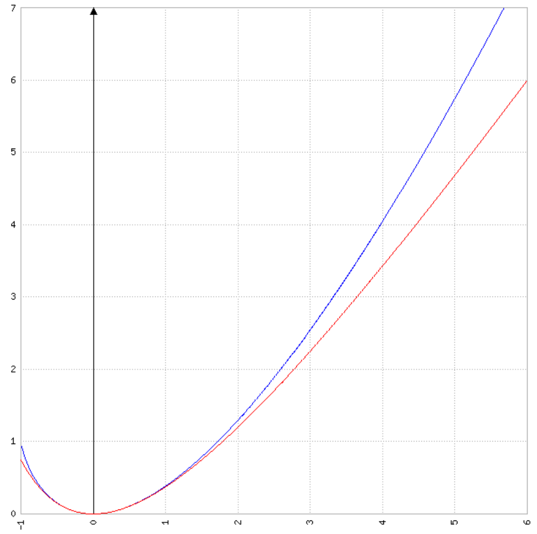


Figure 3.2: Plot of the function $\psi(t) = (1+t)\log(1+t) - t$ in blue and $t \to \frac{t^2}{2(1+t/3)}$ in red.

Pinsker inequality ensures that D_ϕ is 1-strongly convex with respect to the ℓ^1 norm on the simplex S_d. Let us now bound the gradients in terms of the dual norm. The dual norm of the ℓ^1 norm is the ℓ^∞ norm. Under the hypothesis (3.15) of Theorem 3.4, we have $|g_t|_\infty \leq L$ for $t = 1, \ldots, T$.

It remains to bound $D_\phi(\theta^*, \theta_1)$ for $\theta_1 = \mathbf{1}/d$. We observe that for any $\theta^* \in C$,

$$D_\phi(\theta^*, \theta_1) = \sum_{j=1}^{d} \theta_j^* \log(\theta_j^* d) \leq \sum_{j=1}^{d} \theta_j^* \log(d) = \log(d).$$

Hence, the assumptions of Theorem 3.6 hold with $R^2 = \log(d)$ and $\alpha = 1$, so Theorem 3.6 ensures that

$$\frac{1}{T}\sum_{t=1}^{T}(f_t(\theta_t) - f_t(\theta^*)) \leq L\sqrt{\frac{2\log(d)}{T}},$$

for the exponential weights strategy (3.13). We exactly recover the bound of Theorem 3.4.

3.5 Illustration

Let us illustrate the aggregation strategies (3.7) and (3.13) on some pollution data. We will work with a data set gathering O3 concentration, temperature, nebulosity, rain, wind, etc in Brittany at different times of the day. Our goal will be to predict the ozone O3 concentration from weather observations. The R-code can be downloaded at `https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/MathIA/Experts.R`

Assume that we have the forecasts of three experts. A first expert, mister Heat, knows that O3 appears when the temperature is high enough. So he decides to predict the O3 concentration with a linear combination of the O3 concentration of the day before and the temperature at midday. A second expert, misses Sun, knows that some sun is needed in order to have a reaction producing O3. So, she decides to predict O3 concentration with a linear combination of the O3 concentration of the day before and the morning nebulosity index. Finally, a local expert claims ”come on, we are in Brittany, with a lot of wind and rain, no pollution can appear with such conditions”. So he decides to predict O3 concentration with a linear combination of the O3 concentration of the day before, the wind and the rain intensity. The predictions of these 3 experts are displayed in the chart 3.3.

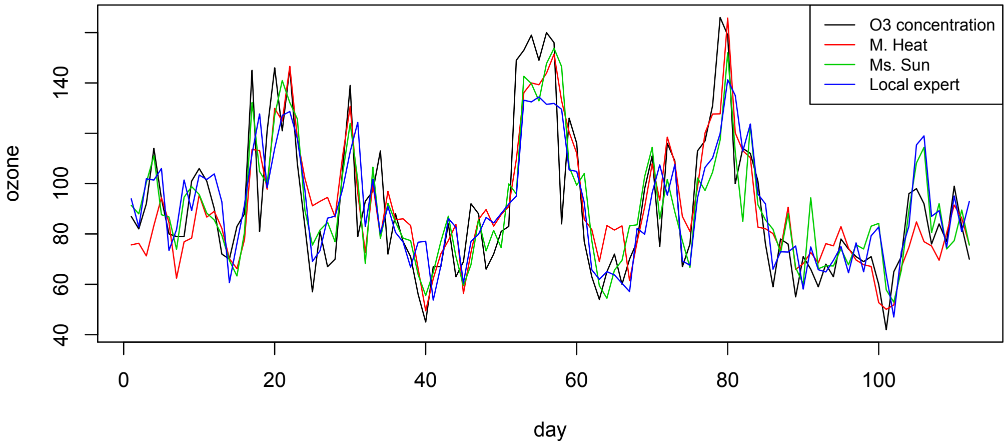


Figure 3.3: In black, the actual ozone concentration. In red the prediction of mister Heat, in green the prediction of misses Sun and in blue the prediction of the local expert.

We consider the aggregation strategies EW1 defined by (3.7) and EW2 defined by (3.13). For each expert, and each aggregation strategy, we compute the sum of the residual square errors (RMSE)

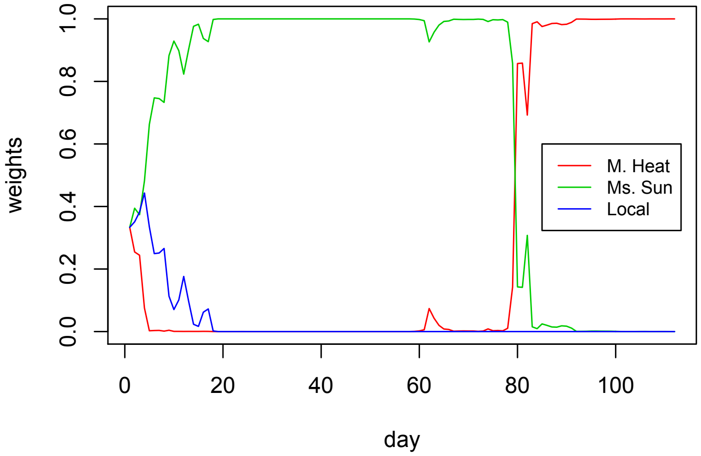
$$\text{RMSE} = \sum_t (\text{predict}(t) - \text{O3}(t))^2,$$

where predict(t) is the prediction for the day t and O3(t) is the actual observation. We obtain the following RMSE.

	M. Heat	Ms. Sun	Local expert	EW1	EW2
RMSE	26348	29013	35930	26577	23381

We observe that, in agreement with the theory, the RMSE of EW1 is almost as good as the RMSE of the best expert (M. Heat), while the RMSE of EW2 is smaller than that of all the experts. This last observation highlights the interest of taking convex combinations of expert predictions, instead of simply selecting one of them.

In the Figure 3.4, we display the weights $(\theta_t(1), \theta_t(2), \theta_t(3))$ of each expert in the aggregations EW1 and EW2. We observe that the weights in EW1 are mostly 0 or 1, with an abrupt change around day 80. The weights in EW2 are more evenly spread, and the local expert is not fully discarded. Keeping advices of all experts seems to be the recipe of the success!



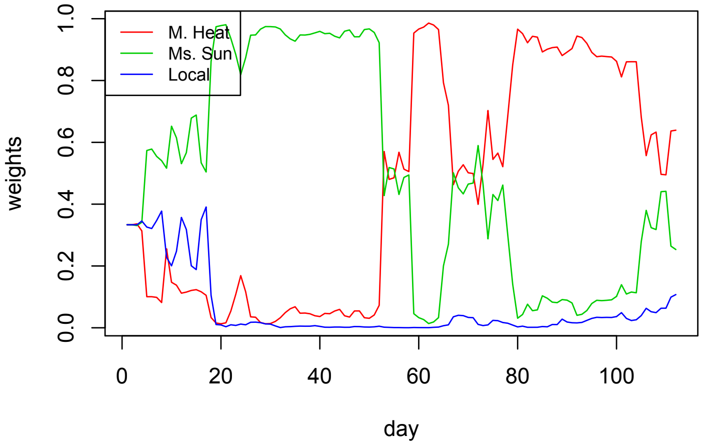


Figure 3.4: Weights $(\theta_t(1), \theta_t(2), \theta_t(3))$ of each expert in the predictions EW1 (top) and EW2 (bottom).

Chapter 4

Multi-Armed bandits

4.1 Setting

4.1.1 Bandits problems

Bandits problems correspond to problems where, at each time $t = 1, 2, \ldots$,

- the learner has to take a decision or choose an action,
- he then receives a reward for his action,
- the only information available at time t for choosing the action, are the rewards collected at the precedent rounds.

Figure 4.1: A one-armed bandit.

Examples:

- **Medical trials. (initial motivation)** A doctor face a new severe epidemic (coronavirus?). She can prescribe different drugs, or the same drug but at different doses to her patients. She does not know which drug or dose is the best. Her goal is to maximize the number of recoveries among her patients. She then faces the following issue: she wants to give as often as possible the best drug or dose in order to get a maximum of recoveries, but this best drug or dose is unknown, so she needs to prescribe the different drugs or doses in order to gain knowledge about the efficiency of the drugs or doses.
- **Recommendation - advertisement. (prominent applications)** Many websites display advertisements or recommendations. A display is successful, if the visitor clicks on the advertisement or recommendation. The system then tries to select among the huge number of possible advertisements or recommendations, the ones leading to the largest number of clicks. Yet, these best advertisements or recommendations are unknown, and the system has to simultaneously learn which ones are the best, and display them as often as possible.
- **Robot or algorithm control. (rising applications)** A robot (or algorithm) can have a rigid program in order to execute a task, or he can have a program which learns from the past the best

strategies for executing complex tasks. The program tries to get as many successes as possible (task completed). The goal is then to simultaneously learn the best strategies and apply them as often as possible. Examples of application include computer game players, air-conditioning regulation in data-centers, or optimisation of networking protocols.

- **A gambler in a casino. (where the name comes from!)** Let us consider the following toy example. A gambler arrives in a casino. He has access to different slot machines[1], each of them having their own mean payoff. The gambler wants to get the largest possible cumulated payoff, so he wants to focus on the slot machine having the largest mean payoff. Yet, the payoffs are unknown, so the gambler has to simultaneously learn the mean payoffs and play as often as possible the best one(s).

In all the above problems, the learner or algorithm has to choose at each round $t = 1, 2, \dots,$ an action A_t among a set $\mathcal{A}$ of possible actions. He then receives a reward $Y(A_t) \in \mathbb{R}$ for this action. The choice A_t is only based on the past rewards $Y(A_1), \dots, Y(A_{t-1})$. The learner or algorithm has to deal with the following issue. He only has access to the outcomes of the past actions, so he needs to try out different actions in order to gain information about his environment. At the same time, he tries to apply as often as possible (one of) the best action. He then faces an *exploration - exploitation* dilemma which is typical in learning problems with unknown environment. At time t, shall he explore the environment to get a better knowledge? or shall he exploit the action which gave the best rewards so far?

4.1.2 Modeling

In most of this chapter, we will focus on the case where there is a finite number K of possible actions. We refer to Section 4.4 for a case with an infinite number of possible actions. In reference to one-armed bandits, the set of possible actions is called the set of *arms* in the bandit literature.

Let us formalize the K-armed bandit problem. When the arm k is pulled (i.e. the action k is chosen) for the n^{th} time, we get a stochastic reward $X_k(n)$.

Rewards for the K arms. Each arm produces a sequence of rewards

- sample of arm 1: $X_1(1), X_1(2), \dots$
- ...
- sample of arm K: $X_K(1), X_K(2), \dots$

Observations. The rewards are observed only when an arm is pulled. At time t, if we choose the arm $A_t \in \{1, \dots, K\}$, we observe

$$Y_t = X_{A_t}(N_{A_t}(t)), \quad \text{where} \quad N_k(t) = \sum_{s=1}^{t} \mathbf{1}_{A_s=k}.$$

External randomness. The choice of the arm A_t may depend on some auxiliary sequence of random numbers $U(1), U(2), \dots \in [0, 1]$. For example, this can be useful to select the first arm at random.

Adaptive choices. The algorithm can adapt his choice A_t from past observations, but he cannot use future observations. In mathematical words, the choice A_t is $\sigma(U_1, Y_1, \dots, Y_{t-1}, U_t)$-mesurable.

Strategy. A strategy corresponds to the prescription of an algorithm, which will run autonomously as time passes. It can be encoded as a set of functions $\psi = (\psi_t)_{t\geq 1}$, with $\psi_t : \mathbb{R}^{2t-1} \to \{1, \dots, K\}$. At time $t = 1, 2, \dots$, the arm A_t is then pulled according to $A_t = \psi_t(U_1, Y_1, \dots, Y_{t-1}, U_t)$.

[1] Lever operated slot machine used to be called one-armed bandit

4.1.3 Regret

Cumulated reward. The cumulated reward collected up to time T is $\sum_{t=1}^{T} Y_t$. Our goal is to design a strategy maximizing the average of this cumulated reward.

Distributional assumption. In all the chapter, we make the following distributional assumption

- all the random variables $(X_k(n), U(n))_{n\geq 1}$ are jointly independent,
- the random variables $(X_k(n))_{n\geq 1}$ are i.i.d. with distribution ν_k, and mean $\mathbb{E}[X_k(n)] = \mu_k$ for $n \geq 1$.

Next lemma connects the expected cumulated reward to the means $\{\mu_k : k = 1, \dots, T\}$ and the expected number of draws of each arm. As a consequence, to assess the performance of strategy, we ”only” have to evaluate the expected numbers of draws for each arm.

Lemma 4.1 Expected cumulated reward.
Under the above distributional assumptions, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{k=1}^{K} \mu_k \mathbb{E}\left[N_k(T)\right].$$

Proof of Lemma 4.1. We have the decomposition

$$\sum_{t=1}^{T} Y_t = \sum_{k=1}^{K} \sum_{n=1}^{N_k(T)} X_k(n),$$

with the convention $\sum_{n=1}^{0} X_k(n) = 0$. Hence, we only need to prove the identity

$$\mathbb{E}\left[\sum_{n=1}^{N_k(T)} X_k(n)\right] = \mathbb{E}\left[N_k(T)\right] \mu_k.$$

Lemma 4.2 Wald formula.
Let $(\mathcal{G}_n)_{n\geq 0}$ be a filtration and let $N, X(1), X(2), \dots$ be random variables such that for all $n \geq 1$,

- *N takes value in $\{0, \dots, T\}$ and $\{N \geq n\} \in \mathcal{G}_{n-1}$;*
- *$X(n)$ is independent of $\mathcal{G}_{n-1}$ and $\mathbb{E}\left[X(n)\right] = \mu$.*

Then, we have

$$\mathbb{E}\left[\sum_{n=1}^{N} X(n)\right] = \mu \mathbb{E}\left[N\right]. \tag{4.1}$$

Proof of Wald formula. Since $\{N \geq n\} \in \mathcal{G}_{n-1}$ and $X(n)$ is independent of $\mathcal{G}_{n-1}$, we have $X(n)$ and $\{N \geq n\}$ independent. Hence

$$\mathbb{E}\left[\sum_{n=1}^{N} X(n)\right] = \mathbb{E}\left[\sum_{n=1}^{T} X(n)\mathbf{1}_{n\leq N}\right] = \sum_{n=1}^{T} \mathbb{E}\left[X(n)\mathbf{1}_{n\leq N}\right] = \sum_{n=1}^{T} \mu \mathbb{E}\left[\mathbf{1}_{n\leq N}\right] = \mu \mathbb{E}\left[\sum_{n=1}^{T} \mathbf{1}_{n\leq N}\right] = \mu \mathbb{E}\left[N\right].$$

The proof of Wald formula is complete. □

Let us set $\mathcal{G}_n = \sigma(X_k(1), \dots, X_k(n), (U(j))_{j\geq 1}, (X_\ell(j))_{j\geq 1,\, \ell\neq k})$. Then $\{N_k(t) \geq n\} \in \mathcal{G}_{n-1}$ and $X_k(n)$ is independent from $\mathcal{G}_{n-1}$ for all $n \geq 1$. So, applying Wald formula, we get the identity (4.1). □

Regret. As in the previous chapters, we will compare a strategy ψ to the best possible fixed strategy, i.e. to the strategy selecting the (unknown) arm with largest mean. Setting $\Delta_k = \mu_{k^*} - \mu_k$, where $\mu_{k^*} = \max_{j=1,\dots,K} \mu_j$, the regret at time T is

$$R(T) = R(\psi, T) = T\mu_{k^*} - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{k=1}^{K} \Delta_k \, \mathbb{E}\left[N_k(T)\right]. \tag{4.2}$$

4.2 UCB strategy

4.2.1 Optimism in face of uncertainty

Failure of a naive strategy based on empirical means

Let us write $\bar{X}_k(n) = (X_k(1) + \ldots + X_k(n))/n$ for the empirical mean of the rewards of the arm k after n pulling.
After time $t-1$, the average observed reward of arm k is $\bar{X}_k(N_k(t-1))$. A first idea that may come to your mind is to apply the following strategy: at time t, select the arm

$$A_t \in \underset{k=1,\dots,K}{\operatorname{argmax}} \bar{X}_k(N_k(t-1)).$$

This seems to be a good idea as it corresponds to the arm with the largest observed reward at time t.

Yet, this strategy can lead to very bad results. Actually, due to the random fluctuations, we may observe at some time t_0 a very small mean reward $\bar{X}_{k^*}(N_{k^*}(t_0))$ for the best arm compared to the expected reward μ_{k^*}, possibly much smaller than the expected reward μ_k of another arm. This can in particular happen in the early stages where the arm k^* has only been sampled a small amount of time. Then, if the observed mean reward $(\bar{X}_k(N_k(t)) : t \geq t_0)$ of the arm k does not deviate to much from below from the expected reward μ_k, the observed mean reward $\bar{X}_k(N_k(t))$ will always stay above the observed reward $\bar{X}_{k^*}(N_{k^*}(t_0))$ and hence the arm k^* will not be pulled anymore. Such a situation will lead to a cumulated regret linear in T.

Where confidence bounds kick in

A benefit of statistical inference is to provide some measures of uncertainty. In the above example, the naive strategy failed because the observed mean $\bar{X}_{k^*}(N_{k^*}(t_0))$ was not reliable and we trusted too much this value. Instead of focusing only on the observed mean, we shall consider instead confidence intervals.

Assume for example that $X_k(1), X_k(2), \ldots$ are i.i.d. with $\mathcal{N}(\mu_k, 1)$ distribution. Then, we have

$$\mathbb{P}\left[\mu_k \in \left[\bar{X}_k(n) - \sqrt{2L/n}\ ,\ \bar{X}_k(n) + \sqrt{2L/n}\right]\right] \geq 1 - 2e^{-L}.$$

In addition to the value $\bar{X}_k(n)$, the above confidence interval provides a measure $\sqrt{2L/n}$ of the uncertainty of this value as an estimator of μ_k. This measure is useful to know how much we can trust the value $\bar{X}_k(n)$. In particular, this measure informs the algorithm not to trust too much the empirical mean for small sample sizes n. The question is: How can we use efficiently this information?

A popular recipe for this problem is the ”optimism in face of uncertainty”. This recipe states that ”you should consider each action as being as good as it can possibly be given the observations so far, and choose the best action according to this assessment”. In our bandit problem, it means that we should consider each arm to be as good as the upper confidence bound $\bar{X}_k(N_k(t-1)) + \sqrt{2L/N_k(t-1)}$ of the confidence interval, and hence choose the arm with the largest upper confidence bound $\bar{X}_k(N_k(t-1)) + \sqrt{2L/N_k(t-1)}$. Why does such a strategy make sense?

We observe that the largest upper confidence bound $\bar{X}_k(N_k(t-1)) + \sqrt{2L/N_k(t-1)}$ can be large for one of the two following reasons. Either the expected reward μ_k is large, so it is a good reason to pull it (exploitation). Or the uncertainty $\sqrt{2L/N_k(t-1)}$ is large, because the arm has not been explored much. Again, it is worth to pull it in order to reduce the uncertainty (exploration). This principle provides a simple way to trade-off between exploration and exploitation.

4.2.2 Fixed time horizon UCB

In this section, we analyse the UCB algorithm. Our goal is to get a simple understanding of UCB. In particular, the constants in the theorem below can be improved with a more delicate analysis.

The UCB algorithm can be described in a general form as follows.

Fixed horizon UCB: Let $(\delta_T(n))_{n\geq 1}$ be a positive sequence decreasing in n.

- initialization: sample the K arms ones;
- iterations: for $t = K+1, \ldots, T$, take $A_t \in \text{argmax}_{k=1,\ldots,K} U_k(t)$, where

$$U_k(t) = \bar{X}_k(N_k(t-1)) + \delta_T(N_k(t-1)).$$

For any positive sequence $(\delta_T(n))_{n\geq 1}$ decreasing in n, we have the following upper-bound on the regret of Fixed horizon UCB.

Theorem 4.3 *For $T \geq K+1$, we set*

$$\Omega_{k,T} = \left\{ \max_{1\leq n\leq T} \left(\bar{X}_k(n) - \delta_T(n)\right) \leq \mu_k \right\} \cap \left\{ \mu_{k^*} < \min_{1\leq n\leq T} \left(\bar{X}_{k^*}(n) + \delta_T(n)\right) \right\}.$$

Then

$$R(T) \leq \sum_{k\neq k^*} \Delta_k \left(T\mathbb{P}\left(\Omega^c_{k,T}\right) + \delta_T^{-1}(\Delta_k/2) \right), \tag{4.3}$$

where $\delta_T^{-1}(x) = \min\{n \geq 1 : \delta_T(n) \leq x\}$.

Discussion. Before proving this result, let us comment on it. The bound (4.3) gives us some directions for the choice of the sequence $(\delta_T(n))_{n\geq 1}$. This choice must balance the size of the two terms $T\mathbb{P}(\Omega^c_{k,T})$ and $\delta_T^{-1}(\Delta_k/2)$. So $\delta_T(n)$ must be large enough, so that $\mathbb{P}(\Omega^c_{k,T}) = O(1/T)$, but not too large in order to keep $\delta_T^{-1}(\Delta_k/2)$ under control. So, the best is to select $\delta_T(n)$ as small as possible such that the probability of the event $\Omega^c_{k,T}$ is at most $O(1/T)$. We refer to Corollary 4.5 below for such an example.

Proof of Theorem 4.3. The theorem directly follows from (4.2), the trivial bound $N_k(T) \leq T$ and the next lemma.

Lemma 4.4 *On the event $\Omega_{k,T}$, we have $N_k(T) \leq \delta_T^{-1}(\Delta_k/2)$.*

Proof of Lemma 4.4. Assume that at some time $t \in [K+1, T]$ we have $\delta_T(N_k(t)) \leq \Delta_k/2$. Then, on the event $\Omega_{k,T}$ we have for all $t' \in [t, T]$

$$\bar{X}_k(N_k(t)) + \delta_T(N_k(t)) \leq \mu_k + 2\delta_T(N_k(t)) \leq \mu_{k^*} < \bar{X}_{k^*}(N_{k^*}(t')) + \delta_T(N_{k^*}(t')).$$

This sequence of inequalities implies that the arm t is not pulled anymore up to time T, so $N_k(T)$ cannot become larger than $1 \vee \delta_T^{-1}(\Delta_k/2) = \delta_T^{-1}(\Delta_k/2)$. □

Corollary 4.5 *Assume that the distributions of the rewards of each arm k are in $subG(\mu_k, \sigma^2)$. Setting*

$$\delta_T(n) = \sqrt{4\sigma^2 \log(T)/n},$$

we get

$$R(T) \leq \sum_{k \neq k^*} \left(3\Delta_k + \frac{16\sigma^2 \log(T)}{\Delta_k} \right). \tag{4.4}$$

Proof of Corollary 4.5. We have for all $k, n \geq 1$

$$\mathbb{P}\left(\bar{X}_k(n) \geq \mu_k + \delta_T(n)\right) \leq e^{-n\delta_T^2(n)/2\sigma^2} = e^{-2\log T} = \frac{1}{T^2}.$$

Hence, with a union bound

$$\mathbb{P}\left(\Omega_{k,T}^c\right) \leq 2 \sum_{n=1}^{T} \frac{1}{T^2} = \frac{2}{T}.$$

In addition, $\delta_T(n) \leq \Delta_k/2$ if and only if $n \geq 16\sigma^2 \log(T)/\Delta_k^2$, so

$$\delta_T^{-1}(\Delta_k/2) \leq 1 + \frac{16\sigma^2 \log(T)}{\Delta_k^2}.$$

4.2.3 Horizon free UCB

In the above section, the algorithm has the undesirable feature to depend on the time horizon T via δ_T. Is it possible to have an horizon free algorithm?
With a refinement of the proof of the previous theorem, we show in this section that we can replace δ_T by δ_t at time t.

Let $(\delta_t(n))_{t \geq 1, n \geq 1}$ be a positive sequence non-decreasing in t and decreasing in n.

Horizon-free UCB:

- initialization: sample the K arms ones;
- iterations: for $t \geq K + 1$, take $A_t \in \text{argmax}_{k=1,\ldots,K} U_k(t)$, where

$$U_k(t) = \bar{X}_k(N_k(t-1)) + \delta_t(N_k(t-1)).$$

We have the following upper-bound on the regret of horizon-free UCB.

Theorem 4.6 *Let $(T_\ell)_{\ell \geq 0}$ be an increasing sequence of integers, with $T_0 = 0$ and set*

$$\Omega_{k,\ell} = \left\{ \max_{1 \leq n \leq T_{\ell+1}} \left(\bar{X}_k(n) - \delta_{T_{\ell+1}}(n) \right) \leq \mu_k \right\} \cap \left\{ \mu_{k^*} < \min_{1 \leq n \leq T_{\ell+1}} \left(\bar{X}_{k^*}(n) + \delta_{T_\ell}(n) \right) \right\}.$$

Then for any $T \in [T_L + 1, T_{L+1}]$ we have

$$R(T) \leq \sum_{k \neq k^*} \sum_{\ell=0}^{L} \Delta_k \left((T_{\ell+1} - T_\ell) \mathbb{P}\left(\Omega_{k,\ell}^c\right) + \delta_{T_{\ell+1}}^{-1}(\Delta_k/2) \right).$$

We observe first that for any $T \in [T_L + 1, T_{L+1}]$ we have

$$R(T) \leq \sum_{k=1}^{K} \Delta_k \sum_{\ell=0}^{L} \mathbb{E}\left[N_k(T_{\ell+1}) - N_k(T_\ell)\right].$$

Hence the theorem follows from next lemma.

Lemma 4.7 *On the event $\Omega_{k,\ell}$, we have $N_k(T_{\ell+1}) - N_k(T_\ell) \leq \delta^{-1}_{T_{\ell+1}}(\Delta_k/2)$.*

Proof of Lemma 4.7. Assume that at some time $t \in [T_\ell, T_{\ell+1}]$ we have $\delta_{T_{\ell+1}}(N_k(t)) \leq \Delta_k/2$. Then, on the event $\Omega_{k,\ell}$ we have for all $t' \in [t, T_{\ell+1}]$

$$\begin{aligned}\bar{X}_k(N_k(t)) + \delta_{t'}(N_k(t)) &\leq \bar{X}_k(N_k(t)) + \delta_{T_{\ell+1}}(N_k(t)) \leq \mu_k + 2\delta_{T_{\ell+1}}(N_k(t))\\ &\leq \mu_{k^*} < \bar{X}_{k^*}(N_{k^*}(t')) + \delta_{T_\ell}(N_{k^*}(t')) \leq \bar{X}_{k^*}(N_{k^*}(t')) + \delta_{t'}(N_{k^*}(t')).\end{aligned}$$

This sequence of inequalities implies that the arm t is not pulled anymore up to time $T_{\ell+1}$, so $N_k(T_{\ell+1})$ cannot become larger than $N_k(T_\ell) \vee \delta^{-1}_{T_{\ell+1}}(\Delta_k/2)$. The conclusion follows. □

Corollary 4.8 *Assume that the distributions of the rewards of each arm k are in $subG(\mu_k, \sigma^2)$. Setting*

$$\delta_t(n) = \sqrt{8\sigma^2 \log(t)/n}$$

we have

$$R(T) \leq \sum_{k:k\neq k^*} \left(\Delta_k(6 + 3\log_2\log_2(T)) + \frac{128\sigma^2\log(T)}{\Delta_k}\right).$$

Proof of the corollary. We set $T_\ell = 2^{2^{\ell-1}}$ for $\ell \geq 1$. Then

$$\mathbb{P}\left(\Omega^c_{k,\ell}\right) \leq 2T_{\ell+1}e^{-8\log(T_\ell)/2} = \frac{2T_{\ell+1}}{T_\ell^4} = \frac{2}{T_{\ell+1}}.$$

As $\delta_t(n) \leq \Delta_k/2$ iff $n \geq 32\sigma^2\log(t)/\Delta_k^2$, we get that

$$\delta^{-1}_{T_{\ell+1}}(\Delta_k/2) \leq 1 + \frac{32\sigma^2\log(T_{\ell+1})}{\Delta_k^2}.$$

To conclude, we observe that

$$\sum_{\ell=0}^{L}\log(T_{\ell+1}) = \sum_{\ell=1}^{L+1} 2^{\ell-1}\log(2) \leq 2^{L+1}\log(2) = 4\log(T_L) \leq 4\log(T),$$

and $L = 1 + \log_2\log_2(T_L) \leq 1 + \log_2\log_2(T)$. □

4.3 Lower bounds

4.3.1 Kullback-Leibler divergence

Kullback-Leibler divergence. Let $\mathbb{P}, \mathbb{Q}$ be two probability distributions defined on a common measurable space and fulfilling $\mathbb{P} \ll \mathbb{Q}$. The KL-divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$KL(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right], \tag{4.5}$$

with the convention $0\log(0) = 0$. By convention, we set $KL(\mathbb{P}, \mathbb{Q}) = +\infty$ when $\mathbb{P}$ is not dominated by $\mathbb{Q}$.

The Kullback-Leibler divergence can also be written as

$$KL(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left[\log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right],$$

with the implicit convention that $\log(0) = 0$.

Exercise. Positivity.
Let $\varphi : [0, +\infty)$ be defined by $\varphi(x) = x\log(x)$ for $x > 0$, and $\varphi(0) = 0$. Check that φ is convex and conclude that $KL(\mathbb{P},\mathbb{Q}) \geq 0$.

Examples.

1. Let $\mathcal{B}(p)$ and $\mathcal{B}(q)$ denote two Bernoulli distributions with parameters p and q in $(0, 1)$. Then,

$$kl(p,q) := KL(\mathcal{B}(p),\mathcal{B}(q)) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

2. The Kullback-Leibler divergence between two Gaussian distributions $\mathcal{N}(\mu, 1)$ and $\mathcal{N}(\mu', 1)$ is

$$KL\left(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\mu',\sigma^2)\right) = \frac{1}{2\sigma^2}(\mu-\mu')^2.$$

Exercise. Pinsker inequality.
Adapt the proof of Lemma 3.9 to prove that for two probability distributions $\mathbb{P}$, $\mathbb{Q}$, with $\mathbb{P} \ll \mathbb{Q}$, we have

$$|\mathbb{P}-\mathbb{Q}|_1 := \mathbb{E}_{\mathbb{Q}}\left[\left|\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right|\right] \leq \sqrt{2\,KL(\mathbb{P},\mathbb{Q})}.$$

Check also the following additive property.

Exercise. Tensorization of the Kullback-Leibler divergence.
For four probability distributions $\mathbb{P}_1, \mathbb{P}_2, \mathbb{Q}_1, \mathbb{Q}_2$, with $\mathbb{P}_1 \ll \mathbb{Q}_1$ and $\mathbb{P}_2 \ll \mathbb{Q}_2$, we have

$$KL(\mathbb{P}_1\otimes\mathbb{P}_2, \mathbb{Q}_1\otimes\mathbb{Q}_2) = KL(\mathbb{P}_1,\mathbb{Q}_1) + KL(\mathbb{P}_2,\mathbb{Q}_2).$$

Next proposition provides an upper-bound on the difference between the expectations of a random variable under two different probability distributions in terms of the Kullback-Leibler divergence. It will be handy for analyzing the minimal regret in multi-armed bandit problems.

Proposition 4.9 Processing inequality.
Let Z be a random variable taking values in $[0, 1]$, and $\mathbb{P} \ll \mathbb{Q}$. Then

$$kl(\mathbb{E}_{\mathbb{P}}[Z], \mathbb{E}_{\mathbb{Q}}[Z]) \leq KL(\mathbb{P},\mathbb{Q}).$$

Proof of Proposition 4.9. We first prove that for any $\mathbb{P} \ll \mathbb{Q}$ and any event A, we have

$$kl(\mathbb{P}(A),\mathbb{Q}(A)) \leq KL(\mathbb{P},\mathbb{Q}), \tag{4.6}$$

with $kl(0, 0) = 0 = kl(1, 1)$.

We first observe that if $\mathbb{Q}(A) = 0$ (resp. $\mathbb{Q}(A^c) = 0$) then $\mathbb{P}(A) = 0$ (resp. $\mathbb{P}(A^c) = 0$), hence (4.6) trivially holds for $\mathbb{Q}(A) \in \{0, 1\}$. Hence, we assume below that $\mathbb{Q}(A) \in (0, 1)$.

The function $\varphi : [0, +\infty)$, defined by $\varphi(x) = x\log(x)$ for $x > 0$, and $\varphi(0) = 0$, is convex. From Jensen inequality, we get

$$\begin{aligned} KL(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right] &= \mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\Big|A\right]\mathbb{Q}(A) + \mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\Big|A^c\right]\mathbb{Q}(A^c) \\ &\geq \varphi\left(\mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\Big|A\right]\right)\mathbb{Q}(A) + \varphi\left(\mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\Big|A^c\right]\right)\mathbb{Q}(A^c) \end{aligned}$$

We have

$$\mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\middle|A\right] = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\mathbf{1}_A\right]/\mathbb{Q}(A) = \mathbb{P}(A)/\mathbb{Q}(A),$$

so we get

$$KL(\mathbb{P},\mathbb{Q}) \geq \mathbb{P}(A)\log(\mathbb{P}(A)/\mathbb{Q}(A)) + \mathbb{P}(A^c)\log(\mathbb{P}(A^c)/\mathbb{Q}(A^c)) = kl(\mathbb{P}(A),\mathbb{Q}(A)),$$

which proves (4.6).

To prove Proposition 4.9, we apply (4.6) to the event $A = \{(w,u) : u < Z(w)\} \subset \Omega \times [0,1]$, and the probability distributions $\mathbb{P}\otimes\lambda$ and $\mathbb{Q}\otimes\lambda$, where λ is the Lebesgue measure on $[0,1]$

$$kl\big((\mathbb{P}\otimes\lambda)(A), (\mathbb{Q}\otimes\lambda)(A)\big) \leq KL(\mathbb{P}\otimes\lambda, \mathbb{Q}\otimes\lambda).$$

To conclude, we notice that $KL(\mathbb{P}\otimes\lambda,\mathbb{Q}\otimes\lambda) = KL(\mathbb{P},\mathbb{Q}) + KL(\lambda,\lambda) = KL(\mathbb{P},\mathbb{Q})$ and

$$(\mathbb{P}\otimes\lambda)(A) = \int_{w\in\Omega} d\mathbb{P}(w)\int_{u=0}^{1} du\,\mathbf{1}_{Z(w)>u} = \mathbb{E}_{\mathbb{P}}[Z].$$

The proof of Proposition 4.9 is complete. □

Notation. For a set of distributions $\nu = (\nu_1,\ldots,\nu_K)$ for the rewards, we define $\mathbb{P}_\nu$ as the joint distribution of $\big((X_k(n))_{k=1,\ldots,K,\ n\geq 1}, (U_n)_{n\geq 1}\big)$. To a policy ψ, we associate the random variable $I_t = I_t(\psi) = (U_1, Y_1, \ldots, Y_{t-1}, U_t)$ defined as in the introduction. We remind the reader that the arm sampled at step t is $A_t = \psi_t(I_t)$. We denote by $\mathbb{P}_\nu^{I_t}$ the distribution of I_t under $\mathbb{P}_\nu$.

Next lemma provides a useful decomposition of the Kullback-Leibler divergence between the two distributions $\mathbb{P}_\nu^{I_t}$ and $\mathbb{P}_{\nu'}^{I_t}$ associated to two sets of distributions $(\nu_1,\ldots,\nu_K)$ and $(\nu'_1,\ldots,\nu'_K)$ for the rewards.

Lemma 4.10 *For two sets of distributions $\nu = (\nu_1,\ldots,\nu_K)$ and $\nu' = (\nu'_1,\ldots,\nu'_K)$ for the rewards, we have*

$$KL(\mathbb{P}_\nu^{I_{T+1}}, \mathbb{P}_{\nu'}^{I_{T+1}}) = \sum_{k=1}^{K} KL(\nu_k,\nu'_k)\mathbb{E}_\nu[N_k(T)].$$

Proof of Lemma 4.10.
To start with, we observe that the random variable U_{T+1} is independent of (I_T, Y_T) and follows a uniform distribution on $[0,1]$. Hence, $\mathbb{P}_\nu^{I_{T+1}} = \mathbb{P}_\nu^{(I_T,Y_T)}\otimes\lambda$, with λ the Lebesgue measure on $[0,1]$, and

$$KL(\mathbb{P}_\nu^{I_{T+1}}, \mathbb{P}_{\nu'}^{I_{T+1}}) = KL(\mathbb{P}_\nu^{(I_T,Y_T)}, \mathbb{P}_{\nu'}^{(I_T,Y_T)}) + KL(\lambda,\lambda) = KL(\mathbb{P}_\nu^{(I_T,Y_T)}, \mathbb{P}_{\nu'}^{(I_T,Y_T)}).$$

In addition, we decompose the distribution

$$d\mathbb{P}_\nu^{(I_T,Y_T)}(i_T,y_T) = d\mathbb{P}_\nu^{I_T}(i_T)\, d\mathbb{P}_\nu^{Y_T}(y_T|I_T = i_T).$$

Hence, we get

$$\log\left(\frac{d\mathbb{P}_\nu^{(I_T,Y_T)}(i_T,y_T)}{d\mathbb{P}_{\nu'}^{(I_T,Y_T)}(i_T,y_T)}\right) = \log\left(\frac{d\mathbb{P}_\nu^{I_T}(i_T)}{d\mathbb{P}_{\nu'}^{I_T}(i_T)}\right) + \log\left(\frac{d\mathbb{P}_\nu^{Y_T}(y_T|I_T=i_T)}{d\mathbb{P}_{\nu'}^{Y_T}(y_T|I_T=i_T)}\right). \tag{4.7}$$

Integrating the first term in the right-hand side of (4.7), we get

$$\begin{aligned}\int_{i_T}\int_{y_T}\log\left(\frac{d\mathbb{P}_\nu^{I_T}(i_T)}{d\mathbb{P}_{\nu'}^{I_T}(i_T)}\right) d\mathbb{P}_\nu^{(I_T,Y_T)}(i_T,y_T) &= \int_{i_T}\int_{y_T}\log\left(\frac{d\mathbb{P}_\nu^{I_T}(i_T)}{d\mathbb{P}_{\nu'}^{I_T}(i_T)}\right) d\mathbb{P}_\nu^{I_T}(i_T)\, d\mathbb{P}_\nu^{Y_T}(y_T|I_T=i_T)\\ &= \int_{i_T}\log\left(\frac{d\mathbb{P}_\nu^{I_T}(i_T)}{d\mathbb{P}_{\nu'}^{I_T}(i_T)}\right) d\mathbb{P}_\nu^{I_T}(i_T) = KL(\mathbb{P}_\nu^{I_T}, \mathbb{P}_{\nu'}^{I_T}). \qquad (4.8)\end{aligned}$$

As for the second term of (4.7), we observe that under $\mathbb{P}_\nu$ the conditional distribution of Y_T given $I_T = i_T$ is $\nu_{\psi_T(i_T)}$. Hence $d\mathbb{P}_\nu^{Y_T}(y_T|I_T = i_T) = d\nu_{\psi_T(i_T)}(y_T)$. So

$$\begin{aligned}
&\int_{i_T}\int_{y_T} \log\left(\frac{d\mathbb{P}_\nu^{Y_T}(y_T|I_T = i_T)}{d\mathbb{P}_{\nu'}^{Y_T}(y_T|I_T = i_T)}\right) d\mathbb{P}_\nu^{(I_T,Y_T)}(i_T, y_T)\\
&= \int_{i_T}\int_{y_T} \log\left(\frac{d\nu_{\psi_T(i_T)}(y_T)}{d\nu'_{\psi_T(i_T)}(y_T)}\right) d\mathbb{P}_\nu^{I_T}(i_T)\; d\nu_{\psi_T(i_T)}(y_T)\\
&= \int_{i_T} d\mathbb{P}_\nu^{I_T}(i_T) \sum_{k=1}^K \mathbf{1}_{\psi_T(i_T)=k} \underbrace{\int_{y_T} \log\left(\frac{d\nu_k(y_T)}{d\nu'_k(y_T)}\right) d\nu_k(y_T)}_{=KL(\nu_k,\nu'_k)}\\
&= \sum_{k=1}^K KL(\nu_k, \nu'_k)\, \mathbb{E}_\nu\left[\mathbf{1}_{A_T=k}\right], \qquad (4.9)
\end{aligned}$$

where $A_T = \psi_T(I_T)$ is the arm pulled at stage T. Hence combining (4.8) and (4.9), we get by induction that

$$KL(\mathbb{P}_\nu^{I_{T+1}}, \mathbb{P}_{\nu'}^{I_{T+1}}) = KL(\mathbb{P}_\nu^{I_T}, \mathbb{P}_{\nu'}^{I_T}) + \sum_{k=1}^K KL(\nu_k, \nu'_k)\, \mathbb{E}_\nu\left[\mathbf{1}_{A_T=k}\right] = \ldots = \sum_{k=1}^K KL(\nu_k, \nu'_k)\, \mathbb{E}_\nu\left[N_k(T)\right].$$

The proof of Lemma 4.10 is complete. □

We have the following immediate corollary of Proposition 4.9 and Lemma 4.10.

Corollary 4.11 *Let Z be a $\sigma(I_{T+1})$-mesurable random variable taking values in $[0,1]$, and let $\nu = (\nu_1, \ldots, \nu_K)$ and $\nu' = (\nu'_1, \ldots, \nu'_K)$ be two set of distributions for the rewards. Then*

$$kl(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]) \leq \sum_{k=1}^K KL(\nu_k, \nu'_k)\mathbb{E}_\nu\left[N_k(T)\right]. \qquad (4.10)$$

This result will be useful in order to lower bound the expected number of pulling of an arm $\mathbb{E}_\nu[N_k(T)]$.

Proof of Corollary 4.11.
Since Z is $\sigma(I_{T+1})$-mesurable, we have $Z = F(I_{T+1})$ for some measurable $F : \mathbb{R}^{2T+1} \to [0,1]$. We can write the expectation of Z as follows

$$\mathbb{E}_\nu[Z] = \int_{\omega\in\Omega} F(I_{T+1}(\omega))\, d\mathbb{P}_\nu(\omega) = \int_{x\in\mathbb{R}^{2T+1}} F(x)\, d\mathbb{P}_\nu^{I_{T+1}}(x) = \mathbb{E}_{\mathbb{P}_\nu^{I_{T+1}}}[F].$$

Since F takes values in $[0,1]$, Proposition 4.9 and Lemma 4.10 then gives

$$\begin{aligned}
kl(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]) &= kl\Big(\mathbb{E}_{\mathbb{P}_\nu^{I_{T+1}}}[F], \mathbb{E}_{\mathbb{P}_{\nu'}^{I_{T+1}}}[F]\Big)\\
&\leq KL\big(\mathbb{P}_\nu^{I_{T+1}}, \mathbb{P}_{\nu'}^{I_{T+1}}\big) = \sum_{k=1}^K KL(\nu_k, \nu'_k)\mathbb{E}_\nu\left[N_k(T)\right].
\end{aligned}$$

The proof of Corollary 4.11 is complete. □

4.3.2 Asymptotic lower bounds

We are now ready to prove a lower bound for the best possible performance of a policy on a bandit problem with Gaussian rewards. We have seen that for Gaussian rewards $\nu_k = \mathcal{N}(\mu_k, \sigma^2)$, the UCB policy achieves a regret

$$R(\psi^{UCB}, T) = O\left(\sum_{k \neq k^*} \frac{\sigma^2 \log(T)}{\Delta_k}\right).$$

We will prove that, in some sense, no policy can have a better regret.

Warning. For $\nu = (\nu_1, \ldots, \nu_K)$, if the best arm is $k^*(\nu) = 3$, then the policy $\psi^{(3)}$ which only samples arm 3 is optimal and suffers a zero regret. Yet, the policy $\psi^{(3)}$ is very poor when $k^*(\nu) \neq 3$, in which case the regret is $\Delta_3 T$. This motivates the fact that we want policies ψ which are good on a whole class of problems.

Theorem 4.12 below shows that any policy with a $o(T^\alpha)$ regret on all bandit problems with Gaussian rewards, has a regret larger than the regret of UCB, up to a possible multiplicative constant. Let us formalize this result.

Definitions:

- Let $\mathcal{D}_\sigma = \left\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\right\}$ denote the set of Gaussian distributions with variance σ^2;
- Let $\Psi_{\alpha-fast}$ denote the set of policies ψ such that, for any $\nu_1, \ldots, \nu_K \in \mathcal{D}_\sigma$, and any $k \neq k^*(\nu)$, we have $\mathbb{E}_\nu\left[N_k(T)\right] = o(T^\alpha)$ as $T \to \infty$.

Theorem 4.12 *For any $\nu_1, \ldots, \nu_K \in \mathcal{D}_\sigma$, with $\nu_k = \mathcal{N}(\mu_k, \sigma^2)$, for $k = 1, \ldots, K$, for any $\alpha \in (0, 1)$, and for any policy $\psi \in \Psi_{\alpha-fast}$*

$$\liminf_{T \to +\infty} \frac{R(\psi, T)}{\log(T)} \geq (1 - \alpha) \sum_{k \neq k^*(\nu)} \frac{2\sigma^2}{\Delta_k}.$$

Let us comment briefly on this result. We observe that a policy that has a regret $o(T^\alpha)$ uniformly over $(\nu_1, \ldots, \nu_K) \in \mathcal{D}_\sigma^K$, cannot have a regret smaller than $O(\log(T))$ anywhere in $\mathcal{D}_\sigma^K$. In addition, the lower bound on the regret is proportional to

$$\sum_{k \neq k^*} \frac{\sigma^2 \log(T)}{\Delta_k},$$

which is the sum appearing in (4.4), up to constants. Hence, no policy can perform better (up to constants) than the UCB policy, uniformly over $(\nu_1, \ldots, \nu_K) \in \mathcal{D}_\sigma^K$.

Proof of Theorem 4.12.
According to (4.2), we only need to prove that

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\nu\left[N_k(T)\right]}{\log(T)} \geq \frac{2(1 - \alpha)\sigma^2}{\Delta_k^2}, \quad \text{for all } k \neq k^*(\nu), \tag{4.11}$$

where $k^*(\nu)$ is the best arm for the set of distribution ν.
Let us fix $k \neq k^*(\nu)$. The recipe for proving the lower bound is to compare, via Corollary 4.11, the performance of ψ on ν and on ν', where ν' is such that $k^*(\nu') = k$. We define the collection $(\nu'_1, \ldots, \nu'_K) \in \mathcal{D}_\sigma^K$ as follows. For $\delta > 0$, we set $\nu'_k = \mathcal{N}(\mu_{k^*} + \delta, \sigma^2)$ and for all $j \neq k$ we set $\nu'_j = \nu_j$. Hence, the difference between $(\nu'_1, \ldots, \nu'_K)$ and $(\nu_1, \ldots, \nu_K)$ is only at the distribution ν'_k.
According to (4.10) with $Z = N_k(T)/T$, we have

$$\mathbb{E}_\nu\left[N_k(T)\right] KL(\nu_k, \nu'_k) \geq kl\left(\frac{\mathbb{E}_\nu[N_k(T)]}{T}, \frac{\mathbb{E}_{\nu'}[N_k(T)]}{T}\right).$$

We also have

$$\begin{aligned} kl(p,q) &= (1-p)\log\left(\frac{1}{1-q}\right) + \underbrace{p\log(1/q)}_{\geq 0} + \underbrace{p\log(p) + (1-p)\log(1-p)}_{\geq -\log(2)} \\ &\geq (1-p)\log\left(\frac{1}{1-q}\right) - \log(2), \end{aligned}$$

so

$$\mathbb{E}_\nu\left[N_k(T)\right] KL(\nu_k, \nu_k') \geq \left(1 - \frac{\mathbb{E}_\nu[N_k(T)]}{T}\right)\log\left(\frac{T}{T - \mathbb{E}_{\nu'}[N_k(T)]}\right) - \log(2),$$

We observe that k is sub-optimal under $\mathbb{P}_\nu$, but it is the best arm under the distribution $\mathbb{P}_{\nu'}$. So, since $\psi \in \Psi_{\alpha-fast}$, we have $\mathbb{E}_\nu\left[N_k(T)\right] = o(T^\alpha)$ and $\mathbb{E}_{\nu'}\left[N_j(T)\right] = o(T^\alpha)$ for all $j \neq k$. Hence

$$\mathbb{E}_\nu\left[N_k(T)\right] = o(T^\alpha) \quad \text{and} \quad T - \mathbb{E}_{\nu'}\left[N_k(T)\right] = \sum_{j:j\neq k} \mathbb{E}_{\nu'}\left[N_j(T)\right] = o(T^\alpha).$$

It follows that

$$\mathbb{E}_\nu\left[N_k(T)\right] KL(\nu_k, \nu_k') \geq \left(1 - o(T^{-(1-\alpha)})\right)(1-\alpha)\log(T) - \log(2) - \log(o(1))$$

Direct computations give

$$KL(\nu_k, \nu_k') = KL(\mathcal{N}(\mu_k, \sigma^2), \mathcal{N}(\mu_{k^*} + \delta, \sigma^2)) = \frac{1}{2\sigma^2}\left(\mu_{k^*} + \delta - \mu_k\right)^2,$$

so taking $\delta = \varepsilon\Delta_k$, with $\varepsilon > 0$, we get

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu\left[N_k(T)\right]}{\log(T)} \geq \frac{2(1-\alpha)\sigma^2}{(1+\varepsilon)^2\Delta_k^2}.$$

Since this lower bound is valid for any $\varepsilon > 0$, the lower bound (4.11) is proved. The proof of Theorem 4.12 is complete. □

4.4 Problem: X-armed bandits

We consider now the case where there is an infinite number of arms, indexed by $x \in [0,1]$. We assume that the arm x produces rewards which are in $[0,1]$ with mean denoted by $\mu(x)$.

Without further assumptions, there is no hope to get a non-trivial regret bound, as we cannot even sample all arms ones. Hence, we will consider the case where $x \to \mu(x)$ is regular. More precisely, we assume that μ is (β, L)-Hölder,

$$|\mu(x) - \mu(y)| \leq L|x-y|^\beta : \quad \text{for all } x, y \in [0,1],$$

for some $L \in \mathbb{R}^+$ and $\beta \in (0,1]$.

As μ is regular, arms with close indices x and y have close mean rewards $\mu(x)$ and $\mu(y)$. Hence, an idea is to split $[0,1] = J_1 \cup \ldots \cup J_K$ into K intervals of length $1/K$ and then to cluster the arms accordingly into K groups. We can define then a K-arms bandit as follows: the arm k corresponds to sampling a value x chosen uniformly at random in J_k. Hence, the mean reward of the arm k is

$$m_k = K\int_{J_k} \mu(x)\,dx.$$

As the rewards of the arm k are in $[0,1]$, the distribution of the rewards of the arm k is in

$subG(m_k, 1/4)$. According to Corollary 4.5, the regret of fixed horizon UCB for this K-arms bandit problem is upper-bounded by

$$R_K(T) = T \max_{j=1,\dots,K} m_j - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] \leq \sum_{k \neq k^*} \left(3\Delta_k + \frac{4\log(T)}{\Delta_k}\right),$$

where $\Delta_k = \max_{j=1,\dots,K} m_j - m_k = m_{k^*} - m_k$.

In this problem, you will work out this bound in order to get a bound on the regret for the original problem

$$R^*(T) = T \max_{x \in [0,1]} \mu(x) - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right].$$

Theorem 4.13 *When μ is (β, L)-Hölder, for a suitable choice of K (depending on β), the algorithm described above fulfills the regret bound*

$$R^*(T) \leq C_{L,\beta}\, T^{\frac{\beta+1}{2\beta+1}} (\log T)^{\frac{\beta}{2\beta+1}},$$

for some constant $C_{L,\beta} > 0$ depending only on (L, β).

Proof of Theorem 4.13. Prove the theorem by solving the five next questions.

1. Prove that $\max_{x \in [0,1]} \mu(x) - \max_{j=1,\dots,K} m_j \leq LK^{-\beta}$.
2. Let us choose some $D > 0$. We can split the regret $R_K(T)$ into two pieces

   $$R_K(T) = \sum_{k:\Delta_k \leq D} \Delta_k \mathbb{E}\left[N_k(T)\right] + \sum_{k:\Delta_k > D} \Delta_k \mathbb{E}\left[N_k(T)\right].$$

   Check that the first sum can be simply upper bounded by DT.
3. Check that the second sum is upper bounded by

   $$\sum_{k:\Delta_k > D} \Delta_k \mathbb{E}\left[N_k(T)\right] \leq \frac{4K\log(T)}{D} + 3KL.$$

4. Putting pieces together, check that

   $$R^*(T) \leq LK^{-\beta}T + DT + \frac{4K\log(T)}{D} + 3KL.$$

5. Optimizing the value D and then partially optimizing the number K of blocks, conclude the proof of Theorem 4.13.

4.5 Illustration of UCB

Let us visualize the UCB algorithm on a simulated example. The R-code can be downloaded at `https://www.imo.universite-paris-saclay.fr/~giraud/Orsay/MathIA/Bandits.R`

We consider 4 arms following a Bernoulli distribution with means

$$\mu_1 = 0.1, \quad \mu_2 = 0.5, \quad \mu_3 = 0.3, \quad \mu_4 = 0.4.$$

We run UCB as in Corollary 4.5 for $T = 1000$ time steps. Here $\sigma^2 = 1/4$.

In Figure 4.2, we display the regret as time passes $t \longrightarrow R(t)$ and the arms sampled at each time step.

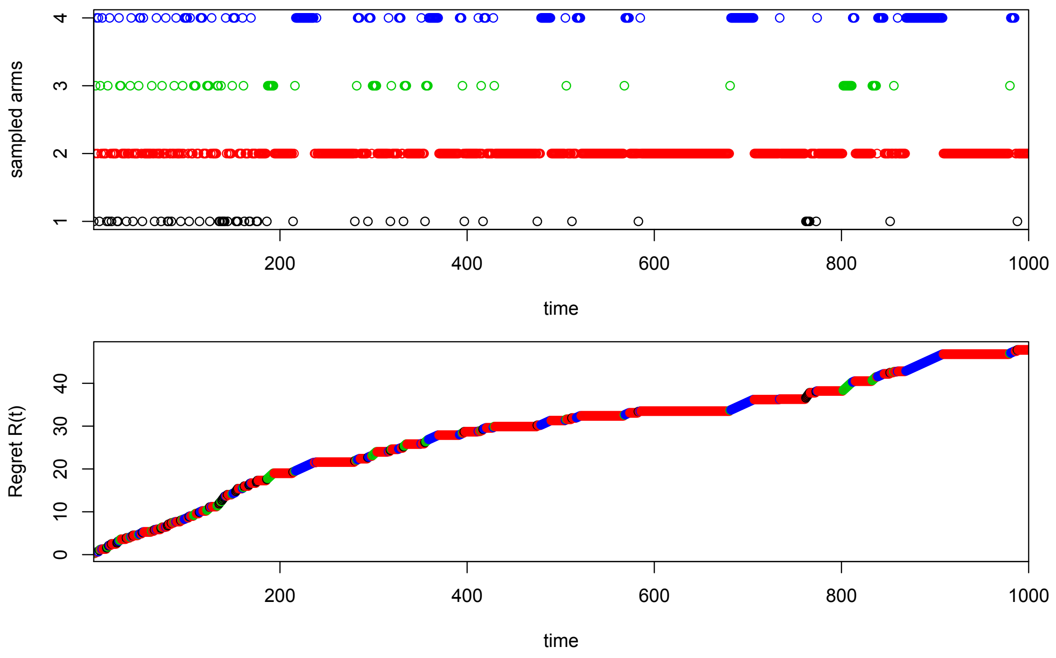


Figure 4.2: Top: Sampled arms when time passes. Arm1 in black, Arm 2 in red, Arm 3 in green, Arm 4 in blue. Below: regret as time passes. The color of the dot at time t corresponds to the color of the arm sampled A_t

In Figure 4.3, we display the Upper Confidence Bound $U_k(t)$ (triangles) and the empirical mean $\bar{X}_k(N_k(t-1))$ (crosses) for each arm $k = 1, \dots, 4$, at three time steps $t = 100, 300, 1000$. You can observe that the empirical means $\bar{X}_k(N_k(t-1))$ are slowly converging to the true means μ_k, and that the 4 upper confidence bounds $U_1(t), \dots, U_4(t)$ are almost at the same level at each time steps (why?).

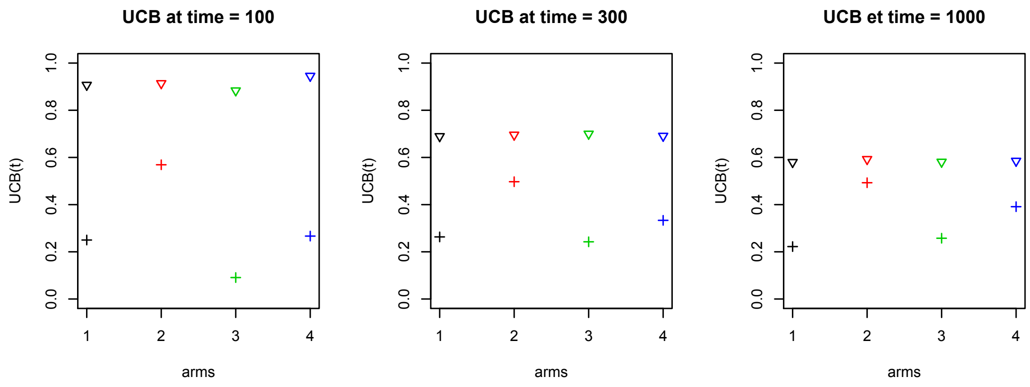


Figure 4.3: Upper confidence bounds $U_k(t)$ (triangles) and empirical means $\bar{X}_k(N_k(t-1))$ (crosses) for each arm $k = 1, \dots, 4$. Arm1 in black, Arm 2 in red, Arm 3 in green, Arm 4 in blue. Left: $t = 100$. Center: $t = 300$. Right: $t = 1000$.

Part II

Matrix analysis for Machine Learning

Chapter 5

Singular Value Decomposition

Linear algebra and matrix analysis play an important role in machine learning. They are involved in many different topics, including dimension reduction, clustering or regression.

5.1 Reminder on spectral decomposition of symmetric real matrices

Spectral decomposition of symmetric real matrices plays an important role for constructive computations.

Theorem 5.1 Spectral decomposition of symmetric real matrices.
Let $A \in \mathbb{R}^{n\times n}$ be a symmetric real matrix. Then, there exists $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_n \in \mathbb{R}$ and an orthonormal basis $\{u_1, \ldots, u_n\}$ of $\mathbb{R}^n$ such that

$$A = \sum_{k=1}^{n} \lambda_k u_k u_k^T.$$

The spectral decomposition can also be written $A = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^T$ with $U = [u_1 u_2 \cdots u_n]$.

Proof of Theorem 5.1. Let us give an analytic proof of the spectral decomposition. Let $F : \mathbb{R}^n \to \mathbb{R}$ be defined by $F(u) = \frac{1}{2}u^T Au$. As F is continuous and as the unit sphere $\partial B_{\mathbb{R}^n}(0,1)$ is compact, there exists at least one maximizer u_1 of F in $\partial B_{\mathbb{R}^n}(0,1)$

$$u_1 \in \underset{u_1 \in \partial B_{\mathbb{R}^n}(0,1)}{\mathrm{argmax}} \; F(u).$$

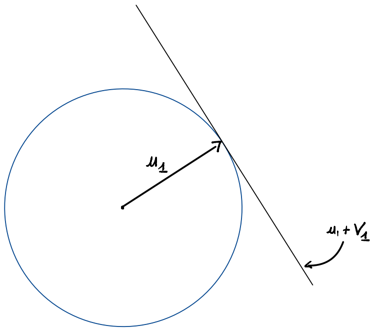


Figure 5.1: Tangent plane $u_1 + V_1$ to $\partial B_{\mathbb{R}^n}(0,1)$ in u_1.

The tangent plane to $\partial B_{\mathbb{R}^n}(0,1)$ in u_1 is $u_1 + V_1$, where $V_1 =< u_1 >^{\perp}$ with $< u_1 >$ the line spanned by u_1. Hence, as u_1 is a maximizer on the sphere, we have $\nabla F(u_1) \perp V_1$. So, there exists $\lambda_1 \in \mathbb{R}$ such that $\nabla F(u_1) = \lambda_1 u_1$. As $\nabla F(u_1) = Au_1$, we have $Au_1 = \lambda_1 u_1$.

We can decompose $\mathbb{R}^n =< u_1 > +V_1$. For any $u \in V_1$, we have

$$\langle Au, u_1\rangle = \langle u, A^T u_1\rangle = \langle u, Au_1\rangle = \lambda_1\langle u, u_1\rangle = 0.$$

So $AV_1 \subset V_1$. Hence, we can apply the same argument as above to $F_1 : V_1 \to \mathbb{R}$, $F_1(u) = u^T Au/2$, which gives $u_2 \in V_1 \cap \partial B_{\mathbb{R}^n}(0,1)$ such that $Au_2 = \lambda_2 u_2$ for some $\lambda_2 \in \mathbb{R}$.

We then get $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_n$ and the orthonormal family $\{u_1, \ldots, u_n\}$ of eigenvectors of A by induction. Finally, we observe that since any $x \in \mathbb{R}^n$ can be decomposed as $x = \sum_k (u_k^T x)\, u_k$, we have

$$Ax = \sum_{k=1}^{n} (u_k^T x)\, Au_k = \sum_{k=1}^{n} \lambda_k (u_k^T x)\, u_k = \sum_{k=1}^{n} \lambda_k u_k u_k^T x.$$

The proof of Theorem 5.1 is complete. □

Positive semi-definite matrices. We remind the reader that a symmetric real matrix A is positive semidefinite (p.s.d.) if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$. Since

$$x^T Ax = \sum_{k=1}^{n} \lambda_k \langle x, u_k\rangle^2,$$

a symmetric real matrix A is positive semi-definite if and only if $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_n \geq 0$.

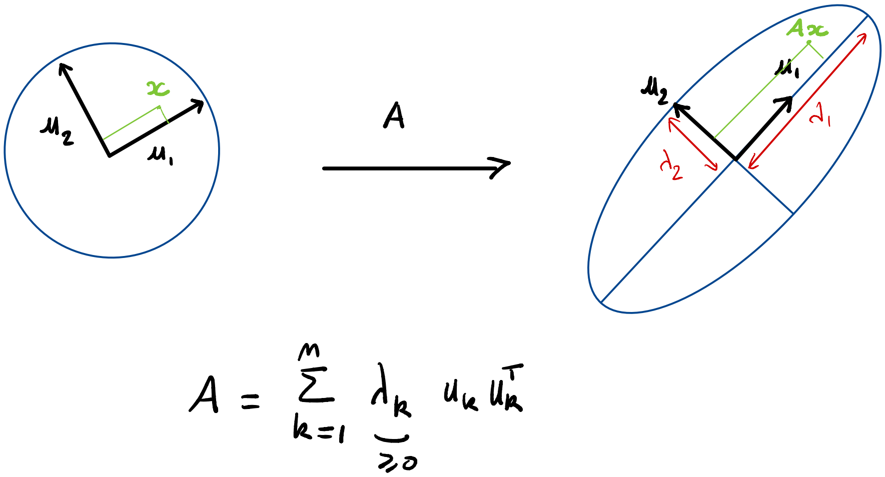

x
u_2
u_1
A
Ax
u_2
u_1
λ_1
λ_2
$A = \sum_{k=1}^{n} \underbrace{\lambda_k}_{\geq 0} u_k u_k^T$

Figure 5.2: Geometric representation of the spectral theorem for p.s.d matrices.

5.2 Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is a matrix decomposition that is very useful in many fields of applied mathematics. In the following, we will use that, for any $n \times p$ matrix A, the matrices $A^T A$ and AA^T are symmetric positive semidefinite.

Theorem 5.2 Singular value decomposition
Any $n \times p$ matrix A of rank r can be decomposed as

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T, \quad \textit{where} \tag{5.1}$$

- $r = \mathrm{rank}(A)$,
- $\sigma_1 \geq \ldots \geq \sigma_r > 0$,
- $\{\sigma_1^2, \ldots, \sigma_r^2\}$ *are the nonzero eigenvalues of* $A^T A$ *(they are also the nonzero eigenvalues of* AA^T*), and*
- $\{u_1, \ldots, u_r\}$ *and* $\{v_1, \ldots, v_r\}$ *are two orthonormal families of* $\mathbb{R}^n$ *and* $\mathbb{R}^p$*, such that*

$$AA^T u_j = \sigma_j^2 u_j \quad \textit{and} \quad A^T A v_j = \sigma_j^2 v_j.$$

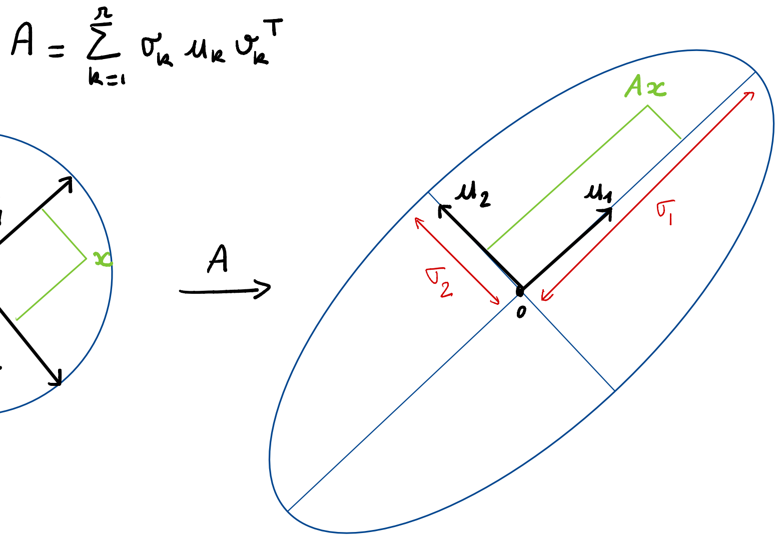


Figure 5.3: Geometric representation of the SVD. In blue, $x = 0.75 v_1 + 0.3 v_2$ is plotted on the left hand figure and $Ax = 0.75\sigma_1 u_1 + 0.3\sigma_2 u_2$ is plotted on the right hand side figure.

The values $\sigma_1, \ldots, \sigma_r$ are called the singular values of A. The vectors $\{u_1, \ldots, u_r\}$ and $\{v_1, \ldots, v_r\}$ are said to be left-singular vectors and right-singular vectors, respectively. The decomposition (5.1) is called a Singular Value Decomposition (SVD) of A.

Proof. Let us prove that such a decomposition exists. We remind first the following decomposition of $\mathbb{R}^p$.

Lemma 5.3 *For any matrix $A \in \mathbb{R}^{n \times p}$, we have the orthogonal decomposition*

$$\mathbb{R}^p = ker(A) \obot range(A^T).$$

Proof of Lemma 5.3. First, we observe that $\ker(A) \perp \mathrm{range}(A^T)$. Indeed, for any $y = A^T x \in$

range(A^T) and $x_0 \in \ker(A)$, we have

$$\langle x_0, y\rangle = \langle x_0, A^T x\rangle = \langle \underbrace{Ax_0}_{=0}, x\rangle = 0.$$

Since dim(range(A^T)) = rank(A^T) = rank(A) = p − dim(ker(A)), and since ker(A) ⊥ range(A^T), the conclusion follows. □

According to Lemma 5.3, the range of A and the range of AA^T coincide, so rank(AA^T) = rank(A) = r. Since AA^T is positive semidefinite with rank r, we have a spectral decomposition

$$AA^T = \sum_{j=1}^{r} \lambda_j u_j u_j^T,$$

with $\lambda_1 \geq \ldots \geq \lambda_r > 0$ and $\{u_1, \ldots, u_r\}$ an orthonormal family of $\mathbb{R}^n$. Let us define $v_1, \ldots, v_r$ by $v_j = \lambda_j^{-1/2} A^T u_j$ for $j = 1, \ldots, r$. We have

$$\langle v_i, v_j\rangle = \lambda_i^{-1/2}\lambda_j^{-1/2} u_i^T AA^T u_j = \lambda_i^{-1/2}\lambda_j^{1/2} u_i^T u_j = \delta_{i,j},$$

and

$$A^T A v_j = \lambda_j^{-1/2} A^T (AA^T) u_j = \lambda_j^{1/2} A^T u_j = \lambda_j v_j,$$

so $\{v_1, \ldots, v_r\}$ is an orthonormal family of eigenvectors of $A^T A$. Setting $\sigma_j = \lambda_j{}^{1/2}$, we obtain

$$\begin{aligned}\sum_{j=1}^{r} \sigma_j u_j v_j^T &= \sum_{j=1}^{r} \lambda_j^{1/2}\, \lambda_j^{-1/2} u_j u_j^T A \\ &= \left(\sum_{j=1}^{r} u_j u_j^T\right) A.\end{aligned}$$

Writing $P = \sum_{j=1}^{r} u_j u_j^T$, it remains to check that $PA = A$. We notice that P is the projection onto the range of AA^T. According to Lemma 5.3, the range of A and the range of AA^T coincide so P is also the projection onto the range of A. Hence

$$\sum_{j=1}^{r} \sigma_j u_j v_j^T = \left(\sum_{j=1}^{r} u_j u_j^T\right) A = PA = \mathrm{Proj}_{\mathrm{range}(A)}\, A = A.$$

The proof of Lemma 5.2 is complete. □

In the following, we denote by $\sigma_1(A) \geq \sigma_2(A) \geq \ldots$ the singular values of A.

Exercise: For any $n \times p$ matrix A, prove the equalities

$$\sigma_1(A) = \sup_{\|x\|=1} \|Ax\| = \sup_{\|x\|=1, \|y\|=1} \langle Ax, y\rangle.$$

The next result is a geometric characterization of the singular values.

Theorem 5.4 Min–Max / Max–Min formulas
For any $n \times p$ matrix A and $k \leq r = rank(A)$, we have

$$\sigma_k(A) = \max_{S:\dim(S)=k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|}, \tag{5.2}$$

where the maximum is taken over all the linear spans $S \subset \mathbb{R}^p$ with dimension k. Symmetrically, we have

$$\sigma_k(A) = \min_{S:codim(S)=k-1} \max_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|}, \tag{5.3}$$

where the minimum is taken over all the linear spans $S \subset \mathbb{R}^p$ with codimension $k-1$.

Proof. We start from the singular value decomposition $A = \sum_{j=1}^r \sigma_j(A) u_j v_j^T$ and we consider $\left\{v_{r+1}, \ldots, v_p\right\}$, such that $\left\{v_1, \ldots, v_p\right\}$ is an orthonormal basis of $\mathbb{R}^p$. We define $S_k = \mathrm{span}\{v_1, \ldots, v_k\}$ and $W_k = \mathrm{span}\left\{v_k, \ldots, v_p\right\}$. For any linear span $S \subset \mathbb{R}^p$ with dimension k, we have $\dim(S) + \dim(W_k) = p+1$, so $S \cap W_k \neq \{0\}$. For any nonzero $x \in S \cap W_k$ we have

$$\frac{\|Ax\|^2}{\|x\|^2} = \frac{\sum_{j=k}^r \sigma_j(A)^2 \langle v_j, x\rangle^2}{\sum_{j=k}^p \langle v_j, x\rangle^2} \leq \sigma_k(A)^2,$$

so

$$\max_{S:\dim(S)=k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|} \leq \sigma_k(A).$$

Conversely, for all $x \in S_k \setminus \{0\}$, we have

$$\frac{\|Ax\|^2}{\|x\|^2} = \frac{\sum_{j=1}^k \sigma_j(A)^2 \langle v_j, x\rangle^2}{\sum_{j=1}^k \langle v_j, x\rangle^2} \geq \sigma_k(A)^2,$$

with equality for $x = v_k$. As a consequence,

$$\max_{S:\dim(S)=k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sigma_k(A),$$

with equality for $S = S_k$, which proves (5.2). The min–max formula (5.3) is proved similarly. □

5.3 Matrix analysis

5.3.1 Matrix Norms

Several interesting norms are related to singular values.

Frobenius norm. The standard scalar product on matrices is $\langle A, B\rangle_F = \sum_{i,j} A_{ij} B_{ij}$. It induces the Frobenius norm

$$\|A\|_F^2 = \sum_{i,j} A_{ij}^2 = \mathrm{Tr}(A^T A) = \sum_k \sigma_k(A)^2.$$

The last equality follows from the fact that the $\sigma_k(A)^2$ are the eigenvalues of $A^T A$.

We remind the reader two useful properties of the Frobenius scalar product. The proof of these properties is left to the reader.

Lemma 5.5 *For any matrices A, B, C with compatible dimensions, we have*

$$\langle AB, C\rangle_F = \langle A, CB^T\rangle_F = \langle B, A^T C\rangle_F \quad \textit{and} \quad \langle A, I\rangle_F = Tr(A).$$

Operator norm. The $\ell^2 \to \ell^2$ operator norm is defined by

$$|A|_{\mathrm{op}} = \sup_{\|x\|\leq 1} \|Ax\| = \sigma_1(A).$$

The last equality has been proved in the exercise page 62.

Nuclear norm. The nuclear norm is defined by

$$|A|_* = \sum_{k=1}^{r} \sigma_k(A).$$

Ky–Fan (p,q)**-norm.** For $p \geq 1$ and $q \in \mathbb{N}$, the Ky–Fan (p,q)-norm is defined by

$$\|A\|_{(p,q)} = \left(\sum_{k=1}^{q} \sigma_k(A)^p\right)^{1/p}, \tag{5.4}$$

We observe that $\|A\|_{(2,q)} \leq \|A\|_F$, with strict inequality if $q < \mathrm{rank}(A)$.

The three following inequalities are very useful.

Lemma 5.6 *We have*

1. $|A|_* \leq \sqrt{\mathrm{rank}(A)}\ \|A\|_F$,
2. $\langle A, B\rangle_F \leq |A|_*\, |B|_{\mathrm{op}}$,
3. $\|AB\|_F \leq |A|_{\mathrm{op}}\, \|B\|_F$.

Proof. The first inequality is simply Cauchy–Schwartz inequality. For the second inequality, we start from

$$\langle A, B\rangle_F = \sum_k \sigma_k(A)\langle u_k v_k^T, B\rangle_F = \sum_k \sigma_k(A)\langle u_k, Bv_k\rangle$$

and since $\|u_k\| = \|v_k\| = 1$, we notice that $\langle u_k, Bv_k\rangle \leq \|Bv_k\| \leq |B|_{\mathrm{op}}$. The inequality

$$\langle A, B\rangle_F \leq \sum_k \sigma_k(A)\, |B|_{\mathrm{op}} = |A|_*\, |B|_{\mathrm{op}}$$

then follows. Let us turn to the third inequality. We denote by B_j the j-th column of B. We observe that $\|B\|_F^2 = \sum_j \|B_j\|^2$, so

$$\|AB\|_F^2 = \sum_j \|(AB)_j\|^2 = \sum_j \|AB_j\|^2 \leq \sum_j |A|_{\mathrm{op}}^2\, \|B_j\|^2 = |A|_{\mathrm{op}}^2\, \|B\|_F^2.$$

The proof of Lemma 5.6 is complete. □

5.3.2 Low rank projection

We present in this section some useful results on singular values and SVD.

Lemma 5.7 *For an $n \times p$ matrix A and $k \leq \min(n,p)$, we have for any $q \times n$ matrix B*

$$\sigma_k(BA) \leq |B|_{\mathrm{op}}\, \sigma_k(A). \tag{5.5}$$

Similarly, we have for any $p \times q$ matrix B

$$\sigma_k(AB) \leq |B|_{\mathrm{op}}\, \sigma_k(A). \tag{5.6}$$

Proof. From the definition of the operator norm, we have $\|BAx\| \leq |B|_{\mathrm{op}} \|Ax\|$. The inequality (5.5) then follows from (5.2). Furthermore, we have $\sigma_k(AB) = \sigma_k(B^T A^T) \leq \left|B^T\right|_{\mathrm{op}} \sigma_k(A^T) = |B|_{\mathrm{op}} \sigma_k(A)$, which gives (5.6). □

The second result provides an improvement of the classical Cauchy–Schwartz inequality $\langle A, B\rangle_F \leq \|A\|_F \|B\|_F$ in terms of the Ky–Fan $(2, q)$-norm, with $q = \mathrm{rank}(A) \wedge \mathrm{rank}(B)$.

Lemma 5.8 *For any matrices $A, B \in \mathbb{R}^{n\times p}$, we set $q = \mathrm{rank}(A) \wedge \mathrm{rank}(B)$. We then have*

$$\langle A, B\rangle_F \leq \|A\|_{(2,q)} \|B\|_{(2,q)},$$

where the Ky–Fan $(2, q)$-norm $\|A\|_{(2,q)}$ is defined in (5.4).

Proof. By symmetry, we can assume, that the rank of B is not larger than the rank of A. Let us denote by q the rank of B and P_B the projection on the range of B. We have $B = P_B B$, so

$$\langle A, B\rangle_F = \langle A, P_B B\rangle_F = \langle P_B A, B\rangle_F \leq \|P_B A\|_F \|B\|_F.$$

The rank of $P_B A$ is at most q and previous lemma ensures that $\sigma_k(P_B A) \leq \sigma_k(A)$, so

$$\|P_B A\|_F^2 = \sum_{k=1}^{q} \sigma_k(P_B A)^2 \leq \sum_{k=1}^{q} \sigma_k(A)^2 = \|A\|_{(2,q)}^2.$$

Since $q = \mathrm{rank}(B)$, we have $\|B\|_F = \|B\|_{(2,q)}$, and the lemma is proved. □

The last result characterizes the "projection" on the set of matrices of rank r.

Theorem 5.9 *For $A = \sum_{k=1}^{r} \sigma_k(A) u_k v_k^T$ and $q < r$, we have*

$$\min_{B:\mathrm{rank}(B)\leq q} \|A - B\|_F^2 = \sum_{k=q+1}^{r} \sigma_k(A)^2,$$

where the minimum is achieved for

$$B = \sum_{k=1}^{q} \sigma_k(A) u_k v_k^T.$$

Proof. According to Lemma 5.8, for any matrix B of rank $q < r$, we have

$$\|A - B\|_F^2 = \|A\|_F^2 - 2\langle A, B\rangle_F + \|B\|_F^2 \geq \|A\|_F^2 - 2\|A\|_{(2,q)} \|B\|_F + \|B\|_F^2.$$

The right-hand side is minimum for $\|B\|_F = \|A\|_{(2,q)}$, so

$$\|A - B\|_F^2 \geq \|A\|_F^2 - \|A\|_{(2,q)}^2 = \sum_{k=q+1}^{r} \sigma_k(A)^2.$$

Finally, we observe that this lower bound is achieved for $B = \sum_{k=1}^{q} \sigma_k(A) u_k v_k^T$. □

5.4 Explicit computations with SVD decomposition

5.4.1 Moore–Penrose Pseudo-Inverse

The Moore–Penrose pseudo-inverse A^+ of a matrix A generalizes the notion of inverse for singular matrices. It is a matrix such that $AA^+y = y$ for all y in the range of A and $A^+Ax = x$ for all x in the range of A^+. Furthermore, the matrices AA^+ and A^+A are symmetric. When A is nonsingular, we have the identity $A^+ = A^{-1}$. We first describe A^+ for diagonal matrices, then for symmetric matrices, and finally for arbitrary matrices.

Diagonal matrices
The Moore–Penrose pseudo-inverse of a diagonal matrix D is a diagonal matrix D^+, with diagonal entries $[D^+]_{jj} = 1/D_{jj}$ when $D_{jj} \neq 0$ and $[D^+]_{jj} = 0$ otherwise.

Symmetric matrices
Write $A = UDU^T$ for a spectral decomposition of A with D diagonal and U unitary[1]. The Moore–Penrose pseudo-inverse of A is given by $A^+ = UD^+U^T$.

Arbitrary matrices
Write $A = \sum_{j=1}^r \sigma_j(A)u_jv_j^T$ for a singular value decomposition of A with $r = \text{rank}(A)$. The Moore–Penrose pseudo-inverse of A is given by

$$A^+ = \sum_{j=1}^r \sigma_j(A)^{-1}v_ju_j^T.$$

We notice that

$$A^+A = \sum_{j=1}^r v_jv_j^T = \text{Proj}_{\text{range}(A^T)} \quad \text{and} \quad AA^+ = \sum_{j=1}^r u_ju_j^T = \text{Proj}_{\text{range}(A)},$$

so $AA^+y = y$ for all y in the range of A, and $A^+Ax = x$ for all x in the range of A^+. In particular, when A is nonsingular, we have $AA^+ = A^+A = I$, so $A^+ = A^{-1}$.

5.4.2 Problem: Ridge regression

Preliminaries on random vectors

Let Z be a random vector in $\mathbb{R}^p$ and A be a (non-random) $n \times p$ matrix. Prove that

$$\mathbb{E}\left[\|Z\|^2\right] = \|\mathbb{E}\left[Z\right]\|^2 + \text{Tr}(\text{Cov}(Z)),$$

and

$$\text{Cov}(AZ) = A\text{Cov}(Z)A^T.$$

These two formulas are very useful and should be known by heart.

Ridge regression

We consider the linear model $Y = \mathbf{X}\beta + \varepsilon$, with $Y, \varepsilon \in \mathbb{R}^n$ et $\beta \in \mathbb{R}^p$. The matrix $\mathbf{X}$ and the vector β are non-random. We assume that $\mathbf{E}[\varepsilon] = 0$ and $\mathbf{Cov}(\varepsilon) = \sigma^2 I_n$.

We only observe the vector $Y \in \mathbb{R}^n$ and the matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$. Our goal is to estimate the vector β. In the following, the dimension p can be larger than the dimension n.

For $\lambda > 0$, the ridge estimator $\widehat{\beta}_\lambda$ is defined by

$$\widehat{\beta}_\lambda \in \underset{\beta\in\mathbb{R}^p}{\text{argmin}}\, \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2. \tag{5.7}$$

[1] U unitary if $U^TU = UU^T = I$.

1. Check that $\mathcal{L}$ is strictly convex and has a unique minimum in $\mathbb{R}^p$. Is-it still true when $\lambda = 0$?
2. Prove that $\widehat{\beta}_\lambda = A_\lambda Y$ with $A_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^T$.
3. Let $\sum_{k=1}^r \sigma_k u_k v_k^T$ be a singular value decomposition of $\mathbf{X}$. Prove that

$$A_\lambda = \sum_{k=1}^r \frac{\sigma_k}{\sigma_k^2 + \lambda} v_k u_k^T \overset{\lambda \to 0+}{\to} \mathbf{X}^+$$

   where A^+ is the Moore–Penrose pseudo-inverse of A.

4. Check that we have

$$\mathbf{X}\widehat{\beta}_\lambda = \sum_{k=1}^r \frac{\sigma_k^2}{\sigma_k^2 + \lambda} \langle u_k, Y\rangle \, u_k. \tag{5.8}$$

5. Check that we have

$$\mathbb{E}\left[\widehat{\beta}_\lambda\right] = \sum_{k=1}^r \frac{\sigma_k^2}{\sigma_k^2 + \lambda} \langle v_k, \beta\rangle v_k.$$

6. Prove that the mean square error $\mathbb{E}\left[\|\widehat{\beta}_\lambda - \beta\|^2\right]$ of the Ridge estimator can be decomposed as

$$\mathbb{E}\left[\|\widehat{\beta}_\lambda - \beta\|^2\right] = \left\|\beta - \mathbb{E}\left[\widehat{\beta}_\lambda\right]\right\|^2 + \mathrm{Tr}(\mathrm{Cov}(\widehat{\beta}_\lambda)).$$

   The first term is the norm of the bias of the Ridge estimator and the second term is the variance $\mathbb{E}\left[\|\widehat{\beta}_\lambda - \mathbb{E}[\widehat{\beta}_\lambda]\|^2\right]$ of the Ridge estimator.

7. Let us denote by $P = \sum_{j=1}^r v_j v_j^T$ the projection on the range of $\mathbf{X}^T$. Prove the following formula for the bias term

$$\left\|\beta - \mathbb{E}\left[\widehat{\beta}_\lambda\right]\right\|^2 = \|\beta - P\beta\|^2 + \sum_{k=1}^r \left(\frac{\lambda}{\lambda + \sigma_k^2}\right)^2 \langle v_k, \beta\rangle^2.$$

8. Check that the variance of the ridge estimator is given by

$$\mathrm{Tr}(\mathrm{Cov}(\widehat{\beta}_\lambda)) = \sigma^2 \mathrm{Tr}(A_\lambda A_\lambda^T) = \sigma^2 \sum_{k=1}^r \left(\frac{\sigma_k}{\sigma_k^2 + \lambda}\right)^2.$$

9. How do the bias and the variance of $\widehat{\beta}_\lambda$ vary when λ increases?

Remark. We notice from (5.8) that $\mathbf{X}\widehat{\beta}_\lambda$ shrinks Y in the directions u_k where $\sigma_k \ll \lambda$, whereas it leaves Y almost unchanged in the directions u_k where $\sigma_k \gg \lambda$.

Chapter 6

Perturbation bounds

In many situations in statistics and machine learning, we have access to a matrix A of observations, which is a noisy version $A = B + E$ of an unknown signal matrix B of interest. Since we only have access to A, what SVD or spectral properties of B can we learn from the SVD or spectral properties of A?

In this chapter, we will provide some perturbation bounds which relates the SVD or spectral properties of A and B. The first results hold for any perturbation E and then the case of random perturbations are investigated.

6.1 Singular values localization

Weyl inequalities provide some relationships between the singular values of A and B. The first result states that the singular values are 1-Lipschitz with respect to the operator norm.

Theorem 6.1 Weyl inequality
For two $n \times p$ matrices A and B, we have for any $k \leq \min(n, p)$

$$|\sigma_k(A) - \sigma_k(B)| \leq \sigma_1(A - B) = |A - B|_{\mathrm{op}}\,. \tag{6.1}$$

Proof. For any $x \in \mathbb{R}^p \setminus \{0\}$, we have

$$\frac{\|Ax\|}{\|x\|} \leq \frac{\|Bx\|}{\|x\|} + \frac{\|(A-B)x\|}{\|x\|} \leq \frac{\|Bx\|}{\|x\|} + \sigma_1(A - B).$$

The Inequality (6.1) follows by applying the Max–Min formula (5.2). □

The Inequality (6.1) can be generalized as follows.

Theorem 6.2 Generalized Weyl inequalities
For two $n \times p$ matrices A and B, and any i, j with $i + j - 2 \leq \min(n, p)$, we have

$$\sigma_{i+j-1}(B) \leq \sigma_i(A) + \sigma_j(A - B).$$

We refer to the Exercise 6.4.1 page 76 for a proof of these inequalities.

6.2 Eigenspaces localization

As the left (respectively right) singular vectors of a matrix M are the eigenvectors of the symmetric matrix MM^T (resp. M^TM), it is enough to get perturbation bounds for the eigenvectors of symmetric matrices. This is the topic of this section.

Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices and let $A = \sum_k \lambda_k u_k u_k^T$ and $B = \sum_k \rho_k v_k v_k^T$

be their eigenvalue decomposition with $\lambda_1 \geq \cdots \geq \lambda_n$ and $\rho_1 \geq \cdots \geq \rho_n$. The vectors $\{u_1, \dots, u_n\}$ and $\{v_1, \dots, v_n\}$ are two orthonormal bases of $\mathbb{R}^n$. We want to compare the eigenspaces span $\{u_1, \dots, u_r\}$ and span $\{v_1, \dots, v_r\}$, spanned by the r leading eigenvectors of A and B.

A first idea could be to compare the two matrices $U_r = [u_1, \dots, u_r]$ and $V_r = [v_1, \dots, v_r]$. Yet, there are some orthogonal transformation R such that span $\{Ru_1, \dots, Ru_r\}$ = span $\{u_1, \dots, u_r\}$, but $RU_r \neq U_r$, so a directed comparison of U_r and V_r is not suited. Instead, we will compare

$$U_r U_r^T = \sum_{k=1}^{r} u_k u_k^T \quad \text{and} \quad V_r V_r^T = \sum_{k=1}^{r} v_k v_k^T,$$

which are the orthogonal projectors in $\mathbb{R}^n$ onto the linear spans span $\{u_1, \dots, u_r\}$ and span $\{v_1, \dots, v_r\}$ respectively. Next lemma relates the Frobenius distance between $U_r U_r^T$ and $V_r V_r^T$ to the Frobenius norm of $U_{-r}^T V_r$.

Lemma 6.3 *Let $U_{-r} = [u_{r+1}, \dots, u_n]$ and $V_{-r} = [v_{r+1}, \dots, v_n]$. Then, we have*

$$\|U_r U_r^T - V_r V_r^T\|_F^2 = 2\|V_{-r}^T U_r\|_F^2 = 2\|U_{-r}^T V_r\|_F^2.$$

Proof of Lemma 6.3. We first expand the squares and use that the Frobenius norm of a projector is equal to its rank

$$\begin{aligned}\|U_r U_r^T - V_r V_r^T\|_F^2 &= \|U_r U_r^T\|_F^2 + \|V_r V_r^T\|_F^2 - 2\langle U_r U_r^T, V_r V_r^T\rangle_F \\ &= 2r - 2\langle U_r U_r^T, V_r V_r^T\rangle_F\end{aligned}$$

Then, since span$\{v_{r+1}, \dots, v_n\}$ is the orthogonal complement of span$\{v_1, \dots, v_r\}$, we have $V_r V_r^T = I_n - V_{-r} V_{-r}^T$. So, as $\langle U_r U_r^T, I_n\rangle = \mathrm{Tr}(\mathrm{U_r U_r^T}) = \mathrm{r}$,

$$\begin{aligned}\|U_r U_r^T - V_r V_r^T\|_F^2 &= 2r - 2\langle U_r U_r^T, I_n - V_{-r} V_{-r}^T\rangle_F \\ &= 2\langle V_{-r}^T U_r, V_{-r}^T U_r\rangle = 2\|V_{-r}^T U_r\|_F^2.\end{aligned}$$

The second equality of Lemma 6.3 follows by symmetry. □

Next result is the main theorem of this chapter. It provides a (classical) upper-bound on the norm $\|U_{-r}^T V_r\|_F^2$.

Theorem 6.4 Davis-Kahan perturbation bound.
Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices and let $A = \sum_k \lambda_k u_k u_k^T$ and $B = \sum_k \rho_k v_k v_k^T$ be their eigenvalue decomposition with $\lambda_1 \geq \cdots \geq \lambda_n$ and $\rho_1 \geq \cdots \geq \rho_n$.
Let $U_r = [u_1, \dots, u_r]$, $U_{-r} = [u_{r+1}, \dots, u_n]$ and similarly $V_r = [v_1, \dots, v_r]$, $V_{-r} = [v_{r+1}, \dots, v_n]$. Then, we have

$$\begin{aligned}\|U_{-r}^T V_r\|_F &\leq \frac{(\sqrt{r}\,|A-B|_{\mathrm{op}}) \wedge \|A-B\|_F}{(\rho_r - \lambda_{r+1}) \vee (\lambda_r - \rho_{r+1})} && (6.2)\\ &\leq 2\,\frac{(\sqrt{r}\,|A-B|_{\mathrm{op}}) \wedge \|A-B\|_F}{\lambda_r - \lambda_{r+1}}. && (6.3)\end{aligned}$$

In many cases, we only wish to compare the two leading eigenvectors of A and B, which corresponds to the case $r = 1$.

Corollary 6.5 Comparing leading eigenvectors.

$$\sqrt{1 - \langle u_1, v_1\rangle^2} \leq \frac{2\inf_{\lambda\in\mathbb{R}} |A + \lambda I - B|_{\mathrm{op}}}{\lambda_1 - \lambda_2}. \qquad (6.4)$$

Proof of Corollary 6.5.
We first observe that

$$\|U_{-1}^T v_1\|^2 = v_1^T U_{-1} U_{-1}^T v_1 = v_1^T (I - u_1 u_1^T) v_1 = 1 - (u_1^T v_1)^2.$$

In addition, the eigenvectors of A and $A + \lambda I$ are the same, while the eigenvalues are all translated by λ, preserving the eigengap between the two largest eigenvalues. So, for any $\lambda \in \mathbb{R}$, the Inequality (6.3) applied to $A + \lambda I$ and B gives

$$\sqrt{1 - \langle u_1, v_1 \rangle^2} \le \frac{2\,|A + \lambda I - B|_{\text{op}}}{\lambda_1 - \lambda_2}.$$

The proof Corollary 6.5 is complete. □

Proof of Theorem 6.4.
We first observe that the Bound (6.3) directly follows from (6.2) and the inequalities

$$\begin{aligned}\lambda_r - \lambda_{r+1} &= \lambda_r - \rho_{r+1} - (\rho_r - \rho_{r+1}) + \rho_r - \lambda_{r+1} \\ &\le (\lambda_r - \rho_{r+1}) + (\rho_r - \lambda_{r+1}) \le 2((\rho_r - \lambda_{r+1}) \vee (\lambda_r - \rho_{r+1})).\end{aligned}$$

Let us prove (6.2). Since $\lambda_r - \lambda_{r+1} \ge 0$, we notice that the inequality above also gives

$$(\rho_r - \lambda_{r+1}) \vee (\lambda_r - \rho_{r+1}) = (\rho_r - \lambda_{r+1})_+ \vee (\lambda_r - \rho_{r+1})_+.$$

In addition, we observe from Lemma 6.3 that the role of A and B are symmetric. Hence, we only need to prove

$$\|U_{-r}^T V_r\|_F \le \frac{(\sqrt{r}\,|A - B|_{\text{op}}) \wedge \|A - B\|_F}{(\rho_r - \lambda_{r+1})_+}. \tag{6.5}$$

When $\rho_r \le \lambda_{r+1}$ the right-hand side is infinite, so we only need to focus on the case where $\rho_r > \lambda_{r+1}$.
We have the decomposition

$$\|U_{-r}^T V_r\|_F^2 = \sum_{k=1}^{r} \|U_{-r}^T v_k\|^2, \tag{6.6}$$

so we will start by bounding the square norms $\|U_{-r}^T v_k\|^2$. Since

$$A = \sum_{k=1}^{n} \lambda_k u_k u_k^T = U_r \operatorname{diag}(\lambda_1, \ldots, \lambda_r) U_r^T + U_{-r} \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_n) U_{-r}^T,$$

we have $U_{-r}^T A = \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_n) U_{-r}^T$. Hence, with $Bv_k = \rho_k v_k$, we have for $k = 1, \ldots, r$

$$\begin{aligned}\rho_k U_{-r}^T v_k &= U_{-r}^T B v_k = U_{-r}^T (A + B - A) v_k \\ &= \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_n)\, U_{-r}^T v_k + U_{-r}^T (B - A) v_k.\end{aligned}$$

Hence,

$$\operatorname{diag}(\rho_k - \lambda_{r+1}, \ldots, \rho_k - \lambda_n)\, U_{-r}^T v_k = U_{-r}^T (B - A) v_k,$$

and then, since $\rho_k \ge \rho_r > \lambda_{r+1}$,

$$\begin{aligned}\|U_{-r}^T v_k\|^2 &\le \left|\operatorname{diag}(\rho_k - \lambda_{r+1}, \ldots, \rho_k - \lambda_n)^{-1} U_{-r}^T\right|_{\text{op}}^2 \|(B - A) v_k\|^2 \\ &\le \frac{\|(B - A) v_k\|^2}{(\rho_k - \lambda_{r+1})^2} \le \frac{\|(B - A) v_k\|^2}{(\rho_r - \lambda_{r+1})^2},\end{aligned}$$

for $k = 1, \dots, r$. To conclude, we observe that

$$\sum_{k=1}^{r} \|(B-A)v_k\|^2 \leq |B-A|_{\text{op}}^2 \sum_{k=1}^{r} \|v_k\|^2 = r\,|B-A|_{\text{op}}\,,$$

since $\|v_k\| = 1$. So, with (6.6) we get

$$\|U_{-r}^T V_r\|_F^2 \leq \frac{r\,|A-B|_{\text{op}}^2}{(\rho_r - \lambda_{r+1})^2}. \tag{6.7}$$

In addition, since $\sigma_k((B-A)V_r) \leq |V_r|\sigma_k(B-A) = \sigma_k(B-A)$

$$\sum_{k=1}^{r} \|(B-A)v_k\|^2 = \|(B-A)V_r\|_F^2 = \sum_{k=1}^{n} \sigma_k((B-A)V_r)^2 \leq \sum_{k=1}^{n} \sigma_k(B-A)^2 = \|B-A\|_F^2.$$

Combining this bound with (6.6), we get

$$\|U_{-r}^T V_r\|_F^2 \leq \frac{\|A-B\|_F^2}{(\rho_r - \lambda_{r+1})^2}. \tag{6.8}$$

Combining (6.7) and (6.8) we get (6.5), so the proof of Theorem 6.4 is complete. □

6.3 Operator norm of random matrices

As mentioned in the preamble of this chapter, in statistics and machine learning, we often observe a matrix A which is the sum of a signal matrix B and a random noise matrix E. According to the Weyl inequality and the Davis-Kahan inequality, the difference between the spectral or singular value decomposition of $A = B + E$ and B is controlled in terms of the operator norm $|E|_{\text{op}}$. So to understand the size of the perturbation induced by the noise, we need to understand the size of the operator norm of a random matrix E.

For simplicity, we focus on this chapter on the case where E has i.i.d. entries E_{ij} with $\mathcal{N}(0,\sigma^2)$ distribution. All the results of this section are valid for i.i.d. entries with $subG(\sigma^2)$ distribution, at the price of larger numerical constants.

6.3.1 Concentration of quadratic forms of Gaussian vectors

To bound the operator norm of a random matrix, we will need to evaluate quadratic forms of Gaussian vectors. Next lemma gathers two simple versions of Hanson-Wright inequality for Gaussian vectors.

Theorem 6.6 Hanson-Wright inequality for Gaussian vectors

Symmetric forms. *Let ε be a standard Gaussian random variable $\mathcal{N}(0, I_p)$ in $\mathbb{R}^p$ and S be a real symmetric $p \times p$ matrix. Then, we have for any $L \geq 0$*

$$\mathbb{P}\left[\varepsilon^T S \varepsilon - \text{Tr}(S) > \sqrt{8\|S\|_F^2 L} \vee (8\,|S|_{\text{op}}\,L)\right] \leq e^{-L}. \tag{6.9}$$

Cross products. *Let $\varepsilon, \varepsilon'$ be two independent standard Gaussian random variable $\mathcal{N}(0, I_p)$ in $\mathbb{R}^p$ and A be* any *real $p \times p$ matrix. Then, we have for any $L \geq 0$*

$$\mathbb{P}\left[\varepsilon^T A \varepsilon' > \sqrt{4\|A\|_F^2 L} \vee (4\,|A|_{\text{op}}\,L)\right] \leq e^{-L}. \tag{6.10}$$

Remarks:

1. The Inequality (6.9) is equivalent to the following statement: there exists an exponential random variable ξ with parameter 1 such that
$$\varepsilon^T S\varepsilon - \mathrm{Tr}(S) \leq \sqrt{8\|S\|_F^2\xi} \vee (8\,|S|_{\mathrm{op}}\,\xi).$$
Using this compact formulation can be useful in complex proofs.
2. The Hanson-Wright inequalities remain valid (with worst constants) when ε has independent subG(1) entries. For example, in Chapter 1 we have proved such a bound in (1.2), page 5, for a diagonal matrix $S = \mathrm{diag}(a_1, \dots, a_n)$.

Proof of Theorem 6.6.
Symmetric forms. The proof is based on the classical Chernoff argument. Next lemma provides an upper-bound on the Laplace transform of a square Gaussian random variable.

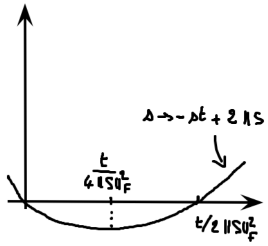
Lemma 6.7 *Let Z be a* $\mathcal{N}(0, 1)$ *standard Gaussian random variable. Then, for any* $|s| \leq 1/4$*, we have*
$$\mathbb{E}\left[\exp(s(Z^2 - 1))\right] \leq e^{2s^2}.$$

Proof of Lemma 6.7. Since $-\log(1 - x) \leq x + x^2$ for $|x| \leq 1/2$, we have
$$\mathbb{E}\left[\exp(s(Z^2 - 1))\right] = \frac{e^{-s}}{(1 - 2s)^{1/2}} \leq e^{2s^2}.$$

Le proof of Lemma 6.7 is complete. □

Since S is symmetric, we can diagonalize it, $S = \sum_{k=1}^p \lambda_k v_k v_k^T$ and
$$\varepsilon^T S\varepsilon = \sum_{k=1}^p \lambda_k (v_k^T \varepsilon)^2.$$

Since the eigenvectors $\{v_1, \dots, v_p\}$ form an orthonormal basis of $\mathbb{R}^p$, the matrix $V = [v_1, \dots, v_p]$ fulfills $V^T V = I$. Hence, $Z = V^T\varepsilon$ follows a $\mathcal{N}(0, I)$ distribution, which means that the random variables $Z_k = v_k^T\varepsilon$, for $k = 1, \dots, p$ are i.i.d. $\mathcal{N}(0, 1)$-random variables.
Applying Markov inequality, we get for $t \geq 0$ and $|s| \leq (4\,|S|_{\mathrm{op}})^{-1}$
$$\begin{aligned}\mathbb{P}\left[\varepsilon^T S\varepsilon - \mathrm{Tr}(S) > t\right] &\leq e^{-st}\,\mathbb{E}\left[e^{s(\varepsilon^T S\varepsilon - \mathrm{Tr}(S))}\right]\\ &\leq e^{-st}\prod_{k=1}^p \mathbb{E}\left[\exp\left(s\lambda_k(Z_k^2 - 1)\right)\right]\\ &\leq \exp\left(-st + 2s^2\sum_{k=1}^p \lambda_k^2\right) = \exp\left(-st + 2\|S\|_F^2 s^2\right)\end{aligned}$$



The minimum of $s \to -st + 2\|S\|_F^2 s^2$ over $|s| \leq (4\,|S|_{\text{op}})^{-1}$ is achieved for $s = \frac{1}{4}(t/\|S\|_F^2) \wedge (1/|S|_{\text{op}})$ and hence

$$\min_{|s|\leq(4|S|_{\text{op}})^{-1}} \left(-st + 2\|S\|_F^2 s^2\right) = -\frac{t^2}{8\|S\|_F^2}\mathbf{1}_{\{t\leq\|S\|_F^2/|S|_{\text{op}}\}} + \left(\frac{\|S\|_F^2}{8\,|S|_{\text{op}}^2} - \frac{t}{4\,|S|_{\text{op}}}\right)\mathbf{1}_{\{t>\|S\|_F^2/|S|_{\text{op}}\}}$$
$$\leq -\frac{1}{8}\left(\frac{t^2}{\|S\|_F^2} \wedge \frac{t}{|S|_{\text{op}}}\right).$$

The Bound (6.9) follows.

Cross products. The trick for (6.10) is to notice that

$$\varepsilon^T A \varepsilon' = \begin{bmatrix}\varepsilon\\ \varepsilon'\end{bmatrix}^T S \begin{bmatrix}\varepsilon\\ \varepsilon'\end{bmatrix}, \quad \text{with} \quad S = \frac{1}{2}\begin{bmatrix}0 & A\\ A^T & 0\end{bmatrix}.$$

Since S is symmetric we can apply (6.9). The conclusion follows by checking that $\text{Tr}(S) = 0$, $|S|_{\text{op}} = |A|_{\text{op}}/2$ and $\|S\|_F^2 = \|A\|_F^2/2$. □

6.3.2 Concentration of random Gram matrices

In this section, we derive some bounds on the operator norm $|E|_{\text{op}} = \sigma_1(E)$ of a random matrix with entries E_{ij} following an i.i.d. $\mathcal{N}(0,\sigma^2)$ distribution. As $|E|_{\text{op}} = \sqrt{\left|EE^T\right|_{\text{op}}}$, we focus on the random Gram matrix EE^T.

We first observe that

$$\mathbb{E}\left[E_{ik}E_{jk}\right] = \begin{cases} \mathbb{E}[E_{ik}]\,\mathbb{E}[E_{jk}] = 0 & \text{for} \quad i \neq j \\ \mathbb{E}\left[E_{ik}^2\right] = \sigma^2 & \text{for} \quad i = j. \end{cases}$$

Hence, we have $\mathbb{E}[EE^T] = p\sigma^2 I_n$. Instead of simply upper-bounding $\left|EE^T\right|_{\text{op}}$, we will give a more precise result by bounding the fluctuations of EE^T around its expectation $\mathbb{E}[EE^T] = p\sigma^2 I_n$.

Theorem 6.8 Concentration of EE^T.
Assume that $E \in \mathbb{R}^{n\times p}$ has i.i.d. entries following a $\mathcal{N}(0,\sigma^2)$ distribution. Then, there exists a random variable ξ with exponential distribution with parameter 1 such that

$$\left|EE^T - p\sigma^2 I_n\right|_{\text{op}} \leq 4\sigma^2\sqrt{p(6n+2\xi)} + (48n + 16\xi)\sigma^2. \tag{6.11}$$

We can derive from this theorem the following control on $|E|_{\text{op}}$.

Corollary 6.9 *Under the hypotheses of Theorem 6.8, there exists a random variable ξ with exponential distribution with parameter 1 such that*

$$|E|_{\text{op}} \leq \sigma\left(\sqrt{p} + 7\sqrt{n+\xi}\right).$$

Proof of Corollary 6.9.
By the triangular inequality, we have

$$\sigma_1(E)^2 = \left|EE^T\right|_{\text{op}} \leq \left|p\sigma^2 I_n\right|_{\text{op}} + \left|EE^T - p\sigma^2 I_n\right|_{\text{op}}$$
$$\leq \sigma^2\left(p + 4\sqrt{p(6n+2\xi)} + 48n + 16\xi\right)$$
$$\leq \sigma^2\left(\sqrt{p} + 7\sqrt{n+\xi}\right)^2.$$

The Corollary follows. □

Proof of Theorem 6.8.
Sketch of the proof. Before starting the proof of Theorem 6.8, let us sketch the main lines.
First of all, dividing both sides of (6.11) by σ^2, we can assume with no loss of generality that $\sigma^2 = 1$.
As $EE^T - pI_n$ is a symmetric matrix, we have

$$\left|EE^T - pI_n\right|_{\text{op}} = \sup_{x \in \partial B_{\mathbb{R}^n}(0,1)} |\langle (EE^T - pI_n)x, x\rangle|.$$

For a given $x \in \partial B_{\mathbb{R}^n}(0,1)$, the scalar product $\langle (EE^T - pI_n)x, x\rangle$ is a quadratic form of independent Gaussian random variables, and hence its random fluctuations can be controlled by Hanson-Wright inequality (6.9).
Then, we have to handle the supremum of the scalar products $\langle (EE^T - pI_n)x, x\rangle$ over all $x \in \partial B_{\mathbb{R}^n}(0,1)$. The supremum of n random variables $Z_1, \ldots, Z_n$ can be handled easily with a union bound

$$\mathbb{P}\left[\max_{i=1,\ldots,n} Z_i > t\right] \le \sum_{i=1}^{n} \mathbb{P}[Z_i > t].$$

Here, we have a supremum over an infinite (even uncountable) set $\partial B_{\mathbb{R}^n}(0,1)$, so we cannot implement directly such an union bound. Yet, we notice that for two close x and y, the random values $\langle (EE^T - pI_n)x, x\rangle$ and $\langle (EE^T - pI_n)y, y\rangle$ are also close. Hence, the recipe is to discretize the ball $\partial B_{\mathbb{R}^n}(0,1)$ and to control the supremum over $\partial B_{\mathbb{R}^n}(0,1)$ by a supremum over the discretization of the ball plus the error made when replacing $\partial B_{\mathbb{R}^n}(0,1)$ by its discretization.

The proof then proceeds into three steps: first a discretization of the supremum over $\partial B_{\mathbb{R}^n}(0,1)$, then a concentration bound on the scalar product $\langle (EE^T - pI_n)x, x\rangle$ based on Hanson-Wright inequality and finally a union bound to conclude.

Step 1: Discretization. Let $\partial B_{\mathbb{R}^n}(0,1)$ denote the unit sphere in $\mathbb{R}^n$. For any symmetric matrix A, the operator norm of A is equal to

$$|A|_{\text{op}} = \sup_{x \in \partial B_{\mathbb{R}^n}(0,1)} |\langle Ax, x\rangle|.$$

As explained above, since $\partial B_{\mathbb{R}^n}(0,1)$ is an infinite set, we cannot directly use an union bound in order to control the fluctuation of the supremum. Instead, we use a discretized version of the above equality, in order to be able to apply an union bound.
A set $\mathcal{N}_\varepsilon \subset \partial B_{\mathbb{R}^n}(0,1)$ is called an ε-net of $\partial B_{\mathbb{R}^n}(0,1)$, if for any $x \in \partial B_{\mathbb{R}^n}(0,1)$, there exists $y \in \mathcal{N}_\varepsilon$ such that $\|x - y\| \le \varepsilon$. Next lemma links the operator norm of a matrix to a supremum over an ε-net.

Lemma 6.10 *For any symmetric matrix $A \in \mathbb{R}^{n\times n}$ and any ε-net of $\partial B_{\mathbb{R}^n}(0,1)$, we have*

$$|A|_{\text{op}} \le \frac{1}{1-2\varepsilon} \sup_{x \in \mathcal{N}_\varepsilon} |\langle Ax, x\rangle|. \tag{6.12}$$

Proof of Lemma 6.10.
Let $x^* \in \partial B_{\mathbb{R}^n}(0,1)$ be such that $|A|_{\text{op}} = |\langle Ax^*, x^*\rangle|$ and let $y \in \mathcal{N}_\varepsilon$ fulfilling $\|x^* - y\| \le \varepsilon$. According to the decomposition

$$\langle Ax^*, x^*\rangle = \langle Ay, y\rangle + \langle A(x^* - y), y\rangle + \langle Ax^*, x^* - y\rangle,$$

and the triangular inequality, we have

$$|A|_{\mathrm{op}} = |\langle Ax^*, x^*\rangle| \leq |\langle Ay, y\rangle| + |\langle A(x^* - y), y\rangle| + |\langle Ax^*, x^* - y\rangle|$$
$$\leq \sup_{y\in\mathcal{N}_\varepsilon} |\langle Ay, y\rangle| + 2\,|A|_{\mathrm{op}}\,\varepsilon.$$

The Bound (6.12) then follows. □

Next lemma provides an upper bound on the cardinality of a minimal ε-net of $\partial B_{\mathbb{R}^n}(0,1)$.

Lemma 6.11 *For any $n \in \mathbb{N}$ and $\varepsilon > 0$, there exists an ε-net of $\partial B_{\mathbb{R}^n}(0,1)$ with cardinality upper bounded by*

$$|\mathcal{N}_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^n.$$

We refer to the Exercise 6.4.2 for a proof of this lemma based on volumetric arguments. Choosing $\varepsilon = 1/4$, we get the existence of an 1/4-net $\mathcal{N}_{1/4}$ of $\partial B_{\mathbb{R}^n}(0,1)$ with cardinality at most 9^n and such that

$$\left|EE^T - pI_n\right|_{\mathrm{op}} \leq 2 \max_{x\in\mathcal{N}_{1/4}} |\langle(EE^T - pI_n)x, x\rangle| = 2 \max_{x\in\mathcal{N}_{1/4}} \left|\|E^Tx\|^2 - p\right|. \tag{6.13}$$

Step 2: concentration of the quadratic forms. Let $E_{:i}$ denotes the ith column of E. We observe that $x^TE_{:i} \sim \mathcal{N}(0, x^Tx)$ and that the $(x^TE_{:i})_{i=1,\ldots,p}$ are independent since the columns $E_{:i}$ are independent. Hence, since $\mathcal{N}_{1/4} \subset \partial B_{\mathbb{R}^n}(0,1)$, the coordinates $[E^Tx]_i = E_{:i}^Tx$ are i.i.d. $\mathcal{N}(0,1)$, and the random vector $\varepsilon_x = E^Tx$ follows a standard Gaussian distribution $\mathcal{N}(0, I_p)$ in $\mathbb{R}^p$. Hanson-Wright inequality (6.9) with $S = I_p$ and $S = -I_p$ ensures that there exist two exponential random variables ξ_x, ξ'_x, such that

$$\|E^Tx\|^2 - p \leq \sqrt{8p\xi_x} \vee 8\xi_x \quad \text{and} \quad p - \|E^Tx\|^2 \leq \sqrt{8p\xi'_x} \vee 8\xi'_x.$$

Therefore, combining with (6.13), we obtain the concentration bound

$$\left|EE^T - pI_n\right|_{\mathrm{op}} \leq 2\left(\sqrt{8p \max_{x\in\mathcal{N}_{1/4}} (\xi_x \vee \xi'_x)} + 8 \max_{x\in\mathcal{N}_{1/4}} (\xi_x \vee \xi'_x)\right). \tag{6.14}$$

Step 3: Union bound. An union bound gives

$$\mathbb{P}\left[\max_{x\in\mathcal{N}_{1/4}} (\xi_x \vee \xi'_x) > \log(2|\mathcal{N}_{1/4}|) + t\right] \leq \sum_{x\in\mathcal{N}_{1/4}} \left(\mathbb{P}\left[\xi_x > \log(2|\mathcal{N}_{1/4}|) + t\right] + \mathbb{P}\left[\xi'_x > \log(2|\mathcal{N}_{1/4}|) + t\right]\right)$$
$$\leq 2|\mathcal{N}_{1/4}|\ \exp(-\log(2|\mathcal{N}_{1/4}|) - t) = e^{-t},$$

so there exists an exponential random variable ξ with parameter 1 such that

$$\max_{x\in\mathcal{N}_{1/4}} (\xi_x \vee \xi'_x) \leq \log(2|\mathcal{N}_{1/4}|) + \xi \leq 3n + \xi.$$

Combining this bound with (6.14), we obtain (6.11). The proof of Theorem 6.8 is complete. □

6.4 Exercices

6.4.1 Generalized Weyl inequalities

In this exercise, you will prove the Theorem 6.2.

1. Check that there exists two linear spans $S_i \subset \mathbb{R}^p$ and $S_j \subset \mathbb{R}^p$ of codimension $i-1$ and $j-1$ such that
$$\max_{x \in S_i \setminus \{0\}} \frac{\|Ax\|}{\|x\|} \leq \sigma_i(A) \quad \text{and} \quad \max_{x \in S_j \setminus \{0\}} \frac{\|(B-A)x\|}{\|x\|} \leq \sigma_j(B-A).$$
2. Check that the codimension of $S_i \cap S_j$ is not larger than $i+j-2$.
3. Prove that for any S with dimension $i+j-1$, we have $S \cap S_i \cap S_j \neq \{0\}$.
4. Conclude with the Max–Min formula (5.2).

6.4.2 Cardinality of an ε-net

In this exercise, we prove the Lemma 6.11. Let us define $\mathcal{N}_\varepsilon$ as follows. Start from any $x_1 \in \partial B_{\mathbb{R}^n}(0,1)$, and for $k = 2, 3, \ldots$ choose recursively any $x_k \in \partial B_{\mathbb{R}^n}(0,1)$ such that $x_k \notin \cup_{j=1,\ldots,k-1} B_{\mathbb{R}^n}(x_j, \varepsilon)$. When no such x_k remains, stop and define $\mathcal{N}_\varepsilon = \{x_1, x_2, \ldots\}$.

1. Assume that the algorithm has been able to perform k steps (without stopping). Prove that
   i) the balls $\left\{B_{\mathbb{R}^n}(x_j, \varepsilon/2) : j = 1, \ldots, k\right\}$ are disjoint;
   ii) $\bigcup_{j=1,\ldots,k} B_{\mathbb{R}^n}(x_j, \varepsilon/2) \subset B_{\mathbb{R}^n}(0, 1+\varepsilon/2)$.
2. By comparing the volume of the balls $B_{\mathbb{R}^n}(x, \varepsilon/2)$ and $B_{\mathbb{R}^n}(0, 1+\varepsilon/2)$, prove that
$$k \leq \left(1 + \frac{2}{\varepsilon}\right)^n.$$
3. Prove that $\mathcal{N}_\varepsilon$ is an ε-net of $\partial B_{\mathbb{R}^n}(0,1)$ and that its cardinality is upper-bounded by $\left(1 + \frac{2}{\varepsilon}\right)^n$.

6.4.3 Limit distribution of singular values of random matrices

As the singular values of E and E^T are the same, we can assume with no loss of generality that $p \geq n$. We consider the case where the entries E_{ij} are i.i.d., centered, with variance 1. When n, p go to infinity with the limiting ratio $n/p \to \beta \leq 1$, the empirical distribution
$$\frac{1}{n} \sum_{k=1}^{n} \delta_{p^{-1/2}\sigma_k(E)}(x)$$
of the singular values of the matrix $p^{-1/2}E \in \mathbb{R}^{n \times p}$ converges almost surely to the the Marchenko-Pastur distribution [**?**], which has a density on $[1-\sqrt{\beta}, 1+\sqrt{\beta}]$ given by
$$f_\beta(x) = \frac{1}{\pi \beta x} \sqrt{\left(x^2 - \left(1-\sqrt{\beta}\right)^2\right)\left(\left(1+\sqrt{\beta}\right)^2 - x^2\right)}. \tag{6.15}$$

This classical result of random matrix theory is illustrated in Figure 6.1.

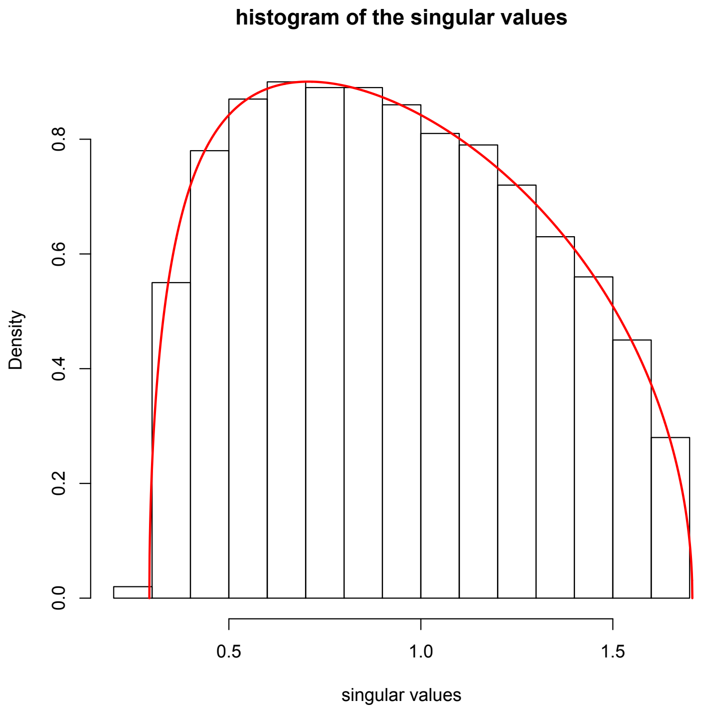


Figure 6.1: Plot of the histogram of the singular values of $p^{-1/2}E$ and of the Marchenko-Pastur distribution (6.15) for $\beta = 1/2$ (in red).

You can reproduce this plot with the following R-code.
The first step is to download the R software at `https://cran.r-project.org`.
Then enter the next lines of R code.

```
# Generate the matrix E
n <- 1000
beta <- 1/2
p <- n/beta
E <- matrix(rnorm(n*p),ncol=n)

# Plot histogram of the singular values
hist(svd(E,nu=0,nv=0)$d/sqrt(p),freq=FALSE,xlab="singular values")

# Superimpose the Marchenko-Pastur distribution
x <- 1-sqrt(beta)+(0:p)/p*2*sqrt(beta)
f <- sqrt((x**2-(1-sqrt(beta))**2)*((1+sqrt(beta))**2-x**2))/(pi*beta*x)
points(x,f,type="l",col=2,lwd=3)
```

Chapter 7

Principal Component Analysis

In many cases, we have some observations $X^{(1)}, \ldots, X^{(n)}$ which are in a space $\mathbb{R}^p$ of high-dimension p. Dealing with high-dimensional observations is an issue for two reasons. First, high-dimensional data come with a high level of fluctuations (this phenomenon is known as the curse of dimensionality), so classical estimation procedures fail in this context. Second, numerical computing with high-dimensional data is very resource intensive. A solution to bypass these issues is to perform dimension reduction. The goal of dimension reduction is to represent data in a lower dimensional space, with a minimum of distorsion. The most simple dimension reduction technique is the Principal Component Analysis (PCA). This technique, which is the subject of this chapter, is one of the most widely used method in data analysis. As we will see, the Principal Component Analysis (PCA) is tightly linked to the Singular Value Decomposition (SVD) introduced in Chapter 5.

7.1 Principal Component Analysis

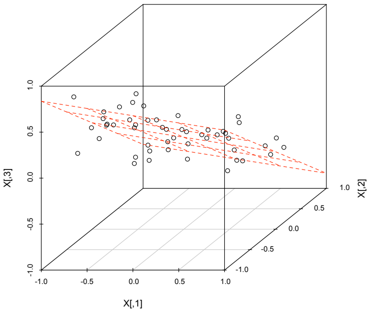
7.1.1 Finding the best low dimensional linear representation of data

The principle of Principal Component Analysis (PCA) is to seek for a linear span $\widehat{V}_d$ in $\mathbb{R}^p$, with a prescribed dimension d such that the data point $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ are as close as possible to their projection on $\widehat{V}_d$.

For data points $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ and any dimension $d \leq p$, the PCA computes the linear span $\widehat{V}_d$ in $\mathbb{R}^p$ minimizing

$$\widehat{V}_d \in \underset{V:\dim(V)\leq d}{\operatorname{argmin}} \sum_{i=1}^{n} \|X^{(i)} - \mathrm{Proj}_V X^{(i)}\|^2, \tag{7.1}$$

where the minimum is over all the subspaces $V \subset \mathbb{R}^p$ with dimension at most d and Proj_V is the orthogonal projection matrix onto the linear span V.



$\widehat{V}_2$ in dimension $p = 3$.

Let us stack the data $X^{(1)}, \ldots, X^{(n)}$ into a $n \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} (X^{(1)})^T \\ \vdots \\ (X^{(n)})^T \end{pmatrix} \in \mathbb{R}^{n\times p},$$

and let us denote by $\mathbf{X} = \sum_{k=1}^{r} \sigma_k u_k v_k^T$ a SVD of $\mathbf{X}$ with the usual convention $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.

Theorem 7.1 PCA algorithm.
The solution to (7.1) is the linear span $\widehat{V}_d = \text{span}\,\{v_1, \ldots, v_d\}$.

In addition, the coordinates of $\text{Proj}_{\widehat{V}_d} X^{(i)}$ in the orthonormal basis $(v_1, \ldots, v_d)$ of $\widehat{V}_d$ are given by $(c_1(i), \ldots, c_d(i))$, where $c_k(i)$ denotes the i-th entry of the vector $c_k := \sigma_k u_k \in \mathbb{R}^n$.

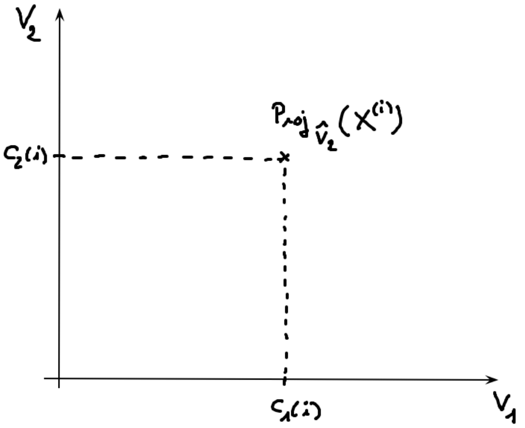


Figure 7.1: The data point $X^{(i)}$ projected on $\widehat{V}_2$, represented in the axes v_1, v_2.

Comments.

1. We observe that performing a PCA only amounts to compute the d first terms of the SVD of $\mathbf{X}$.
2. The projection $\text{Proj}_{\widehat{V}_d} X^{(i)} \in \mathbb{R}^p$ lies in $\widehat{V}_d$, but it is still a vector in $\mathbb{R}^p$, hence with p coordinates. In order to handle a vector with only d coordinates, we must work with the d-tuple $(c_1(i), \ldots, c_d(i)) \in \mathbb{R}^d$ of the coordinates of the projection on the orthonormal basis $\{v_1, \ldots, v_d\}$.
3. Since $\widehat{V}_d$ is a linear span and not an affine span, it is highly recommended to first center the data points
$$\widetilde{X}^{(i)} = X^{(i)} - \frac{1}{n}\sum_{j=1}^{n} X^{(j)}$$
and then proceed with a PCA on the $\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(n)}$.

Terminology: The right-singular vectors $v_1, \ldots, v_r$ are called the principal axes. The vectors $c_k = \sigma_k u_k$ for $k = 1, \ldots, r$ are called the principal components. As $\mathbf{X} v_k = \sigma_k u_k$, the principal component c_k is obtained as the image of v_k by the matrix $\mathbf{X}$. We emphasize that the principal axes are orthonormal and the principal components are orthogonal.

Proof of Theorem 7.1. To start with, we observe that
$$\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2 = \|\mathbf{X} - \mathbf{X}\,\text{Proj}_V\|_F^2.$$

For any linear span V of dimension d, the rank of the matrix $\mathbf{X}\,\text{Proj}_V$ is not larger than d, so according to Theorem 5.9 in Chapter 5, page 65,
$$\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2 = \|\mathbf{X} - \mathbf{X}\,\text{Proj}_V\|_F^2 \geq \min_{\text{rank}(B)\leq d} \|\mathbf{X} - B\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2. \tag{7.2}$$

Furthermore, for $\widehat{V}_d = \text{span}\{v_1, \ldots, v_d\}$, we have

$$\mathbf{X}\,\text{Proj}_{\widehat{V}_d} = \sum_{k=1}^{r} \sigma_k u_k v_k^T \sum_{j=1}^{d} v_j v_j^T = \sum_{k=1}^{d} \sigma_k u_k v_k^T.$$

So

$$\|\mathbf{X} - \mathbf{X}\,\text{Proj}_{\widehat{V}_d}\|_F^2 = \Big\| \sum_{k=d+1}^{r} \sigma_k u_k v_k^T \Big\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2. \tag{7.3}$$

Comparing (7.2) and (7.3), we find that $\widehat{V}_d = \text{span}\{v_1, \ldots, v_d\}$ is solution to (7.1).

In addition, the coordinate of $\text{Proj}_{\widehat{V}_d} X^{(i)}$ over v_k is obtained by taking the scalar product between $X^{(i)}$ and v_k,

$$\langle X^{(i)}, v_k\rangle = \langle \mathbf{X}^T e_i, v_k\rangle = \langle e_i, \mathbf{X} v_k\rangle = \sigma_k \langle e_i, u_k\rangle,$$

where we used for the last equality that $\mathbf{X} v_k = \sigma_k u_k$. Hence, the coordinates of $\text{Proj}_{\widehat{V}_d} X^{(i)}$ in the orthonormal basis $(v_1, \ldots, v_d)$ of $\widehat{V}_d$ are given by $(c_1(i), \ldots, c_d(i))$, where $c_k := \sigma_k u_k$. □.

7.1.2 Illustration

PCA is a popular and powerful dimension reduction technique. Let us illustrate PCA with a visual example based on the Mixed National Institute of Standards and Technology (MNIST) data set [**?**], which gathers 1100 scans of each digit. Each scan is a 16×16 image, which can be encoded as a vector in $\mathbb{R}^{256}$. The Figure 7.2 illustrates the compressed images, when they are projected on the linear span $\widehat{V}_{10}$ output by PCA with $d = 10$.

Let us describe this example with more details. Let us focus on a single digit, say 8. The preliminary step is to center each image $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^{256}$ of the digit 8 according to

$$\widetilde{X}^{(i)} = X^{(i)} - \frac{1}{n}\sum_{j=1}^{n} X^{(j)}.$$

Then, we proceed with a PCA on the matrix

$$\widetilde{\mathbf{X}} = \begin{pmatrix} (\widetilde{X}^{(1)})^T \\ \vdots \\ (\widetilde{X}^{(n)})^T \end{pmatrix} = \sum_{k=1}^{\tilde{r}} \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^T,$$

by computing $\widehat{V}_{10} = \text{span}\{\widetilde{v}_1, \ldots \widetilde{v}_{10}\}$. The vectors $\text{Proj}_{\widehat{V}_{10}}(\widetilde{X}^{(1)}), \ldots, \text{Proj}_{\widehat{V}_{10}}(\widetilde{X}^{(n)}) \in \mathbb{R}^{256}$ give the best approximation of $\widetilde{\mathbf{X}}^{(1)}, \ldots, \widetilde{\mathbf{X}}^{(n)}$ by a projection on a linear span of dimension 10. To obtain the final compressed images, we de-center the images

$$\text{compressed}(X^{(i)}) = \text{Proj}_{\widehat{V}_{10}}(\widetilde{X}^{(i)}) + \frac{1}{n}\sum_{j=1}^{n} X^{(j)}, \quad i = 1, \ldots, n. \tag{7.4}$$

These compressed images, together with the original images are plotted in Figure 7.2. While we observe a loss in the compression, the digit can still be identified. The benefit of the compression is that each compressed image is now described with only 10 parameters.

We emphasize that this example is for visual illustration only. In practice, there are some more powerful algorithms for image compression based on discrete Fourier or wavelets transforms (jpeg, jpeg2000, etc).

PCA in action

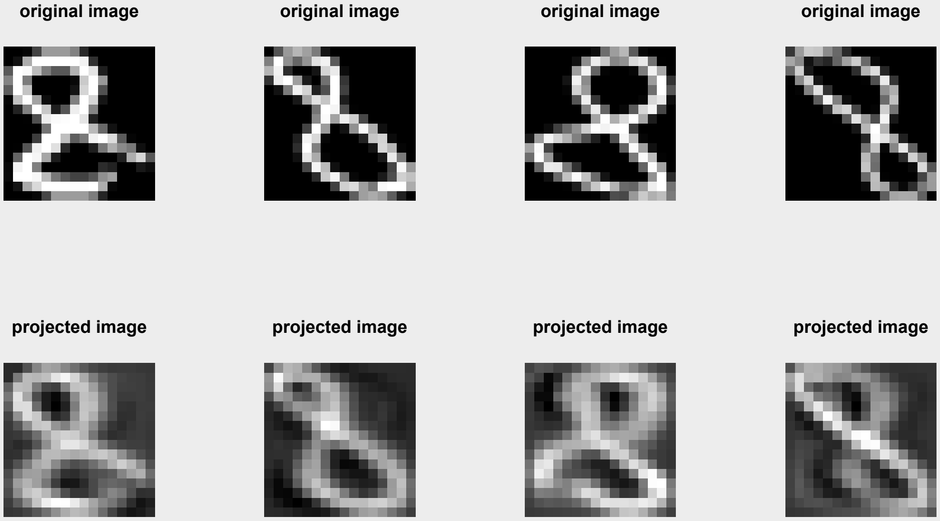


Figure 7.2: Projection on $\widehat{V}_{10}$ of some scans of the digit 8.

The Mixed National Institute of Standards and Technology (MNIST) data set [**?**] gathers 1100 scans of each digit. Each scan is a 16×16 image that is encoded by a vector in $\mathbb{R}^{256}$. The above pictures represent the projections of four images onto the space $\widehat{V}_{10}$, computed according to (7.4) from the 1100 scans of the digit 8 in the MNIST database. The original images are displayed in the first row, with their projection onto $\widehat{V}_{10}$ in the second row.

7.2 Interpreting PCA

Formulas for projection

Let us denote by Z the projected data: $Z^{(i)} = \mathrm{Proj}_{\widehat{V}_d} X^{(i)}$, for $i = 1, \ldots, n$. We have seen that

$$Z_a^{(i)} = \sum_{k=1}^{d} c_k(i) v_k(a) = \langle c(i), v(a) \rangle,$$

where $c(i) = (c_1(i), \ldots, c_d(i))$ and $v(a) = (v_1(a), \ldots, v_d(a))$.

Let us denote by $X_a = (X_a^{(1)}, \ldots, X_a^{(n)}) \in \mathbb{R}^n$ the vector gathering the observations for the variable a. Similarly as for the individual points $X^{(1)}, \ldots, X^{(n)}$, we may wish to project the variables $X_1, \ldots, X_p$ onto a linear span of dimension d. To do so, we only need to replace the matrix $\mathbf{X}$ by its transpose

$$\mathbf{X}^T = \sum_k \sigma_k v_k u_k^T,$$

and apply PCA to $\mathbf{X}^T$. Theorem 7.1 ensures that the best possible approximation space is $\widehat{U}_d = \mathrm{span}\,\{u_1, \ldots, u_d\}$ and

$$\mathrm{Proj}_{\widehat{U}_d} X_a = \sum_{k=1}^{d} \sigma_k v_k(a) u_k = \sum_{k=1}^{d} v_k(a) c_k.$$

Hence, $v(a)$ represents the coordinates of $\text{Proj}_{\widehat{U}_d} X_a$ in the orthogonal (but not orthonormal!) basis $(c_1, \ldots, c_d)$ of $\widehat{U}_d$.

PCA can be performed for two different purposes: reducing the dimension before further statistical analysis (as with the MNIST data set), or visualizing the data (as in the next heptathlon example).

Dimension reduction

When the goal is to reduce the dimension, then emerges the question of choosing d. From the proof of Theorem 7.1, we get the following measure of the quality of approximation

$$\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_{\widehat{V}_d} X^{(i)}\|^2 = \|\mathbf{X} - \mathbf{X}\,\text{Proj}_{\widehat{V}_d}\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2.$$

Hence, in order to evaluate the fraction of variance not explained by the projection on the d first principal axes, we only have to look at the ratio

$$\frac{\|\mathbf{X} - \mathbf{X}\,\text{Proj}_{\widehat{V}_d}\|_F^2}{\|\mathbf{X}\|_F^2} = \frac{\sum_{k=d+1}^{r} \sigma_k^2}{\sum_{k=1}^{r} \sigma_k^2} = 1 - \frac{\sum_{k=1}^{d} \sigma_k^2}{\sum_{k=1}^{r} \sigma_k^2}.$$

Accordingly, it is classical to plot the square singular values $\sigma_1^2 \geq \sigma_2^2, \ldots$ and look for an "elbow" in the plot. We then choose d corresponding to this elbow: the fraction of unexplained variance decreases fast before this elbow and more slowly after it. In the Section 7.3, we provide some theoretical choices of d when the signal can be decomposed as a signal part and a Gaussian noise part.

Data visualization

It is hard to visualize data points in a high-dimensional space. PCA is frequently used for this purpose. When the goal is to visualize data points, we choose $d = 2$ (possibly $d = 3$) and we represent the cloud of points $X^{(1)}, \ldots, X^{(n)}$, by their projection on $\widehat{V}_2$. More precisely, for $i = 1, \ldots, n$, we plot the vector $c(i)$ of coordinates of $Z^{(i)} = \text{Proj}_{\widehat{V}_2} X^{(i)}$ on the orthonormal basis $\{v_1, v_2\}$ of $\widehat{V}_2$. We can then observe the repartition of the data points: do we see some "clusters" or some "outliers", or some other patterns?

It is also important to compare the norm of $Z^{(i)}$ and the norm of $X^{(i)}$, in order to check if the point i is well represented by its projection on $\widehat{V}_2$. If the ratio $\|c(i)\|/\|X^{(i)}\|$ is smaller than, say 0.8, then the point i is not well represented by $c(i)$.

We can also visualize the variables by plotting their projection on $\widehat{U}_2$. It is interesting to note that

$$\langle \text{Proj}_{\widehat{U}_2} X_a, \text{Proj}_{\widehat{U}_2} X_b \rangle = \sum_{k=1}^{2} \sigma_k^2 v_k(a) v_k(b),$$

so we can plot the vectors $[\sigma_k v_k(a)]_{k=1,2}$ and $[\sigma_k v_k(b)]_{k=1,2}$ in order to visualize the correlations between the variables a and b. Again, it is good to check if the ratio $\|\text{Proj}_{\widehat{U}_2} X_a\|/\|X_a\|$ is close to one, in order to trust or not the visualization of the variable a.

A popular plot is the so called biplot of c and v, where we plot simultaneous the $c(i)$ and $v(a)$. In this case $c(i)$ represents the projection of the data point i and $v(a)$ represents the projection of the variable a. We emphasize that there is a distorsion in the representation of the variable a, as $v(a)$ corresponds to the coordinates of $\text{Proj}_{\widehat{U}_2} X_a$ in the orthogonal, but not orthonormal basis $\{c_1, c_2\}$ of $\widehat{U}_2$. We can observe on a biplot the correlation between individuals and variables. As the projection $Z^{(i)} = \text{Proj}_{\widehat{V}_2} X^{(i)}$ of $X^{(i)}$ on $\widehat{V}_2$ is given by $Z_a^{(i)} = \langle c(i), v(a) \rangle$, we can visualize on the biplot the size of the entry $Z_a^{(i)}$ by looking at the scalar product $\langle c(i), v(a) \rangle$. If $c(i)$ and $v(a)$ are well aligned, then the entry $Z_a^{(i)}$ will be large, while if $c(i)$ and $v(a)$ are orthogonal, then the entry $Z_a^{(i)}$ will be small.

Example: heptathlon data set

Let us illustrate PCA on a second example. The R-code for analyzing this example is given in Section 7.4.2.

The results of the heptathlon event at the 1988 Seoul Olympic Games are displayed in the table below.

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02
Hautenauve (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43

This table provides the scores of $n = 25$ athletes at $p = 7$ disciplines

1. hurdles: scores 100m hurdles.
2. highjump: scores high jump.
3. shot: scores shot.
4. run200m: scores 200m race.
5. longjump: scores long jump.
6. javelin: scores javelin.
7. run800m: scores 800m race.

The athletes are ranked according to their final combined score (Joyner-Kersee won the gold medal).

Here, the scores of the different disciplines are incommensurables. In order to get scores in the same scale, the data points are first centered

$$X_{ia} \leftarrow X_{ia} - \frac{1}{n}\sum_{j=1}^{n} X_{ja},$$

and then normalized to have unit norm

$$X_a \leftarrow X_a/\|X_a\|.$$

To start with, we can plot the square singular values $\sigma_1^2 \geq \ldots \geq \sigma_r^2$, and check that we will catch most of the variance with $d = 2$.

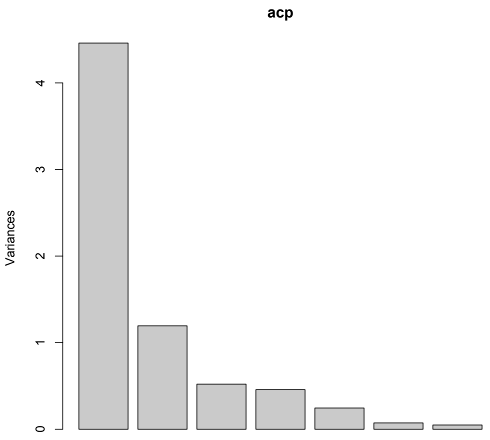


Figure 7.3: Plot of the square singular values $\sigma_1^2, \ldots, \sigma_r^2$

Then, we can check if the variables are well approximated by plotting $\text{Proj}_{\widehat{U}_2} X_a / \|X_a\|$. These vectors are close to the unit circle, so each discipline is well represented by the projection on $\widehat{U}_2$.

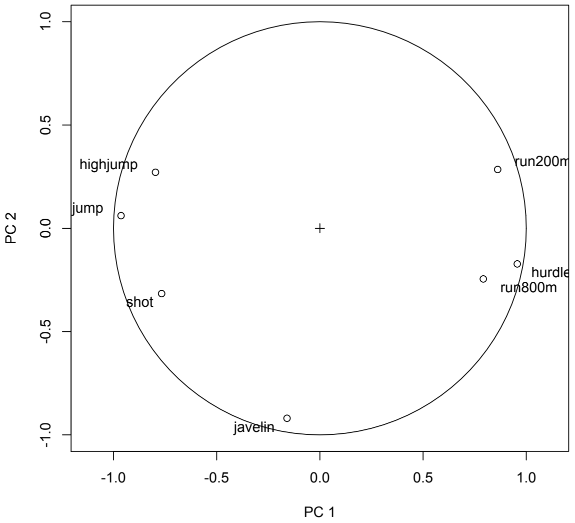


Figure 7.4: Plot of the vectors $\text{Proj}_{\widehat{U}_2} X_a / \|X_a\|$ relative to the unit circle.

Finally, we draw the biplot of the athletes (in black) and the disciplines (in red).

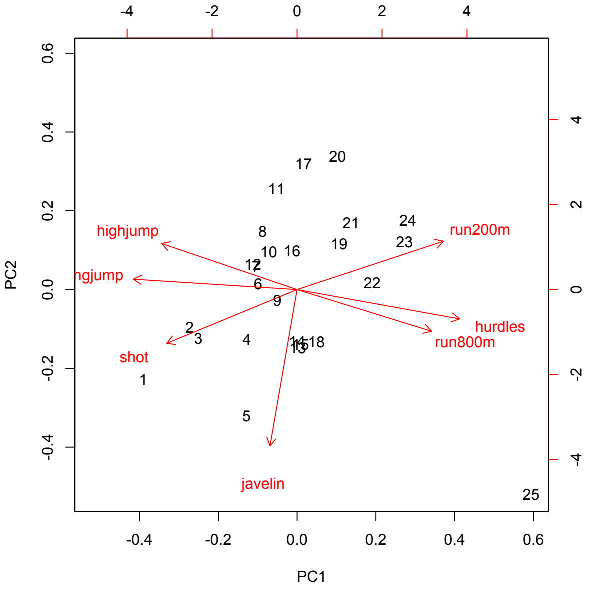


Figure 7.5: Biplot of the athletes (in black) and the disciplines (in red). Each athlete is represented with is final rank.

You can observe that the disciplines where you should have the smallest possible score (run200m, hurdles, run800m) are in the opposite side of the disciplines where you should have the largest possible score (highjump, longjump, shot). The javelin discipline looks quite orthogonal to all the others. This may be due to the strong technical nature of this discipline. You can also observe that the discipline requiring powerful qualities (run200m and shot) are well aligned, as well as those corresponding to more lanky athletes (highjump, longjump, hurdles, run800m).

7.3 Theory for PCA

7.3.1 Recovering a low dimensional signal

PCA makes sense if the data points $X^{(1)}, \ldots, X^{(n)}$ lie in the vicinity of a d-dimensional space. In many cases the data points $X^{(i)} = b^{(i)} + \varepsilon^{(i)}$ can be decomposed as a component $b^{(i)}$ lying in a low dimensional space plus some fluctuation $\varepsilon^{(i)}$. The existence of a low-dimensional component $b^{(i)}$ is related to the physical nature of the data. For example, pictures can be represented in lower dimensions (compression) due to the geometric structures in images; social or economical variables can be represented in low-dimension due to the strong social and economical structures relating the different variables; biological data reflect the biological networks producing them, etc. Most of the time, the low-dimensional component $b^{(i)}$ is the signal of interest, and the goal is to recover it.

The decomposition $X^{(i)} = b^{(i)} + \varepsilon^{(i)}$ for $i = 1, \ldots, n$, gives rise to the decomposition $\mathbf{X} = B + E$, where the i-th rows of B and E are given by $(b^{(i)})^T$ and $(\varepsilon^{(i)})^T$, respectively. Let us consider the SVD of $\mathbf{X}$ and B

$$\mathbf{X} = \sum_{k=1}^{\widehat{r}} \widehat{\sigma}_k \widehat{u}_k \widehat{v}_k^T, \quad \text{and} \quad B = \sum_{k=1}^{r} \sigma_k u_k v_k^T.$$

We have set some "hats" on the SVD of $\mathbf{X}$ in order to emphasize that these quantities can be computed from the data $\mathbf{X}$, while the matrix B and its SVD are not observed. We estimate B by the projection of the data $\mathbf{X}$ on the d first principal axes

$$\widehat{B}_d = \mathbf{X}\,\mathrm{Proj}_{\widehat{V}_d} = \sum_{k=1}^{d} \widehat{\sigma}_k \widehat{u}_k \widehat{v}_k^T. \tag{7.5}$$

Next theorem provides a bound on the estimation error $\|\widehat{B}_d - B\|_F^2$ in terms of the operator norm $|E|_{\mathrm{op}}$ of the fluctuations and in terms of the best possible approximation error of B by a matrix of rank d (see Theorem 5.9, page 65)

$$\min_{M:\mathrm{rank}(M)\leq d} \|B-M\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2.$$

Theorem 7.2 Recovering low rank component.
The estimator $\widehat{B}_d$ fulfills the error bound

$$\|\widehat{B}_d - B\|_F^2 \leq 9\left(d\,|E|_{\mathrm{op}}^2 + \sum_{k=d+1}^{r} \sigma_k^2\right). \tag{7.6}$$

Let us illustrate this result, by considering the case where the entries E_{ij} of the $n \times p$ matrix E are i.i.d. Gaussian with $\mathcal{N}(0,\sigma^2)$ distribution. Then, we have the next result which directly follows from Theorem 7.2 and Corollary 6.9, page 74.

Corollary 7.3 Bound for Gaussian noise.
For any $L > 0$, when the entries of the matrix $E \in \mathbb{R}^{n\times p}$ are i.i.d. Gaussian with $\mathcal{N}(0,\sigma^2)$ distribution, we have with probability at least $1-e^{-L}$

$$\|\widehat{B}_d - B\|_F^2 \leq 9d\left(\sqrt{p} + 7\sqrt{n+L}\right)^2 \sigma^2 + 9\sum_{k=d+1}^{r} \sigma_k^2.$$

We observe that in the setting of Corollary 7.3, we have

$$\mathbb{E}\left[\|\mathbf{X}-B\|_F^2\right] = \mathbb{E}\left[\|E\|_F^2\right] = np\sigma^2,$$

which is much larger than $d(p+n)\sigma^2$ if $d \ll n \wedge p$. So, when B is approximately of rank d with $d \ll n \wedge p$, there is a substantial gain in using $\widehat{B}_d$ instead of $\mathbf{X}$ in order to estimate B.

Proof of Theorem 7.2. Before starting the proof, we remind the reader two useful inequalities.

Lemma 7.4 *For any $a > 0$, and $x, y \in \mathbb{R}^n$, we have*

$$2\langle x, y\rangle \leq a\|x\|^2 + a^{-1}\|y\|^2, \tag{7.7}$$
$$\|x+y\|^2 \leq (1+a)\|x\|^2 + (1+a^{-1})\|y\|^2. \tag{7.8}$$

The Inequality (7.8) immediately follows from (7.7) and the Inequality (7.7) follows from

$$a\|x\|^2 + a^{-1}\|y\|^2 - 2\langle x, y\rangle = \|a^{1/2}x - a^{-1/2}y\|^2 \geq 0.$$

Let us prove now (7.6). We denote by

$$B_d = \sum_{k=1}^{d} \sigma_k u_k v_k^T,$$

the best approximation of B by a matrix of rank d (see Theorem 5.9, page 65). As $\widehat{B}_d$ is the best approximation of $\mathbf{X}$ by a matrix of rank d, we have

$$\|\mathbf{X}-\widehat{B}_d\|_F^2 \leq \|\mathbf{X}-B_d\|_F^2.$$

Using the decomposition $\mathbf{X}=B+E$ and expanding the squares, the previous inequality is equivalent to

$$\|B-\widehat{B}_d\|_F^2 \leq \|B-B_d\|_F^2 + 2\langle E, \widehat{B}_d - B_d\rangle_F. \tag{7.9}$$

We observe that $\operatorname{rank}(\widehat{B}_d - B_d) \leq 2d$, so according to Lemma 5.8, we can upper-bound the scalar product $\langle E, \widehat{B}_d - B_d\rangle_F$ in terms of the $(2,2d)$-Ky-Fan norm

$$2\langle E, \widehat{B}_d - B_d\rangle_F \leq 2\|E\|_{(2,2d)}\|\widehat{B}_d - B_d\|_{(2,2d)} = 2\|E\|_{(2,2d)}\|\widehat{B}_d - B_d\|_F.$$

Applying Inequality (7.7) with $a=5/2$, then Inequality (7.8) with $a=1/9$, and finally $\|E\|_{(2,2d)}^2 \leq 2d\,|E|_{\text{op}}$, we get

$$\begin{aligned}
2\langle E, \widehat{B}_d - B_d\rangle_F &\leq \frac{5}{2}\|E\|_{(2,2d)}^2 + \frac{2}{5}\|\widehat{B}_d - B_d\|_F^2\\
&\leq \frac{5}{2}\|E\|_{(2,2d)}^2 + \frac{2}{5}\Big((10/9)\|\widehat{B}_d - B\|_F^2 + 10\|B-B_d\|_F^2\Big)\\
&\leq 5d\,|E|_{\text{op}}^2 + \frac{4}{9}\|\widehat{B}_d - B\|_F^2 + 4\|B-B_d\|_F^2.
\end{aligned}$$

Combining this last inequality with (7.9) and

$$\|B-B_d\|_F^2 = \Big\|\sum_{k=d+1}^{r}\sigma_k u_k v_k^T\Big\|_F^2 = \sum_{k=d+1}^{r}\sigma_k^2,$$

we get (7.6). □

7.3.2 Dimension selection

The Bound (7.6) can be used in order to propose a choice for d with theoretical garanties. Indeed, let us notice that the Bound (7.6) can be written as

$$\|\widehat{B}_d - B\|_F^2 \leq 9\left(\sum_{k=1}^{d}(|E|_{\text{op}}^2 - \sigma_k^2) + \sum_{k=1}^{r}\sigma_k^2\right). \tag{7.10}$$

As the second term is independent of d, the integer d minimizing the right-hand side of (7.10) is the integer d minimizing the first sum. As $\sigma_1^2 \geq \sigma_2^2 \geq \ldots$, the first sum is decreasing as long as $\sigma_d^2 \geq |E|_{\text{op}}^2$ and then it increases. Hence in order to minimize the right-hand side of (7.6), the best is to choose the dimension

$$d^* := \max\left\{k : \sigma_k \geq |E|_{\text{op}}\right\}. \tag{7.11}$$

While the operator norm $|E|_{\text{op}}$ can be evaluated in some cases, for example with Corollary 6.9 page 74, the singular values $\sigma_1 \geq \sigma_2 \geq \ldots$ of B are not observed, so we cannot directly use (7.11). Yet, combining Weyl Inequality (6.1), page 69, with the Bound (7.10), we get

$$\|\widehat{B}_d - B\|_F^2 \leq 9\left(\sum_{k=1}^{d}\Big(|E|_{\text{op}}^2 - (\widehat{\sigma}_k - |E|_{\text{op}})_+^2\Big) + \sum_{k=1}^{r}\sigma_k^2\right),$$

with $\widehat{\sigma}_1 \geq \widehat{\sigma}_2 \geq \ldots$ the singular values of $\mathbf{X}$. Following the same reasoning as before, with σ_k replaced by $(\widehat{\sigma}_k - |E|_{\text{op}})_+$, we get the next selection rule for d.

Corollary 7.5 Dimension selection.
For $\widehat{d} = \max\left\{k : \widehat{\sigma}_k \geq 2\,|E|_{\mathrm{op}}\right\}$*, we have*

$$\|\widehat{B}_{\widehat{d}} - B\|_F^2 \leq 9^2 \min_{d\geq 0}\left\{d\,|E|_{\mathrm{op}}^2 + \sum_{k=d+1}^{r} \sigma_k^2\right\}.$$

As before, in the case where the entries E_{ij} of the matrix E are i.i.d. Gaussian with $\mathcal{N}(0,\sigma^2)$ distribution, we can set $\tilde{d} = \max\left\{k : \widehat{\sigma}_k \geq 2\left(\sqrt{p} + 7\sqrt{2n}\right)\sigma\right\}$ and then get with probability at least $1 - e^{-n}$

$$\|\widehat{B}_{\tilde{d}} - B\|_F^2 \leq 9^2 \min_{d\geq 0}\left\{d\left(\sqrt{p} + 7\sqrt{2n}\right)^2\sigma^2 + \sum_{k=d+1}^{r} \sigma_k^2\right\}.$$

Proof of Corollary 7.5. According to Weyl Inequality (6.1) and the definition of $\widehat{d}$, we have

$$\sigma_{\widehat{d}} \geq \widehat{\sigma}_{\widehat{d}} - |E|_{\mathrm{op}} \geq |E|_{\mathrm{op}},$$

so $\widehat{d} \leq d^*$, with d^* defined by (7.11). In addition, Weyl Inequality (6.1) and the definition of $\widehat{d}$ also ensure that for any $k \in [\widehat{d}+1, d^*]$, we have

$$\sigma_k \leq \widehat{\sigma}_k + |E|_{\mathrm{op}} \leq 3\,|E|_{\mathrm{op}},$$

where the second inequality comes from $k \geq \widehat{d}+1$ and the definition of $\widehat{d}$. So, the Bound (7.6) gives

$$\begin{aligned}
\|\widehat{B}_{\widehat{d}} - B\|_F^2 &\leq 9\left(\widehat{d}\,|E|_{\mathrm{op}}^2 + \sum_{k=\widehat{d}+1}^{d^*} \sigma_k^2 + \sum_{k=d^*+1}^{r} \sigma_k^2\right)\\
&\leq 9\left(\widehat{d}\,|E|_{\mathrm{op}}^2 + 9(d^* - \widehat{d})\,|E|_{\mathrm{op}}^2 + \sum_{k=d^*+1}^{r} \sigma_k^2\right)\\
&\leq 9^2\left(d^*\,|E|_{\mathrm{op}}^2 + \sum_{k=d^*+1}^{r} \sigma_k^2\right) = 9^2 \min_{d\geq 0}\left\{d\,|E|_{\mathrm{op}}^2 + \sum_{k=d+1}^{r} \sigma_k^2\right\},
\end{aligned}$$

where the last equality follows from the definition (7.11) of d^*. □

7.4 Exercises

7.4.1 Rank recovery

Let us consider the setting of Section 7.3.1 and let us assume that the rank of B is d. As in Corollary 7.5, let us set for some $\lambda > 0$

$$\widehat{d} = \max\{k : \widehat{\sigma}_k \geq \lambda\}. \tag{7.12}$$

In this exercise, we will give some conditions ensuring that $\widehat{d} = d$ with large probability.

1. Check that
$$\mathbb{P}\left[\widehat{d} \neq d\right] = \mathbb{P}\left[\widehat{\sigma}_{d+1} \geq \lambda \text{ or } \widehat{\sigma}_d < \lambda\right].$$

2. With Weyl Inequality (6.1), page 69, prove that
$$\mathbb{P}\left[\widehat{d} \neq d\right] \leq \mathbb{P}\left[|E|_{\mathrm{op}} \geq \lambda \wedge (\sigma_d - \lambda)\right].$$

3. Let us assume that the entries of the matrix $E \in \mathbb{R}^{n\times p}$ are i.i.d. Gaussian with $\mathcal{N}(0, \sigma^2)$ distribution. We also assume that $\sigma_d \geq 2\left(\sqrt{p} + 7\sqrt{2n}\right)\sigma$ and set $\lambda = \left(\sqrt{p} + 7\sqrt{2n}\right)\sigma$. Prove with Corollary 6.9, page 74, that the integer $\widehat{d}$ defined by (7.12) fulfills

$$\mathbb{P}\left[\widehat{d} = d\right] \geq 1 - e^{-n}.$$

Remark. We have recovered the rank under a condition ensuring that the singular value σ_d is large enough. As $\sigma_{d+1} = 0$, such a condition ensures that there is a large enough gap between 0 and the non-zero singular values of B. Such a gap condition is unavoidable for rank recovery. Indeed, if σ_d is small, much smaller than the fluctuations of $\widehat{\sigma}_d - \sigma_d$, then no procedure can detect if σ_d is or not 0.

7.4.2 Implementing a PCA with R

You can implement a PCA on the Heptathlon example, with the following R-code.
The first step is to download the R software at `https://cran.r-project.org`.
Then, you can enter the following code to download and analyze the data.

```
# download heptathlon dataset
data("heptathlon", package = "HSAUR")
# display the dataset
heptathlon

# plot square singular values
pca <-prcomp(heptathlon[,1:7], scale = TRUE)
plot(pca)

# display principal axes
pca$rotation

# display principal components
pca$x

# biplot
 biplot(pca,xlabs=1:25)
```

Chapter 8

Clustering

In a large fraction of data analysis methodologies, the data are considered as homogeneous: all the observations are assumed to be distributed according to a common statistical model. Such an assumption is valid for data coming from small scale controlled experiments, but it is highly unrealistic at the era of "big data", where data come from multiple sources. A recipe for dealing with such inhomogeneous data, is to consider them as an assemblage of several homogeneous data sets, corresponding to homogeneous "subpopulations". Then each subpopulation can be treated either independently or jointly. The main hurdle in this approach is to recover the unknown subpopulations, which is the main goal of clustering algorithms.

8.1 Cluster model

Assume that we have n observations $X_1, \ldots, X_n \in \mathbb{R}^p$, which are independent, but not identically distributed. We denote by $\mu_i = \mathbb{E}[X_i]$ the mean of X_i and by $\Sigma_i = \mathrm{Cov}(X_i)$ the covariance of X_i. As discussed above, we assume that the distribution of the X_i is homogeneous across some subpopulations. This means that there exists an *unknown* partition $G^* = \left\{G_1^*, \ldots, G_K^*\right\}$ of $\{1, \ldots, n\}$, such that, within a group G_k^* all the means and covariances are equal.

Definition Cluster model.
We assume that

1. *the observations $X_1, \ldots, X_n \in \mathbb{R}^p$, are independent,*
2. *there exists a (minimal) partition $G^* = \left\{G_1^*, \ldots, G_K^*\right\}$ of $\{1, \ldots, n\}$ such that all the random variables $(X_i)_{i \in G_k^*}$ are identically distributed.*

In the following, we denote by $\theta_1, \ldots, \theta_K \in \mathbb{R}^p$, $\Lambda_1, \ldots, \Lambda_K \in \mathbb{R}^{p \times p}$ the vectors and covariances such that

$$\text{for all } i \in G_k^* : \quad \mathbb{E}\left[X_i\right] = \theta_k \ \text{ and } \ \mathrm{Cov}(X_i) = \Lambda_k. \tag{8.1}$$

The mean $\theta_1, \ldots, \theta_K \in \mathbb{R}^p$ are assumed to be all distinct.
As in the previous chapter, we denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the matrix whose i-th row is given by X_i. We define similarly the matrices

- $\mathbf{E} \in \mathbb{R}^{n \times p}$ the matrix whose i-th row is given by $E_i = X_i - \mu_i = X_i - \theta_k$ for $i \in G_k^*$;
- $\Theta \in \mathbb{R}^{K \times p}$ the matrix whose k-th row is given by θ_k;
- and $A \in \mathbb{R}^{n \times K}$ the membership matrix defined by $A_{ik} = \mathbf{1}_{i \in G_k^*}$, for $i = 1, \ldots, n$, and $k = 1, \ldots, K$.

Then we have the compact formula

$$\mathbf{X} = A\Theta + \mathbf{E}. \tag{8.2}$$

Remark: a popular variant of the cluster model, is the mixture model. This model has an additional generating feature compared to the cluster model: Instead of being arbitrary, the partition G^* is generated by sampling for each observation i the label of its group according to a probability distribution π on $\{1, \ldots, K\}$.

8.2 Local algorithms

The main family of local algorithms are the so-called hierarchical clustering algorithms. The hierarchical clustering algorithms cluster data points sequentially, starting from a trivial partition with n singletons (each data point is a cluster on its own) and then merging them step by step until eventually getting a single cluster with all the data points. At the end of the process, we obtain a hierarchical family of nested clusterings and the data scientist can choose her favorite one.

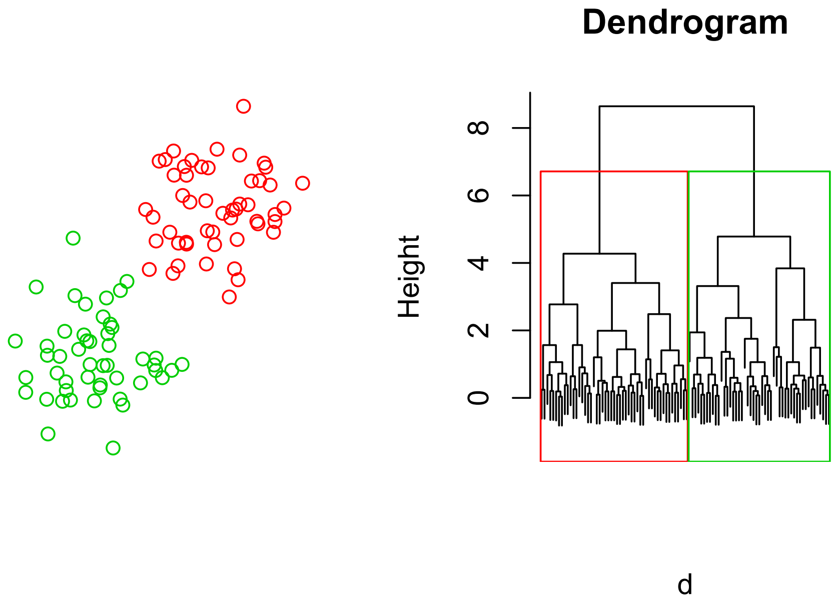


Figure 8.1: Left: data points in $\mathbb{R}^2$. Right: dendrogram of hierarchical clustering with Euclidean distance d and complete linkage ℓ. The colors correspond to the clustering output when selecting $K = 2$ clusters.

Linkage

In hierarchical clustering, the recipe for merging points is quite simple: at each step the algorithm merges the two closest clusters (in a sense to be defined) of the current clustering, letting the other clusters unchanged. This requires the definition of a ”distance” $\ell(G, G')$ between clusters G and G', usually called ”linkage”. Let $d(x, y)$ be any distance on $\mathbb{R}^p$, typically $d(x, y) = \|x - y\|$ or $d(x, y) = |x - y|_1$. Some classical examples of linkage are:

- **Single linkage:** single linkage corresponds to the smallest distance between the points of the two clusters
$$\ell_{single}(G, G') = \min \left\{ d(x_i, x_j) : i \in G,\ j \in G' \right\}.$$
Single linkage clustering tends to produce clusters looking like ”chains”, and we can have within a cluster two data points x, y with $d(x, y)$ very large.
- **Complete linkage:** complete linkage is kind of the opposite of single linkage. It corresponds to the largest distance between the points of the two clusters
$$\ell_{complete}(G, G') = \max \left\{ d(x_j, x_j) : i \in G,\ j \in G' \right\}.$$
Complete linkage clustering tends to produce ”compact” clusters where all data points are close to each other.

- **Average linkage:** average linkage corresponds to the average distance between the points of the clusters G, G'
$$\ell_{average}(G, G') = \frac{1}{|G||G'|} \sum_{i \in G,\ j \in G'} d(x_i, x_j).$$
The clustering produced by average linkage is less ”chainy” than those produced by single linkage and less compact than those produced by complete linkage.

In section 8.4.4, these features are illustrated in a randomly generated example.

Hierarchical clustering algorithm

Hierarchical clustering algorithms start from the trivial partition $G^{(n)} = \{\{1\}, \ldots, \{n\}\}$ with n clusters, and then sequentially merge clusters two by two. At each step, the algorithm merge the two clusters G, G' available at this step with the smallest linkage $\ell(G, G')$. The output is a sequence of clustering $G^{(1)}, \ldots, G^{(n)}$ with $K = 1, \ldots, n$ clusters. These clusterings are nested, in the sense that for $j \leq k$ the partition $G^{(k)}$ is a sub-partition of $G^{(j)}$.

Hierarchical clustering

- Input: data points $X_1, \ldots, X_n$ and a linkage ℓ
- initialization: $G^{(n)} = \{\{1\}, \ldots, \{n\}\}$
- iterations: for $t = n, \ldots, 2$
  - find $(\widehat{a}, \widehat{b}) \in \text{argmin}_{(a,b)}\, \ell(G_a^{(t)}, G_b^{(t)})$
  - build $G^{(t-1)}$ from $G^{(t)}$ by merging $G_{\widehat{a}}^{(t)}$ and $G_{\widehat{b}}^{(t)}$. The other clusters are let unchanged.
- Output: the n partitions $G^{(1)}, \ldots, G^{(n)}$ of $\{1, \ldots, n\}$.

Dendrogram

It is popular to represent the sequence of clustering $G^{(1)}, \ldots, G^{(n)}$ with a dendrogram, which is a tree, rooted in $G^{(1)}$, and whose leaves correspond to $G^{(n)}$. The dendrogram depicts how the merging is performed. The partition $G^{(k)}$ can be read on the dendrogram as follows, see Figure 8.1:

1. locate the level where there are exactly k branches in the dendrogram;
2. cut the dendrogram at this level in order to get k subtrees;
3. each subtree corresponds to one cluster, gathering the points corresponding to its leaves.

The height in the tree represents the distance between two clusters. A classical recipe for choosing the number k of clusters is to look for a level k where the height between two successive merges increases abruptly.

Hierarchical clustering algorithms are popular, as they are simple to understand and to visualize. When the clusters are well separated, they succeed to recover the hidden partition G^*, see Exercise 8.4.2. Yet, hierarchical clustering is based on local informations, and do not take into account global informations on the distribution of the cloud of points, especially at the first steps. As the mistakes in the first steps cannot be repaired in the following steps, it is a strong limitation for clustering in less separated case, see Section 8.4.4 for an illustration. Next section presents another recipe for clustering, based on more global informations, carried by the singular vectors of $\mathbf{X}$.

8.3 Spectral clustering

8.3.1 Spectral clustering recipe

The recipe of spectral clustering algorithms is to compute the K first principal components of $\mathbf{X}$ and then apply some basic clustering algorithms on these vectors. So, spectral clustering algorithms are simply algorithms performing a dimension reduction step (PCA) before proceeding to a clustering.

Many state of the art clustering algorithms are based on spectral clustering. In some settings, spectral clustering alone is able to provide statistically optimal clustering. In some other settings, an additional refinement step is implemented, where each observation is reclassified according to a more specialized algorithm. Hence, spectral clustering algorithm is a good and simple algorithm in order to get a primary estimation of the groups, which can then be refined if needed, by running more specialized algorithms.

Before describing the most basic version of spectral clustering, let us explain why PCA makes sense in this setting. In the clustering model (8.2), the partition G^* is encoded in the rows of the signal $A\Theta \in \mathbb{R}^{n\times p}$. Indeed, for $i \in G^*_k$, the i-th row of $A\Theta$ is given by θ_k, so all the rows belonging to a same cluster are all the same. Hence, if, instead of applying a clustering algorithm on the rows of $\mathbf{X}$, we apply a clustering algorithm on the rows of a good estimator $\widehat{B}$ of $A\Theta$, then we get a better clustering. When the partition has K clusters, the rows of $A\Theta$ are elements of the K dimensional space spanned by $\{\theta_1, \ldots, \theta_K\}$. Hence, in light of the previous chapter, it makes sense to project the data on the space $\widehat{V}_K$ spanned by the K first right singular vectors (principal axes) of $\mathbf{X}$. The coordinates of the projection $\mathrm{Proj}_{\widehat{V}_K}(X_i)$ are given by the i-th coordinates of the K first principal components $\widehat{c}_k = \widehat{\sigma}_k \widehat{u}_k$ of $\mathbf{X}$. This line of reasoning leads to the spectral clustering algorithm described below.

Spectral clustering algorithm

1. Compute the singular value decomposition $\mathbf{X} = \sum_{k\geq 1} \widehat{\sigma}_k \widehat{u}_k \widehat{v}_k^T$;
2. Extract the K first principal components

$$\widehat{C}_K = [\widehat{c}_1, \ldots, \widehat{c}_K] := [\widehat{\sigma}_1 \widehat{u}_1, \ldots, \widehat{\sigma}_K \widehat{u}_K]$$

3. Apply a clustering procedure on the rows of $\widehat{C}_K$ in order to get a partition $\widehat{G}$ of $\{1, \ldots, n\}$.

There are many possible choices of clustering procedure for the last step, for example hierarchical clustering algorithms. In the two clusters problem theoretically investigated in Section 8.3.2, the clustering procedure will simply be based on the sign of the entries of the first left-singular vector.

We observe that computing the K first left-singular vectors of $\mathbf{X}$ involves the whole matrix $\mathbf{X}$, so, contrary to hierarchical clustering, spectral clustering takes into account all points for clustering each single data point.

A variant of spectral clustering

A popular alternative to the clustering of the rows of the principal components matrix $\widehat{C}_K = [\widehat{c}_1, \ldots, \widehat{c}_K]$, is the clustering of the rows of the left-singular vectors matrix $\widehat{U}_K = [\widehat{u}_1, \ldots, \widehat{u}_K]$. Next lemma shows that this clustering also makes sense.

Lemma 8.1 *Let*

$$A\Theta = \sum_{k=1}^{K} \sigma_k u_k v_k^T$$

be a singular decomposition of $A\Theta$*, and set* $U = [u_1, \dots, u_K] \in \mathbb{R}^{n\times K}$*.*
Then, there exist $Z_1, \dots, Z_K \in \mathbb{R}^K$*, such that*

$$U_{i:} = Z_k \ \text{ for all } i \in G_k^*, \quad \text{and} \quad \|Z_k - Z_\ell\|^2 = \frac{1}{|G_k^*|} + \frac{1}{|G_\ell^*|}, \ \text{ for all } k \neq \ell.$$

Proof of Lemma 8.1. Let us define $\Delta = \text{diag}\left(|G_1^*|^{-1/2}, \dots, |G_K^*|^{-1/2}\right)$ and $B := A\Delta$. We notice that the columns of B are

$$b_k = [A\Delta]_{:k} = \left[\frac{\mathbf{1}_{i\in G_k^*}}{|G_k^*|^{1/2}}\right]_{i=1,\dots,n}.$$

In particular, the columns $b_1, \dots, b_K$ of B are orthonormal. As

$$\text{span}\{b_1, \dots, b_K\} = \text{range}(A) \supset \text{range}(A\Theta) = \text{span}\{u_1, \dots, u_K\},$$

with $b_1, \dots, b_K$ orthonormal, the projection on range(A) is given by BB^T and $BB^T u_k = u_k$ for $k = 1, \dots, K$. Hence,

$$U = [u_1, \dots, u_K] = B \underbrace{B^T U}_{=R},$$

with R an orthogonal matrix, since $R^T R = U^T BB^T U = U^T U = I_K$.
From the decomposition $U = BR = A\Delta R$, we obtain

$$U_{ij} = \sum_{\ell=1}^{K} \mathbf{1}_{i\in G_\ell^*} (\Delta R)_{\ell j} = (\Delta R)_{kj} \quad \text{for } i \in G_k^*.$$

Hence, for $i \in G_k^*$, we have $U_{i:} = Z_k$, where the vectors

$$Z_k := [\Delta R]_{k:} = |G_k^*|^{-1/2} R_{k:}, \quad \text{for} \quad k = 1, \dots, K,$$

are orthogonal with square norm $\|Z_k\|^2 = 1/|G_k^*|$, as $\|R_{k:}\|^2 = 1$. Hence,

$$\|Z_k - Z_\ell\|^2 = \frac{1}{|G_k^*|} + \frac{1}{|G_\ell^*|}, \quad \text{for} \ \ \ell \neq k.$$

The proof of Lemma 8.1 is complete. □

Debiasing spectral clustering

The left-singular vectors $\widehat{u}_1, \widehat{u}_2, \dots$ of $\mathbf{X}$ correspond to the eigenvectors of $\mathbf{XX}^T$. Computing the expectation of an entry

$$(\mathbf{XX}^T)_{ij} = (A\Theta\Theta^T A^T)_{ij} + E_i^T E_j + E_i^T(\Theta^T A^T)_j + (A\Theta)_i^T E_j,$$

we get

$$\begin{aligned}\mathbb{E}\left[(\mathbf{XX}^T)_{ij}\right] &= (A\Theta\Theta^T A^T)_{ij} + \mathbb{E}\left[E_i^T E_j\right] \\ &= (A\Theta\Theta^T A^T)_{ij} + \mathbf{1}_{i=j}\,\text{Tr}(\text{Cov}(E_i)).\end{aligned}$$

Let us denote by Γ the diagonal matrix, with entries $\Gamma_{ii} = \mathrm{Tr}(\mathrm{Cov}(E_i))$. We observe that we have in expectation

$$\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = A\Theta\Theta^T A^T + \Gamma.$$

When Γ is proportional to the identity matrix, the eigenvectors of $A\Theta\Theta^T A^T + \Gamma$ and $A\Theta\Theta^T A^T$ are the same, so the eigenvectors of $\mathbf{X}\mathbf{X}^T$ are not biased. When Γ is not proportional to the identity matrix, it is wise to reduce the bias of $\mathbf{X}\mathbf{X}^T$ by considering the eigenvalue decomposition of $\mathbf{X}\mathbf{X}^T - \widehat{\Gamma}$, for some estimator $\widehat{\Gamma}$ of Γ. Yet, unless the matrix Γ is known in advance, it is not straightforward to design a (good) estimator $\widehat{\Gamma}$ of Γ. We refer to the Exercise 8.4.3 for an example of such an estimator.

Let us sum-up the debiased spectral algorithm.

Debiased spectral clustering algorithm

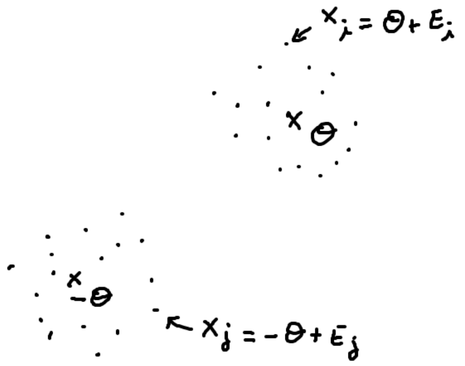
1. Compute the eigenvalue decomposition $\mathbf{X}\mathbf{X}^T - \widehat{\Gamma} = \sum_{k\geq 1} \widehat{d}_k \widehat{u}_k \widehat{u}_k^T$, with eigenvalues ranked in decreasing order;
2. Apply a clustering procedure either on the rows of $\widehat{U}_K = [\widehat{u}_1, \ldots, \widehat{u}_K]$ or on the rows of $\widehat{C}_K = \left[\widehat{d}_1^{1/2}\widehat{u}_1, \ldots, \widehat{d}_K^{1/2}\widehat{u}_K\right]$, in order to get a partition $\widehat{G}$ of $\{1, \ldots, n\}$.

8.3.2 Recovery bounds

In this section, we investigate the ability of spectral clustering to recover the partition G^* from $\mathbf{X}$. In order to avoid an inflation of technicalities, we focus on the most simple setting where there are only two groups with means symmetric with respect to 0 and Gaussian distribution. More precisely, we assume that there exists an unobserved sequence $z_1, \ldots, z_n \in \{-1, +1\}$ of binary labels such that the observations $X_1, \ldots, X_n$ are independent, and the distribution of X_i is a Gaussian distribution $\mathcal{N}(z_i\theta, \sigma^2 I_p)$ for $i = 1, \ldots, n$. Stacking as before the observations $X_1, \ldots, X_n$ into a $n \times p$ matrix $\mathbf{X}$, we then observe

$$\mathbf{X} = z\theta^T + \mathbf{E}, \tag{8.3}$$

where $z \in \{-1, +1\}^n$ and the E_{ij} are i.i.d. with a $\mathcal{N}(0, \sigma^2)$ distribution. The underlying partition is $G^* = \{\{i : z_i = 1\}, \{i : z_i = -1\}\}$.



Let us define for $x \in \mathbb{R}^n$

$$|x|_0 = \sum_{i=1}^n \mathbf{1}_{x_i \neq 0}.$$

A good clustering algorithm, is an algorithm that recovers the vector z, up to a sign change. Hence, if $\widehat{z} \in \{-1, +1\}^n$ encodes the clustering output by this algorithm, ($\widehat{z}_i = 1$ if $i \in \widehat{G}_1$ and $\widehat{z}_i = -1$ if $i \in \widehat{G}_2$), we measure the quality of the clustering by the metric

$$\text{recov}(\widehat{z}) := \frac{1}{n} \min_{\delta \in \{-1,+1\}} |z - \delta\widehat{z}|_0 , \tag{8.4}$$

which counts the proportion of mismatches between $\widehat{G}$ and G^*.

When $\mathbf{X}$ follows the Model (8.3), we have

$$\mathbb{E}\left[\mathbf{X}\mathbf{X}^T\right] = \|\theta\|^2 zz^T + \Gamma,$$

with $\Gamma_{ii} = \text{Tr}(\text{cov}(E_i))$. As all the covariances are assumed to be equal to $\sigma^2 I_p$, the matrix $\Gamma = p\sigma^2 I_n$ is proportional to the identity, and hence, we do not need to debias $\mathbf{X}\mathbf{X}^T$ in the spectral clustering algorithm. Hence, we set $\widehat{\Gamma} = 0$. Since zz^T is of rank one, we only focus on the first eigenvector $\widehat{u}_1$ of $\mathbf{X}\mathbf{X}^T$.

The first eigenvector $\widehat{u}_1$ of $\mathbf{X}\mathbf{X}^T$ does not provide a clustering of $\{1, \ldots, n\}$ into two groups and a clustering procedure is needing (second step of Spectral algorithm). One of the nice feature of Model (8.3) is that we can choose a very simple clustering procedure. Actually, as, hopefully, $\widehat{u}_1 \approx \pm z/\|z\|$, we can simply take the sign of the entries of $\widehat{u}_1$ in order to get a partition of $\{1, \ldots, n\}$ into two groups, corresponding to positive and negative entries of $\widehat{u}_1$. We consider then the following spectral clustering algorithm

$$\widehat{z} = \text{sign}(\widehat{u}_1), \quad \text{with } \widehat{u}_1 \text{ a leading eigenvector of } \frac{1}{n}\mathbf{X}\mathbf{X}^T. \tag{8.5}$$

Theorem 8.2 *Assume that* $\mathbf{X}$ *follows the model (8.3). There exists a numerical constant* $C \geq 1$ *such that, with probability at least* $1 - 2e^{-n/2}$, *the spectral clustering (8.5) fulfills the recovery bound*

$$\mathit{recov}(\widehat{z}) \leq 1 \wedge \frac{C}{s^2} , \tag{8.6}$$

with s^2 *defined by*

$$s^2 = \frac{\|\theta\|^4}{\|\theta\|^2\sigma^2 + \frac{p}{n}\sigma^4} . \tag{8.7}$$

We observe that the upper bound (8.6) is decreasing with the inverse of s^2. It is possible to show that optimal algorithms have a proportion of mismatches decreasing exponentially fast with s^2. In order to get such an optimal rate, we need to improve the spectral clustering with more refined algorithms. This refinement is out of the scope of this monograph.

The remaining of this subsection is devoted to the proof of Theorem 8.2.

Proof of Theorem 8.2.
Let us first connect the Hamming distance $|z - \delta\widehat{z}|_0$ to the square norm $\|z - \delta\sqrt{n}\widehat{u}_1\|^2$.

Lemma 8.3 *For any* $x \in \{-1, 1\}^n$ *and* $y \in \mathbb{R}^n$, *we have*

$$|x - \text{sign}(y)|_0 \leq \min_{\alpha > 0} \|x - \alpha y\|^2.$$

This lemma simply follows from the inequality

$$\mathbf{1}_{x_i \neq \text{sign}(y_i)} = \mathbf{1}_{x_i \neq \text{sign}(\alpha y_i)} \leq |x_i - \alpha y_i|^2,$$

for any $\alpha > 0$ and $i = 1, \ldots, n$.

From Lemma 8.3 with $\alpha = \sqrt{n}$ and $\|z\|^2 = n$, we get

$$\begin{aligned}\frac{1}{n}\min_{\delta=-1,+1}|z-\delta\mathrm{sign}(\widehat{u}_1)|_0 &\leq \frac{1}{n}\min_{\delta=-1,+1}\|z-\delta\sqrt{n}\widehat{u}_1\|^2\\ &= \frac{1}{n}\min_{\delta=-1,+1}(2n-2\delta\sqrt{n}\langle z,\widehat{u}_1\rangle)\\ &= 2(1-|\langle z/\sqrt{n},\widehat{u}_1\rangle|)\\ &\leq 2(1-\langle z/\sqrt{n},\widehat{u}_1\rangle^2),\end{aligned}$$

where we have used that $|\langle z/\sqrt{n},\widehat{u}_1\rangle| \leq \|z/\sqrt{n}\|\|\widehat{u}_1\| = 1$ in the last inequality.

Notice that $z/\sqrt{n}$ is a unit-norm leading eigenvector of $\frac{1}{n}\|\theta\|^2 zz^T$, associated with the eigenvalue $\|\theta\|^2$. Notice also that the second eigenvalue of $\frac{1}{n}\|\theta\|^2 zz^T$ is 0, as $\frac{1}{n}\|\theta\|^2 zz^T$ is a rank one matrix. Combining the previous bound with Davis-Kahan inequality (6.4) with $A = \frac{1}{n}\|\theta\|^2 zz^T$ and $B = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, we get

$$\begin{aligned}\min_{\delta=-1,+1}\frac{1}{n}|z-\delta\widehat{z}|_0 \leq 2(1-\langle z/\sqrt{n},\widehat{u}_1\rangle^2) &\leq 8\inf_{\lambda\in\mathbb{R}}\frac{\left|\lambda I_n + \frac{1}{n}\mathbf{X}\mathbf{X}^T - \frac{1}{n}\|\theta\|^2 zz^T\right|_{\mathrm{op}}^2}{\|\theta\|^4}\\ &\leq 8\,\frac{\left|\frac{1}{n}\mathbf{X}\mathbf{X}^T - \frac{1}{n}\|\theta\|^2 zz^T - \frac{p\sigma^2}{n}I_n\right|_{\mathrm{op}}^2}{\|\theta\|^4}. \qquad (8.8)\end{aligned}$$

It remains to bound from above $\left|\frac{1}{n}\mathbf{X}\mathbf{X}^T - \frac{1}{n}\|\theta\|^2 zz^T - \frac{p\sigma^2}{n}I_n\right|_{\mathrm{op}}$.

Lemma 8.4 *There exists two exponential random variables ξ, ξ' with parameter 1, such that the operator norm of*

$$W = \frac{1}{n}\mathbf{X}\mathbf{X}^T - \frac{\|\theta\|^2}{n}zz^T - \frac{p\sigma^2}{n}I_n$$

is upper-bounded by

$$|W|_{\mathrm{op}} \leq 4\sigma^2\sqrt{\frac{p}{n}\left(6+2\frac{\xi}{n}\right)} + \left(48+\frac{16\xi}{n}\right)\sigma^2 + 2\|\theta\|\sigma\left(1+\sqrt{\frac{8\xi'}{n}}\right). \qquad (8.9)$$

Let us explain how Theorem 8.2 follows from the Bound (8.9). According to (8.8) and (8.9), we have with probability at least $1-2e^{-n/2}$, the upper bound

$$\min_{\delta=-1,+1}\frac{1}{n}|z-\delta\widehat{z}|_0 \leq 1\wedge\left(\frac{30\sqrt{p/n}+159+17\|\theta\|/\sigma}{\|\theta\|^2/\sigma^2}\right)^2.$$

The right-hand side is smaller than 1 only if $17 \leq \|\theta\|/\sigma$, so $159 \leq 10\|\theta\|/\sigma$, from which follows

$$\begin{aligned}\min_{\delta=-1,+1}\frac{1}{n}|z-\delta\widehat{z}|_0 &\leq 1\wedge\left(\frac{30\sqrt{p/n}+27\|\theta\|/\sigma}{\|\theta\|^2/\sigma^2}\right)^2\\ &\leq 1\wedge\left(1800\,\frac{p/n+\|\theta\|^2/\sigma^2}{\|\theta\|^4/\sigma^4}\right) = 1\wedge\frac{1800}{s^2},\end{aligned}$$

which gives (8.6). It remains to prove Lemma 8.4.

Proof of Lemma 8.4.
We have $nW = (\mathbf{E}\mathbf{E}^T - p\sigma^2 I_n) + \mathbf{E}\theta z^T + z\theta^T\mathbf{E}^T$.

The quadratic term $\mathbf{E}\mathbf{E}^T - p\sigma^2 I_n$ is controlled by the bound (6.11) of Theorem 6.8: There exists an exponential random variable ξ with parameter 1 such that

$$\left|\mathbf{E}\mathbf{E}^T - p\sigma^2 I_n\right|_{\text{op}} \leq 4\sigma^2 \sqrt{p(6n+2\xi)} + (48n+16\xi)\sigma^2.$$

Let us now control the cross terms in W.

Lemma 8.5 *There exists an exponential random variable ξ' with parameter 1, such that*

$$\frac{1}{n}\left|z\theta^T\mathbf{E}^T\right|_{\text{op}} = \frac{1}{n}\left|\mathbf{E}\theta z^T\right|_{\text{op}} \leq \|\theta\|\sigma\left(1+\sqrt{\frac{8\xi'}{n}}\right). \tag{8.10}$$

Proof of Lemma 8.5. Dividing left and right-hand side of (8.10) by σ, we can assume with no loss of generality that $\sigma = 1$. Let us set $u = \theta/\|\theta\|$ and $v = z/\sqrt{n}$.

We observe that for x with norm 1,

$$\|\mathbf{E}uv^Tx\| = |v^Tx|\|\mathbf{E}u\| \leq \|\mathbf{E}u\|,$$

with equality for $x = v$. Hence $\left|\mathbf{E}uv^T\right|_{\text{op}} = \|\mathbf{E}u\|$.

For the same reasons as in Step 2 of the proof of Theorem 6.8, the random variable Eu follows a standard Gaussian $\mathcal{N}(0, I_n)$ distribution. Hence, according to Hanson-Wright inequality (6.9) with $S = I_n$, there exists an exponential random variable ξ' with parameter 1, such that

$$\left|\mathbf{E}uv^T\right|_{\text{op}}^2 = \|\mathbf{E}u\|^2 \leq n + \sqrt{8n\xi'} + 8\xi' \leq \left(\sqrt{n} + \sqrt{8\xi'}\right)^2.$$

Since $\frac{1}{n}\left|\mathbf{E}\theta z^T\right|_{\text{op}} = \frac{\|\theta\|}{\sqrt{n}}\left|\mathbf{E}uv^T\right|_{\text{op}}$, the Bound (8.10) follows. □

Combining the Theorem 6.8, the Lemma 8.5, and the decomposition

$$nW = (\mathbf{E}\mathbf{E}^T - p\sigma^2 I_n) + \mathbf{E}\theta z^T + z\theta^T\mathbf{E}^T,$$

we get (8.9). The proof of Lemma 8.4 is complete. □

8.4 Exercises

8.4.1 Sterling numbers of second kind

Let us denote by $S(n, K)$ the number of partitions of $\{1, \ldots, n\}$ into K (non-empty) clusters.

1. What is the value of $S(n, 1)$? of $S(n, n)$?
2. With a combinatorial argument, prove the recursion formula
$$S(n,k) = kS(n-1,k) + S(n-1,k-1), \quad \text{for} \quad 2 \leq k \leq n-1.$$
3. Prove by induction that
$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j C_k^j (k-j)^n,$$
with $C_k^j = k!/(j!(k-j)!)$ the binomial coefficient.
4. With the recursion formula, prove the simple lower bound
$$S(n,k) \geq k^{n-k}.$$

The numbers $S(n,k)$ are called the Sterling numbers of the second kind. The total number $B_n = \sum_{k=1}^n S(n,k)$ of possible partitions of n elements (without constraints on the number of groups) are called the Bell numbers.

For a fixed k, we observe that $S(n,k)$ grows exponentially fast with n, and for $k = n/\log(n)$ the growth is even super-exponential as, for any $0 < c < 1$, we have $S(n,k) \geq \exp(cn\log(n))$ for n large enough. In particular, the Bell number B_n grows super-exponentially fast with n.

8.4.2 Exact recovery with hierarchical clustering

In this exercise, we provide some conditions ensuring that hierarchical clustering exactly recovers the hidden partition in the setting (8.3) of Section 8.3.2. We denote by $\mathcal{W} = \left\{(i, j) : z_i = z_j,\ i < j\right\}$ the set of pairs of points within the same cluster $G_1 = \{i : z_i = -1\}$ or $G_2 = \{i : z_i = 1\}$, and by $\mathcal{B} = \left\{(i, j) : z_i \neq z_j,\ i < j\right\}$ the set of pairs of points between the two clusters.

1. What is the value of $\mathbb{E}\left[\|X_i - X_j\|^2\right]$ for $(i, j) \in \mathcal{W}$? and for $(i, j) \in \mathcal{B}$?
2. Prove that $|\mathcal{W}| \leq n^2/2$ and $|\mathcal{B}| \leq n^2/4$.
3. Prove that

$$\mathbb{P}\left(\max_{(i,j)\in\mathcal{W}} \|X_i - X_j\|^2 \geq 2p\sigma^2 + (\sigma^2\sqrt{96p\log(n)}) \vee (48\sigma^2\log(n))\right) \leq \frac{1}{2n}.$$

4. Let ε be a Gaussian $\mathcal{N}(0, \sigma^2 I_p)$ random variable. Check that

$$\mathbb{P}\left(\langle\theta, \varepsilon\rangle \leq -\sigma\|\theta\|\sqrt{2L}\right) \leq e^{-L}.$$

5. Prove that

$$\mathbb{P}\left(\min_{(i,j)\in\mathcal{B}} \|X_i - X_j\|^2 \leq 2p\sigma^2 + 4\|\theta\|^2 - (\sigma^2\sqrt{96p\log(n)}) \vee (48\sigma^2\log(n)) - 14\sigma\|\theta\|\sqrt{\log(n)}\right) \leq \frac{1}{2n}.$$

6. For $a, b > 0$, prove that the condition $\|\theta\|^2 \geq (a/2) \vee (b/2)^2$ ensures the inequality $4\|\theta\|^2 \geq a + b\|\theta\|$. Conclude that, when we have $\|\theta\|^2 \geq \sigma^2\left(\sqrt{24p\log(n)} \vee (49\log(n))\right)$, then the hierarchical clustering algorithm with Euclidean distance and single or complete linkage recovers the clusters G_1 and G_2 with probability at least $1 - 1/n$.

8.4.3 Estimating Γ

We use in this exercise the notation introduced at the beginning of this chapter: We have the decomposition $X_i = \mu_i + E_i$, with $\mu_i = \mathbb{E}[X_i]$ and $\mu_i = \theta_k$ for all $i \in G_k^*$.

As discussed in Section 8.3.1,

$$(\mathbf{X}\mathbf{X}^T)_{ij} = (A\Theta\Theta^T A^T)_{ij} + E_i^T E_j + E_i^T(\Theta^T A^T)_j + (A\Theta)_i^T E_j,$$

with $\mathbb{E}\left[E_i^T(\Theta^T A^T)_j\right] = 0 = \mathbb{E}\left[(A\Theta)_i^T E_j\right]$, and

$$\mathbb{E}\left[E_i^T E_j\right] = \mathbf{1}_{i=j}\mathbb{E}\left[\|E_i\|^2\right].$$

Let us denote by $\widetilde{\Gamma}$ the diagonal matrix with $\widetilde{\Gamma}_{ii} := \|E_i\|^2$. In order to avoid the systematic bias induced by $\widetilde{\Gamma}$ in the spectral decomposition, we would like to work on the matrix $\mathbf{X}\mathbf{X}^T - \widetilde{\Gamma}$. This is not possible though, since the norm $\|E_i\|$ is not observed. The idea is to compute instead the spectral decomposition of the matrix $\mathbf{X}\mathbf{X}^T - \widehat{\Gamma}$, where $\widehat{\Gamma}$ is built from data and is a good evaluation of $\widetilde{\Gamma}$.

If we knew the partition G^*, "estimating[1]" the random quantity $\widetilde{\Gamma}_{ii}$ (or the parameter Γ_{ii}) would be a simple task. Actually, if $i' \neq i$ belongs to the same group as i, then

$$\langle X_i - X_{i'}, X_i\rangle = \|E_i\|^2 - \langle E_i, E_{i'}\rangle + \langle\mu_i, E_i - E_{i'}\rangle, \tag{8.11}$$

[1] with some abuse of langage, we use the word "estimation" even if $\widetilde{\Gamma}_{ii}$ is a random quantity, not a parameter

with $\mathbb{E}\left[\langle E_i, E_{i'}\rangle\right] = \mathbb{E}\left[\langle \mu_i, E_i - E_{i'}\rangle\right] = 0$. So $\langle X_i - X_{i'}, X_i\rangle$ is an unbiased "estimator" of $\widetilde{\Gamma}_{ii}$ (and Γ_{ii}).

The difficulty is that we do not know G^*, and we need to estimate Γ_{ii} to estimate G^*. To break this vicious spiral, we can build yet on (8.11), by replacing i' by a data-driven choice $\widehat{i}$. We observe that we have the decomposition

$$\langle X_i - X_{\widehat{i}}, X_i\rangle - \langle X_i - X_{i'}, X_i\rangle = \langle X_i - X_{\widehat{i}}, X_{i'}\rangle - \langle X_i - X_{i'}, X_{\widehat{i}}\rangle,$$

and we will be able to control the size of the right hand side, if we are able to control $\max_k |\langle X_i - X_{\widehat{i}}, X_k\rangle|$. This observation motivates the definition of the following estimator

$$\widehat{\Gamma}_{ii} := \langle X_i - X_{\widehat{i}}, X_i\rangle, \quad \text{with} \quad \widehat{i} \in \operatorname*{argmin}_j \max_{k:\, k\neq i,j} |\langle X_i - X_j, X_k\rangle|. \tag{8.12}$$

In this exercise, you will prove the following bound on the error $|\widehat{\Gamma} - \widetilde{\Gamma}|_\infty$.

Proposition 8.6 *Assume that the observations $X_1, \ldots, X_n$ are independent, with X_i following a $\mathcal{N}(\mu_i, \Sigma_i)$ Gaussian distribution.*
Assume also that each group G_k has a cardinality as least 2.
Then, with probability at least $1 - 2e^{-L}$, the diagonal matrix $\widehat{\Gamma}$ fulfills

$$\begin{aligned}&|\widehat{\Gamma} - \widetilde{\Gamma}|_\infty \\ &\leq 6\|\Theta\|_{2\infty}\sqrt{|\Sigma|_{\mathrm{op}}\,(3\log(n) + L)} + 10(\|\Sigma\|_F\,\sqrt{2\log(n) + L}) \vee (|\Sigma|_{\mathrm{op}}\,(4\log(n) + 2L)),\end{aligned}$$

where

$$\|\Theta\|_{2\infty} := \max_k \|\theta_k\|, \quad |\Sigma|_{\mathrm{op}} := \max_i |\Sigma_i|_{\mathrm{op}} \quad \textit{and} \quad \|\Sigma\|_F := \max_i \|\Sigma_i\|_F.$$

While this estimator $\widehat{\Gamma}$ gives good results, it can be improved in order to avoid the dependency on $\|\Theta\|_{2\infty}$. The improved estimator is somewhat more complex, and its analysis is beyond the scope of this monograph.

To prove Proposition 8.6, answer to the following questions.

1. Take $i' \neq i$ in the same group as i. Starting from the decomposition

$$\langle X_i - X_{\widehat{i}}, X_i\rangle = \langle X_i - X_{i'}, X_i\rangle + \langle X_i - X_{\widehat{i}}, X_{i'}\rangle - \langle X_i - X_{i'}, X_{\widehat{i}}\rangle$$

and using the definition of $\widehat{i}$, prove the inequalities

$$\begin{aligned}|\widehat{\Gamma}_{ii} - \widetilde{\Gamma}_{ii}| \leq& |\langle \mu_i, E_i - E_{i'}\rangle| + |\langle E_{i'}, E_i\rangle| + \max_{k\neq i,\widehat{i}} |\langle X_i - X_{\widehat{i}}, X_k\rangle| + \max_{k\neq i,i'} |\langle X_i - X_{i'}, X_k\rangle| \\ \leq& |\langle \mu_i, E_i - E_{i'}\rangle| + |\langle E_{i'}, E_i\rangle| + 2\max_{k\neq i,i'} |\langle X_i - X_{i'}, X_k\rangle| \\ \leq& 3\max_{k,i,i'} |\langle \mu_k, E_i - E_{i'}\rangle| + 5\max_{k,i} |\langle E_i, E_k\rangle|.\end{aligned}$$

2. Check that $\mathrm{Var}(\langle \mu_k, E_i - E_{i'}\rangle) = \mu_k^T(\Sigma_i + \Sigma_{i'})\mu_k \leq 2\,|\Sigma|_{\mathrm{op}}\,\|\Theta\|_{2\infty}$. What is the distribution of $\langle \mu_k, E_i - E_{i'}\rangle$?

3. Prove the bound

$$\mathbb{P}\left[\max_{k,i,i'} |\langle \mu_k, E_i - E_{i'}\rangle| > 2\|\Theta\|_{2\infty}\sqrt{|\Sigma|_{\mathrm{op}}\,(3\log(n) + L)}\right] \leq e^{-L}.$$

4. We can write the scalar product $\langle E_i, E_k\rangle$ as $\varepsilon_i^T \Sigma_i^{1/2}\Sigma_k^{1/2}\varepsilon_k$, with $\varepsilon_i, \varepsilon_k$, two independent standard Gaussian random variables in $\mathbb{R}^p$. With Hanson-Wright inequality (6.10) page 72, proves that

$$\mathbb{P}\left[|\langle E_i, E_k\rangle| > (2\|\Sigma_i^{1/2}\Sigma_k^{1/2}\|_F\,\sqrt{L}) \vee (4\left|\Sigma_i^{1/2}\right|_{\mathrm{op}}\left|\Sigma_k^{1/2}\right|_{\mathrm{op}} L)\right] \leq e^{-L}.$$

5. Deduce from the previous question that

$$\mathbb{P}\left[\max_{i\neq k} |\langle E_i, E_k\rangle| > (2\|\Sigma\|_F \sqrt{2\log(n)+L}) \vee (4\,|\Sigma|_{\mathrm{op}}\,(2\log(n)+L))\right] \le e^{-L},$$

and conclude the proof of Proposition 8.6.

8.4.4 Illustration of hierarchical clustering and spectral clustering

In this section, we illustrate the behavior of different clustering algorithms on synthetic data. We generate the data as follows. Setting $\theta = [0.9, 0.9] \in \mathbb{R}^2$, half of the data points are i.i.d. with a Gaussian $\mathcal{N}(\theta, I_2/2)$ distribution and the other half are i.i.d. with a Gaussian $\mathcal{N}(-\theta, I_2)$ distribution. We compute the clusterings output by Spectral clustering, and by Hierarchical clustering with complete, single and average linkage. We also add the results for clustering output by Lloyd algorithm (another popular clustering algorithm, not covered in theses lectures notes). The results are displayed in Figure 8.2.

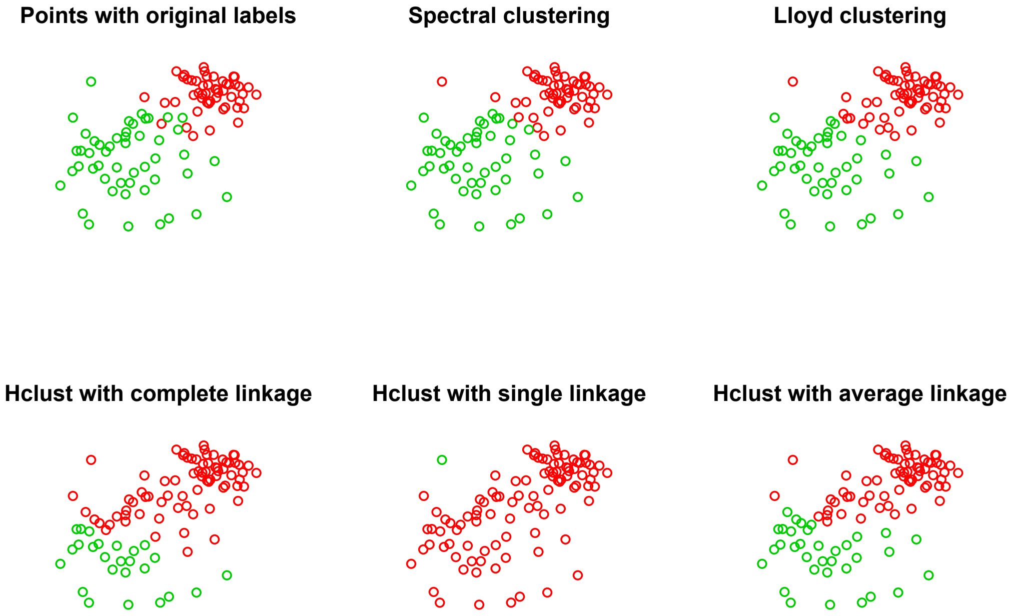


Figure 8.2: First row: original labels (left), spectral clustering (middle) and Lloyd clustering (right). Second row: hierarchical clustering with complete linkage (left), single linkage (middle) and average linkage (right).

We observe that the choice of the linkage has a strong impact on the output of hierarchical clustering. Complete linkage tends to produce clusters with similar width, leading to cluster a fraction of the green points with the red ones. Single linkage cut the data at the largest between points distance, leading to two unbalanced clusters. In this case, it singles out one of the data points.

It is interesting to inspect the dendrogram for the three linkages. They are displayed in Figure 8.3.

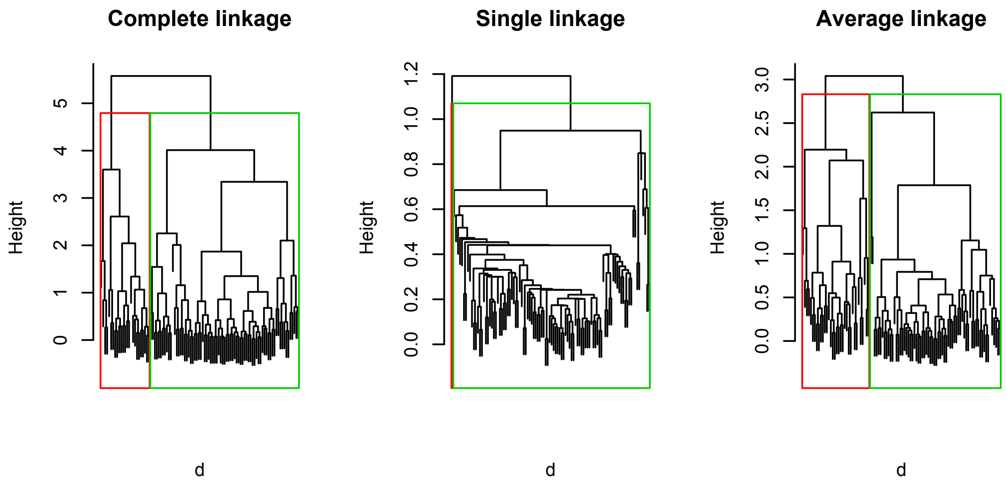


Figure 8.3: Dendrograms built from complete linkage (left), single linkage (middle) and average linkage (right).

You can reproduce these results by running the following R-code.

```
# generate data
n<-100
X<-array(0,c(n,2))
X[1:(n/2),]<-rnorm(n,mean=0.9,sd=0.5)
X[(n/2+1):n,]<-rnorm(n,mean=-0.9)
etiquettes<-c(rep(1,n/2),rep(2,n/2))  # labels of points
d<-dist(X, method = "euclidean")  # matrix of distances

# compute hierarchical clustering with complete linkage
hcomplete <- hclust(d, method = "complete")
G2complete <- cutree(hcomplete,k=2)

# compute hierarchical clustering with single linkage
hsingle <- hclust(d, method = "single")
G2single <- cutree(hsingle,k=2)

# compute hierarchical clustering with average linkage
haverage <- hclust(d, method = "average")
G2average <- cutree(haverage,k=2)

# compute spectral clustering
v<- svd(X,nu=1,nv=0)$u
spect <- sign(v)

# compute Lloyd clustering
lloyd<-kmeans(X,centers=2)

# display the results
par(mfrow=c(2,3))
plot(X,col=1+etiquettes, main="Original")
plot(X,col=2.5+spect/2, main="Spectral")
plot(X,col=1+lloyd$cluster, main="Lloyd")
plot(X,col=G2complete+1,main="Complete linkage")
plot(X,col=G2single+1, main="Single linkage")
plot(X,col=G2average+1, main="Average linkage")

# display the dendrograms
par(mfrow=c(1,3))
plot(hcomplete,main="Complete linkage",label=FALSE)
rect.hclust(hcomplete,k=2,border=2:3)
plot(hsingle,main="Single linkage",label=FALSE)
rect.hclust(hsingle,k=2,border=2:3)
plot(haverage,main="Average linkage",label=FALSE)
rect.hclust(haverage,k=2,border=2:3)
```

When the clusters are better separated, the various algorithms tend to produce similar results. Running the same example for $\mu = [1.5, 1.5]$, we get the results displayed in Figure 8.4.

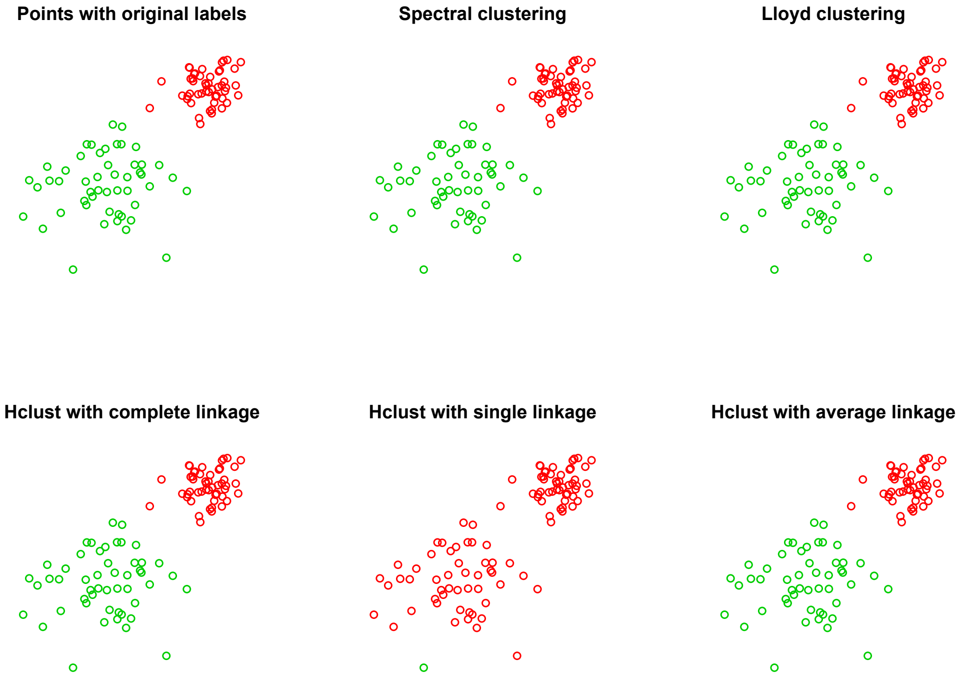


Figure 8.4: Results for well separated clusters ($\mu = [1.5, 1.5]$).

Appendix A

Gaussian Distribution

A.1 Gaussian Random Vectors

A random vector $Y \in \mathbb{R}^d$ is distributed according to the $\mathcal{N}(m, \Sigma)$ Gaussian distribution, with $m \in \mathbb{R}^d$ and $\Sigma \in \mathcal{S}_d^+$ (the set of all $d \times d$ symmetric positive semidefinite matrix), when

$$\mathbb{E}\left[e^{\mathrm{i}\langle\lambda, Y\rangle}\right] = \exp\left(\mathrm{i}\langle\lambda, m\rangle - \frac{1}{2}\lambda^T\Sigma\lambda\right), \quad \text{for all } \lambda \in \mathbb{R}^d. \tag{A.1}$$

When matrix Σ is nonsingular (i.e., positive definite), the $\mathcal{N}(m, \Sigma)$ Gaussian distribution has a density with respect to the Lebesgue measure on $\mathbb{R}^d$ given by

$$\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}}\exp\left(-\frac{1}{2}(y-m)^T\Sigma^{-1}(y-m)\right).$$

Affine transformations of Gaussian distribution are still Gaussian.

Lemma A.1 Affine transformation
Let $Y \in \mathbb{R}^d$ be a random vector with $\mathcal{N}(m, \Sigma)$ Gaussian distribution. Then for any $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$,

$$AY + b \sim \mathcal{N}(Am + b, A\Sigma A^T).$$

In particular, for $a \in \mathbb{R}^d$,

$$\langle a, Y\rangle \sim \mathcal{N}(\langle m, a\rangle, a^T\Sigma a).$$

Proof. The first identity is obtained by computing the characteristic function of $AY + b$

$$\begin{aligned}\mathbb{E}\left[e^{\mathrm{i}\langle\lambda, AY+b\rangle}\right] = \mathbb{E}\left[e^{\mathrm{i}\langle A^T\lambda, Y\rangle + \mathrm{i}\langle\lambda, b\rangle}\right] &= \exp\left(\mathrm{i}\langle A^T\lambda, m\rangle - \frac{1}{2}(A^T\lambda)^T\Sigma A^T\lambda\right)e^{\mathrm{i}\langle\lambda, b\rangle}\\ &= \exp\left(\mathrm{i}\langle\lambda, Am + b\rangle - \frac{1}{2}\lambda^T A\Sigma A^T\lambda\right).\end{aligned}$$

The second identity is obtained with $A = a^T$ and $b = 0$. □

Lemma A.2 Orthogonal projections onto subspaces
Let $Y \in \mathbb{R}^d$ be a random vector with $\mathcal{N}(m, \Sigma)$ Gaussian distribution, and let S and V be two linear spans of $\mathbb{R}^d$ orthogonal with respect to the scalar product induced by Σ. Then the variables $\mathrm{Proj}_S Y$ and $\mathrm{Proj}_V Y$ are independent and follow, respectively, the $\mathcal{N}(\mathrm{Proj}_S m, \mathrm{Proj}_S\Sigma\,\mathrm{Proj}_S)$ and $\mathcal{N}(\mathrm{Proj}_V m, \mathrm{Proj}_V\Sigma\,\mathrm{Proj}_V)$ Gaussian distribution.

Proof. Since the projection matrices Proj_S and Proj_V are symmetric, we obtain that the joint characteristic function of $\text{Proj}_S Y$ and $\text{Proj}_V Y$ is

$$\begin{aligned}
&\mathbb{E}\left[e^{\mathrm{i}\langle\lambda,\text{Proj}_S Y\rangle+\mathrm{i}\langle\gamma,\text{Proj}_V Y\rangle}\right] = \mathbb{E}\left[e^{\mathrm{i}\langle\text{Proj}_S\lambda+\text{Proj}_V\gamma,Y\rangle}\right]\\
&= \exp\left(\mathrm{i}\langle\text{Proj}_S\lambda+\text{Proj}_V\gamma, m\rangle - \frac{1}{2}(\text{Proj}_S\lambda+\text{Proj}_V\gamma)^T\Sigma(\text{Proj}_S\lambda+\text{Proj}_V\gamma)\right)\\
&= \exp\left(\mathrm{i}\langle\lambda, \text{Proj}_S m\rangle - \frac{1}{2}\lambda^T\text{Proj}_S\Sigma\,\text{Proj}_S\lambda\right)\\
&\quad\times\exp\left(\mathrm{i}\langle\gamma, \text{Proj}_V m\rangle - \frac{1}{2}\gamma^T\text{Proj}_V\Sigma\,\text{Proj}_V\gamma\right)\\
&= \mathbb{E}\left[e^{\mathrm{i}\langle\lambda,\text{Proj}_S Y\rangle}\right]\mathbb{E}\left[e^{\mathrm{i}\langle\gamma,\text{Proj}_V Y\rangle}\right].
\end{aligned}$$

We conclude with Lemma A.1. □

A.2 Chi-Square Distribution

Let $Y \in \mathbb{R}^n$ be a random vector with $\mathcal{N}(0, I_n)$ Gaussian distribution. The χ^2 distribution with n degrees of freedom, corresponds to the distribution of $\|Y\|^2$. In particular, the mean of a $\chi^2(n)$ distribution is

$$\mathbb{E}\left[\|Y\|^2\right] = \sum_{i=1}^{n}\mathbb{E}\left[Y_i^2\right] = n.$$

Lemma A.3 Norms of projections
Let $Y \in \mathbb{R}^n$ be a random vector with $\mathcal{N}(0, I_n)$ Gaussian distribution, and let S be a linear subspace of $\mathbb{R}^n$ with dimension d. Then, the variable $\text{Proj}_S Y$ follows the $\mathcal{N}(0, \text{Proj}_S)$ Gaussian distribution and the square-norm $\|\text{Proj}_S Y\|^2$ follows a χ^2-distribution of degree d.

In particular, $\mathbb{E}\left[\|\text{Proj}_S Y\|^2\right] = \dim(S)$.

Proof. The projection Proj_S is symmetric, so $\text{Proj}_S\text{Proj}_S^T = \text{Proj}_S$ and $\text{Proj}_S Y$ follows a $\mathcal{N}(0, \text{Proj}_S)$ Gaussian distribution according to Lemma A.1.

Let $u_1, \dots, u_d$ be an orthonormal basis of S and set $U = [u_1, \dots, u_d]$. Since $U^T U = I_d$, the vector $U^T Y$ follows a $\mathcal{N}(0, I_d)$-distribution and

$$\|\text{Proj}_S Y\|^2 = \sum_{k=1}^{d}(u_k^T Y)^2 = \|U^T Y\|^2$$

follows a χ^2 distribution of degree d. □

A.3 Gaussian Conditioning

We provide in this section a few useful results on Gaussian conditioning.

Lemma A.4
We consider two sets $A = \{1, \dots, k\}$ and $B = \{1, \dots, p\} \setminus A$, and a Gaussian random vector $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \in \mathbb{R}^p$ with $\mathcal{N}(0, \Sigma)$ distribution. We assume that Σ is nonsingular and write $K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix}$ for its inverse.

In the next formulas, K_{AA}^{-1} will refer to the inverse $(K_{AA})^{-1}$ of K_{AA} (and not to $(K^{-1})_{AA} = \Sigma_{AA}$).

Then, the conditional distribution of X_A given X_B is the Gaussian $\mathcal{N}\left(-K_{AA}^{-1}K_{AB}X_B, K_{AA}^{-1}\right)$ distribution. In others words, we have the decomposition

$$X_A = -K_{AA}^{-1}K_{AB}X_B + \varepsilon_A, \quad \text{where } \varepsilon_A \sim \mathcal{N}\left(0, K_{AA}^{-1}\right) \text{ is independent of } X_B. \tag{A.2}$$

Proof. We write $g(x_A, x_B)$, respectively, $g(x_A|x_B)$ and $g(x_B)$, for the density of the distribution of X, respectively, of X_A given $X_B = x_B$ and X_B. We have

$$\begin{aligned} g(x_A|x_B) &= g(x_A, x_B)/g(x_B) \\ &= \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}x_A^T K_{AA} x_A - x_A^T K_{AB} x_B - \frac{1}{2} x_B^T \left(K_{BB} - \Sigma_{BB}^{-1}\right) x_B\right), \end{aligned}$$

with Σ_{BB} the covariance matrix of X_B. Since $\Sigma_{BB}^{-1} = K_{BB} - K_{BA}K_{AA}^{-1}K_{AB}$, we have

$$g(x_A|x_B) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(x_A + K_{AA}^{-1}K_{AB}x_B)^T K_{AA}(x_A + K_{AA}^{-1}K_{AB}x_B)\right).$$

We recognize the density of the Gaussian $\mathcal{N}\left(-K_{AA}^{-1}K_{AB}\, x_B, K_{AA}^{-1}\right)$ distribution. □

Corollary A.5 *For any $a \in \{1, \dots, p\}$, we have*

$$X_a = -\sum_{b\,:\, b \neq a} \frac{K_{ab}}{K_{aa}} X_b + \varepsilon_a, \quad \text{where } \varepsilon_a \sim \mathcal{N}(0, K_{aa}^{-1}) \text{ is independent of } \{X_b : b \neq a\}. \tag{A.3}$$

Proof. We apply the previous lemma with $A = \{a\}$ and $B = A^c$. □

Finally, we derive from (A.2) the following simple formula for the conditional correlation of X_a and X_b given $\{X_c : c \neq a, b\}$, which is defined by

$$\text{cor}(X_a, X_b|X_c : c \neq a, b) = \frac{\text{cov}(X_a, X_b|X_c : c \neq a, b)}{\sqrt{\text{var}(X_a|X_c : c \neq a, b)\, \text{var}(X_b|X_c : c \neq a, b)}}.$$

Corollary A.6 *For any $a, b \in \{1, \dots, p\}$, we have*

$$\text{cor}(X_a, X_b|X_c : c \neq a, b) = \frac{-K_{ab}}{\sqrt{K_{aa}\, K_{bb}}}. \tag{A.4}$$

Proof. The previous lemma with $A = \{a, b\}$ and $B = A^c$ gives

$$\text{cov}(X_A|X_B) = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ab} & K_{bb} \end{pmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{pmatrix} K_{bb} & -K_{ab} \\ -K_{ab} & K_{aa} \end{pmatrix}.$$

Plugging this formula in the definition of the conditional correlation, we obtain Formula (A.4). □

Appendix B

Convex Functions

B.1 Convex functions

A function $F : \mathbb{R}^n \to \mathbb{R}$ is convex if $F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y)$ for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$. An equivalent definition is that the epigraph $\{(x,y),\ x \in \mathbb{R}^n,\ y \in [F(x), +\infty[\}$ is a convex subset of $\mathbb{R}^{n+1}$.

Lemma B.1 *When the function $F : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, we have*

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle\,, \quad \textit{for all } x, y \in \mathbb{R}^n.$$

Proof. Let $x, h \in \mathbb{R}^n$, and define $f : \mathbb{R} \to \mathbb{R}$ by $f(t) = F(x + th)$. Since F is differentiable, so is f and $f'(t) = \langle \nabla F(x + th), h \rangle$. By Taylor's expansion, we have for some $t^* \in [0,1]$

$$F(x+h) - F(x) = \langle \nabla F(x + t^* h), h \rangle = f'(t^*).$$

Since

$$f(\lambda t + (1-\lambda)s) = F(\lambda(x+th) + (1-\lambda)(x+sh)) \leq \lambda f(t) + (1-\lambda) f(s)\,,$$

the function f is convex, so

$$F(x+h) - F(x) = f'(t^*) \geq f'(0) = \langle \nabla F(x), h \rangle.$$

We conclude by setting $h = y - x$. □

B.2 Jensen inequality

Jensen inequality generalizes the two points inequality $F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y)$ to arbitrary convex combination given by expectations.

Lemma B.2 Jensen inequality
For any convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ and any random variable X in $\mathbb{R}^d$, such that $\varphi(X)$ is integrable, we have

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]\,.$$

Proof. Let us denote by $\mathcal{L}_\varphi$ the set of affine functions from $\mathbb{R}^d$ to $\mathbb{R}$, such that $L(x) \leq \varphi(x)$ for all $x \in \mathbb{R}^d$. Since

$$\varphi(x) = \sup_{L \in \mathcal{L}_\varphi} L(x)\,,$$

the linearity of the expectation gives

$$\mathbb{E}[\varphi(X)] = \mathbb{E}\left[\sup_{L \in \mathcal{L}_\varphi} L(X)\right] \geq \sup_{L \in \mathcal{L}_\varphi} \mathbb{E}[L(X)] = \sup_{L \in \mathcal{L}_\varphi} L(\mathbb{E}[X]) = \varphi(\mathbb{E}[X])\,.$$

□

Appendix C

Constrained optimization

Let $f, g_1, \dots, g_N$ be $N+1$ functions from $\mathbb{R}^d$ to $\mathbb{R}$. In this appendix, we recall some basic results related to the minimization problem

$$\min_{x \in C} f(x), \quad \text{where } C = \left\{ x \in \mathbb{R}^d : \ g_1(x) \leq 0, \dots, g_N(x) \leq 0, \ \ell_1(x) = 0, \dots, \ell_m(x) = 0 \right\}. \quad \text{(C.1)}$$

C.1 Dual problem

C.1.1 Lagrangian and dual functions

Two functions play an important role in the investigation of the optimization problem (C.1): The Lagrangian function

$$L(x, \lambda, \mu) = f(x) + \sum_{j=1}^{N} \lambda_j g_j(x) + \sum_{i=1}^{m} \mu_i \ell_i(x), \quad \text{for } (x, \lambda, \mu) \in \mathbb{R}^d \times \mathbb{R}^N \times \mathbb{R}^m, \qquad \text{(C.2)}$$

and the dual function

$$q(\lambda, \mu) = \inf_{x \in \mathbb{R}^d} L(x, \lambda, \mu), \quad \text{for } (\lambda, \mu) \in \mathbb{R}^N \times \mathbb{R}^m. \qquad \text{(C.3)}$$

Since $q(\lambda, \mu)$ is an infimum of affine functions, the dual function q is concave.

C.1.2 Weak duality

Let us abbreviate the N conditions $\lambda_j \geq 0$, $j = 1, \dots, N$ by $\lambda \geq 0$. For any $\lambda \geq 0$, $\mu \in \mathbb{R}^m$ and $x \in C$, we have

$$q(\lambda, \mu) \leq L(x, \lambda, \mu) = f(x) + \sum_{j=1}^{N} \underbrace{\lambda_j}_{\geq 0} \underbrace{g_j(x)}_{\leq 0} + \sum_{i=1}^{m} \mu_i \underbrace{\ell_i(x)}_{=0} \leq f(x).$$

So, for any $\lambda \geq 0$ and any $\mu \in \mathbb{R}^m$, we have

$$q(\lambda, \mu) \leq \min_{x \in C} f(x).$$

We then obtain the following lower bound on the minimization problem (C.1)

$$\sup_{\lambda \geq 0, \ \mu \in \mathbb{R}^m} q(\lambda, \mu) \leq \min_{x \in C} f(x) \qquad \text{(weak duality).} \qquad \text{(C.4)}$$

This inequality is called the weak-duality condition. We observe that while the primal problem (C.1) can be hard to solve numerically in general, when the function q has an explicit expression, the dual problem

$$\sup_{\lambda \geq 0, \ \mu \in \mathbb{R}^m} q(\lambda, \mu) \qquad \text{(dual problem)} \qquad \text{(C.5)}$$

can be solved efficiently, since q is concave and the constraint $\lambda \geq 0$ is linear. The weak-duality can then be a convenient tool to compute efficiently a lower bound on a minimisation problem.

The difference between the value of the primal problem (C.1) and the value of the dual problem (C.5) is called the duality gap. When the inequality in (C.4) is strict, this gap is positive and otherwise it is zero.

C.1.3 Finding a solution

When the functions are convex and differentiable, we can give some simple suffisant conditions for x^* to be a minimizer of the primal problem (C.1).

Lemma C.1 Karush-Kuhn-Tucker (KKT) conditions
Assume that $f, g_1, \ldots, g_N$ are convex and differentiable. Assume also that $\ell_1, \ldots, \ell_m$ are affine. If $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^d \times \mathbb{R}^N \times \mathbb{R}^m$ are such that

$$\begin{aligned} \nabla_x L(x^*, \lambda^*, \mu^*) &= 0 && \textit{(first order condition)} \\ x^* \in C,\ \ \lambda^* &\geq 0 && \textit{(feasibility condition)} \\ \lambda_j^* g_j(x^*) &= 0 && \textit{(slackness condition)} \end{aligned}$$

then, we have

$$q(\lambda^*, \mu^*) = \max_{\lambda \geq 0,\ \mu \in \mathbb{R}^m} q(\lambda, \mu) = \inf_{x \in C} f(x) = f(x^*).$$

The conclusion of Lemma C.1 is twofold. First, if we find (x^*, λ^*, μ^*) fulfilling the first order, the feasibility and the slackness conditions, then x^* minimizes the primal problem (C.1). Second, there is no duality gap, and (λ^*, μ^*) is solution of the dual problem (C.5).

We underline that the slackness and the feasibility conditions enforce two conditions:

1. both λ_j^* and $-g_j(x^*)$ must be non-negative;
2. at least one of these two quantities is zero.

The second condition enforces that if $g_j(x^*) < 0$, then $\lambda_j^* = 0$. In this case, where $g_j(x^*) < 0$, we say that the condition is not active.

Proof of Lemma C.1. Since L is convex in x, and since x^* is a critical point for $x \to L(x, \lambda^*, \mu^*)$, we have

$$L(x^*, \lambda^*, \mu^*) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*, \mu^*) = q(\lambda^*, \mu^*).$$

By the feasibility and the slackness conditions, we then have

$$q(\lambda^*, \mu^*) = L(x^*, \lambda^*, \mu^*) = f(x^*) + \sum_{j=1}^{N} \underbrace{\lambda_j^* g_j(x^*)}_{=0} + \sum_{i=1}^{m} \mu_i^* \underbrace{\ell_i(x^*)}_{=0} = f(x^*). \tag{C.6}$$

Furthermore, since $\lambda^* \geq 0$ and since $x^* \in C$, the weak-duality ensures that

$$q(\lambda^*, \mu^*) \leq \sup_{\lambda \geq 0,\ \mu \in \mathbb{R}^m} q(\lambda, \mu) \leq \min_{x \in C} f(x) \leq f(x^*).$$

Combining this inequality with (C.6), we get the conclusion. □

Index