Mathematics for Artificial Intelligence

Christophe Giraud

Université Paris Saclay, 2020
Foreword

These lecture notes cover the program of the course “Mathematics for Artificial Intelligence II” 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 is the second part of a year-long course on the mathematics of artificial intelligence. It 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 6–9.

While this course can be followed independently of the first course “Mathematics for Artificial Intelligence I”, we highly recommend to work out this first course, in order to get a good introduction to the mathematical theory of machine learning.

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

Enjoy your reading!

Orsay, May 2020

Christophe
Contents

Foreword iii

1 Sub-Gaussian random variables 1
  1.1 Refresher on Gaussian random variable 1
  1.2 SubGaussian random variables 1
    1.2.1 Definition and examples 1
    1.2.2 Tails of subGaussian random variables and Hoeffding inequality 2
    1.2.3 Moments of subGaussian random variables 3
  1.3 Problem: Median of Means 5
    1.3.1 Median of Means estimator 5
    1.3.2 Illustration 6

I Sequential learning 9

2 Sequential learning and optimisation 11
  2.1 Introduction 11
    2.1.1 Batch statistical learning 11
    2.1.2 Sequential statistical learning 12
  2.2 Online learning with stochastic gradient descent 13
    2.2.1 Stochastic gradient descent 13
    2.2.2 SGD for optimisation and batch learning 15
  2.3 Problem: Projected Gradient Descent 17
    2.3.1 Projection over a convex set 17
    2.3.2 Rate for Lipschitz convex functions 17
    2.3.3 Fast rates for strongly convex functions 18

3 Prediction with experts 21
  3.1 Introduction 21
    3.1.1 The learning problem 21
    3.1.2 The regret of a strategy 21
  3.2 Warm-up: aggregation with PGD 22
    3.2.1 PGD for expert aggregation 22
    3.2.2 Linearized problem 22
  3.3 Aggregation with exponential updates 23
    3.3.1 Exponential updates in the linearized problem 23
    3.3.2 Faster aggregation rates for square loss 25
    3.3.3 Exponential updates for the original problem 26
  3.4 Mirror descent 27
    3.4.1 Changing the geometry in gradient descent 27
    3.4.2 Regret bound for Projected Mirror Descent 29
    3.4.3 Problem: Projected Mirror Descent for expert aggregation 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>Regret bounds for stochastic sequences</td>
<td>34</td>
</tr>
<tr>
<td>3.6</td>
<td>Illustration</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>Multi-Armed bandits</td>
<td>39</td>
</tr>
<tr>
<td>4.1</td>
<td>Setting</td>
<td>39</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Bandits problems</td>
<td>39</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Modeling</td>
<td>40</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Regret</td>
<td>41</td>
</tr>
<tr>
<td>4.2</td>
<td>UCB strategy</td>
<td>42</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Optimism in face of uncertainty</td>
<td>42</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Fixed time horizon UCB</td>
<td>43</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Horizon free UCB</td>
<td>44</td>
</tr>
<tr>
<td>4.3</td>
<td>Lower bounds</td>
<td>45</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Kullback-Leibler divergence</td>
<td>45</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Asymptotic lower bounds</td>
<td>48</td>
</tr>
<tr>
<td>4.4</td>
<td>Problem: X-armed bandits</td>
<td>50</td>
</tr>
<tr>
<td>4.5</td>
<td>Illustration of UCB</td>
<td>51</td>
</tr>
<tr>
<td>5</td>
<td>Lower bounds</td>
<td>53</td>
</tr>
<tr>
<td>5.1</td>
<td>Minimax risk</td>
<td>53</td>
</tr>
<tr>
<td>5.2</td>
<td>A recipe for proving lower bounds</td>
<td>54</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Fano’s lemma</td>
<td>54</td>
</tr>
<tr>
<td>5.2.2</td>
<td>From Fano’s lemma to a lower bound over a finite set</td>
<td>57</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Back to the original problem: finding a good discretization</td>
<td>58</td>
</tr>
<tr>
<td>5.3</td>
<td>Illustration</td>
<td>58</td>
</tr>
<tr>
<td>II</td>
<td>Matrix analysis for Machine Learning</td>
<td>61</td>
</tr>
<tr>
<td>6</td>
<td>Singular Value Decomposition</td>
<td>63</td>
</tr>
<tr>
<td>6.1</td>
<td>Reminder on spectral decomposition of symmetric real matrices</td>
<td>63</td>
</tr>
<tr>
<td>6.2</td>
<td>Singular Value Decomposition (SVD)</td>
<td>64</td>
</tr>
<tr>
<td>6.3</td>
<td>Matrix analysis</td>
<td>67</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Matrix Norms</td>
<td>67</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Low rank projection</td>
<td>68</td>
</tr>
<tr>
<td>6.4</td>
<td>Explicit computations with SVD decomposition</td>
<td>70</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Moore–Penrose Pseudo-Inverse</td>
<td>70</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Problem: Ridge regression</td>
<td>70</td>
</tr>
<tr>
<td>7</td>
<td>Perturbation bounds</td>
<td>73</td>
</tr>
<tr>
<td>7.1</td>
<td>Singular values localization</td>
<td>73</td>
</tr>
<tr>
<td>7.2</td>
<td>Eigenspaces localization</td>
<td>73</td>
</tr>
<tr>
<td>7.3</td>
<td>Operator norm of random matrices</td>
<td>76</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Concentration of quadratic forms of Gaussian vectors</td>
<td>76</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Concentration of random Gram matrices</td>
<td>78</td>
</tr>
<tr>
<td>7.4</td>
<td>Exercises</td>
<td>81</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Generalized Weyl inequalities</td>
<td>81</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Cardinality of an ( \varepsilon )-net</td>
<td>81</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Limit distribution of singular values of random matrices</td>
<td>81</td>
</tr>
<tr>
<td>8</td>
<td>Principal Component Analysis</td>
<td>83</td>
</tr>
<tr>
<td>8.1</td>
<td>Principal Component Analysis</td>
<td>83</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>8.1.1</td>
<td>Finding the best low dimensional linear representation of data</td>
<td>83</td>
</tr>
<tr>
<td>8.1.2</td>
<td>Illustration</td>
<td>85</td>
</tr>
<tr>
<td>8.2</td>
<td>Interpreting PCA</td>
<td>86</td>
</tr>
<tr>
<td>8.3</td>
<td>Theory for PCA</td>
<td>90</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Recovering a low dimensional signal</td>
<td>90</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Dimension selection</td>
<td>92</td>
</tr>
<tr>
<td>8.4</td>
<td>Exercises</td>
<td>93</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Rank recovery</td>
<td>93</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Implementing a PCA with R</td>
<td>94</td>
</tr>
<tr>
<td>9</td>
<td>Clustering</td>
<td>95</td>
</tr>
<tr>
<td>9.1</td>
<td>Cluster model</td>
<td>95</td>
</tr>
<tr>
<td>9.2</td>
<td>Local algorithms</td>
<td>96</td>
</tr>
<tr>
<td>9.3</td>
<td>Spectral clustering</td>
<td>98</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Spectral clustering recipe</td>
<td>98</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Recovery bounds</td>
<td>100</td>
</tr>
<tr>
<td>9.4</td>
<td>Exercises</td>
<td>103</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Sterling numbers of second kind</td>
<td>103</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Exact recovery with hierarchical clustering</td>
<td>104</td>
</tr>
<tr>
<td>9.4.3</td>
<td>Estimating $\Gamma$</td>
<td>104</td>
</tr>
<tr>
<td>9.4.4</td>
<td>Illustration of hierarchical clustering and spectral clustering</td>
<td>106</td>
</tr>
<tr>
<td>Appendix A</td>
<td>Gaussian Distribution</td>
<td>109</td>
</tr>
<tr>
<td>A.1</td>
<td>Gaussian Random Vectors</td>
<td>109</td>
</tr>
<tr>
<td>A.2</td>
<td>Chi-Square Distribution</td>
<td>110</td>
</tr>
<tr>
<td>A.3</td>
<td>Gaussian Conditioning</td>
<td>110</td>
</tr>
<tr>
<td>Appendix B</td>
<td>Constrained optimization</td>
<td>113</td>
</tr>
<tr>
<td>B.1</td>
<td>Dual problem</td>
<td>113</td>
</tr>
<tr>
<td>B.1.1</td>
<td>Lagrangian and dual functions</td>
<td>113</td>
</tr>
<tr>
<td>B.1.2</td>
<td>Weak duality</td>
<td>113</td>
</tr>
<tr>
<td>B.1.3</td>
<td>Finding a solution</td>
<td>114</td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>115</td>
</tr>
</tbody>
</table>
Chapter 1

Sub-Gaussian random variables

1.1 Refresher on Gaussian random variable

A real random variable \( X \) follows a \( N(0, \sigma^2) \) 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 \( N(0, 1) \) Gaussian distribution then \( X = \sigma Z \) follows a \( N(0, \sigma^2) \) Gaussian distribution.

The moment generating function, or Laplace transform, of a random variable \( X \) with \( N(0, \sigma^2) \) Gaussian distribution is given by

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

1.2 SubGaussian random variables

1.2.1 Definition and examples

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

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

In addition, we write \( X \in \text{subG}(\mu, \sigma^2) \) if \( X - \mu \in \text{SubG}(\sigma^2) \).

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

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

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 \( X - \mathbb{E}[X] \in \text{subG}(b - a)^2/4) \).*

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

\[
\psi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]} \quad \text{and} \quad \psi''(s) = \frac{\mathbb{E}[X^2e^{sX}]}{\mathbb{E}[e^{sX}]} - \left( \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]} \right)^2.
\]
Setting
\[d\mathbb{P}_s(\omega) = \frac{e^{sX(\omega)}}{\mathbb{E}[e^{sX}]} \, d\mathbb{P}(\omega),\]
we observe that
\[\psi''(s) = \mathbb{E}_s[X^2] - \mathbb{E}_s[X]^2 = \mathbb{E}_s[(X - \mathbb{E}_s[X])^2] \leq \mathbb{E}_s \left[ \left( X - \frac{a + b}{2} \right)^2 \right] \leq \frac{(b - a)^2}{4}.
\]
Hence, Taylor expansion gives
\[
\psi(s) = \psi(0) + s\psi'(0) + \int_0^s \int_{x=0}^u \psi''(x) \, dx \, du \leq 0 + 0 + \frac{(b - a)^2}{8},
\]
so, \(X \in \text{subG}((b - a)^2/4).\)

**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 shows that subGaussian random variables are stable under independent sums.

**Lemma 1.2 Sum of subGaussian random variables.**
If \(X_1, \ldots, X_n \in \text{subG}(1)\) are \(n\) independent random variables, then
\[
\sum_{i=1}^n a_i X_i \in \text{subG}(\|a\|^2).
\]

**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
\[
\mathbb{E} \left[ \exp \left( s \sum_{i=1}^n a_i X_i \right) \right] = \prod_{i=1}^n \mathbb{E} \left[ \exp \left( s a_i X_i \right) \right] \leq \prod_{i=1}^n e^{(a_i s)^2/2} = e^{\|a\|^2 s^2/2}.
\]
Hence \(\sum_{i=1}^n a_i X_i \in \text{subG}(\|a\|^2).\) \(\square\)

### 1.2.2 Tails of subGaussian random variables and Hoeffding 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 \(X \in \text{subG}(\sigma^2)\) fulfills for any \(t \geq 0\)
\[
\mathbb{P}[X \geq t] \leq e^{-t^2/(2\sigma^2)}.
\]

**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. \(\square\)
Corollary 1.4 Tail for sums of subGaussian random variables.
If $X_1, \ldots, X_n \in \text{subG}(\sigma^2)$ 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 $X_i - \mu \in \text{subG}(\sigma^2)$, 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 $X_1, \ldots, X_p \in \text{subG}(\sigma^2)$ be $n$ independent random variables. Then, for any $L \geq 0$, we have

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

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 $s > 0$

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

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

1.2.3 Moments of subGaussian random variables

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

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

$$\mathbb{E} \left[ X^{2k} \right] \leq 2^{k+1} k! \sigma^{2k} \quad \text{for} \ k \geq 1,$$

$$\mathbb{E} \left[ e^{s(X^2 - \mathbb{E}[X^2])] \right] \leq 1 + 64(s\sigma^2)^2 \leq e^{64(s\sigma^2)^2} \quad \text{for} \ (s\sigma^2)^2 \leq 1/32.$$

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

$$\mathbb{E} \left[ |X|^j \right] = \int_0^\infty \mathbb{P} \left[ |X|^j \geq t \right] dt = 2^{j/2-1} \int_0^\infty \mathbb{P} \left[ |X| \geq \sqrt{2u} \right] u^{j/2-1} du \leq 2^{j/2} \int_0^\infty e^{-u} u^{j/2-1} du = 2^{j/2} j (j/2) = 2^{j/2+1} \Gamma(j/2 + 1).$$

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

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

$$\mathbb{E} \left[ e^{s(X^2 - \mathbb{E}[X^2])} \right] \leq \frac{1}{2} \mathbb{E} \left[ e^{2sX^2} + e^{-2s\mathbb{E}[X^2]} \right] \leq \frac{1}{2} \mathbb{E} \left[ e^{2sX^2} + e^{-2sX^2} \right] = \mathbb{E} \left[ \text{ch}(2sX^2) \right] = 1 + \sum_{k \geq 1} \frac{(2s)^{2k}}{(2k)!} \mathbb{E}[X^{4k}].$$

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

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

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

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 $\text{subG}(1)$. Then, for any $t \geq 0$

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

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

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

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 $\phi$ is achieved for $s_\ast = t/(128\|a\|^2)$. Hence, we consider apprart the two cases $32|a|_\infty^2 s_\ast^2 \leq 1$ and $32|a|_\infty^2 s_\ast^2 \geq 1$.

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

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

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

$$\min_{32|a|_\infty^2 s^2 \leq 1} \phi(s) = \phi \left( \frac{1}{|a|_\infty \sqrt{32}} \right) = \frac{-t}{|a|_\infty \sqrt{32}} + \frac{64\|a\|^2}{32|a|_\infty^2} \geq \frac{-t}{|a|_\infty \sqrt{32}} + \frac{t}{2|a|_\infty \sqrt{32}} = \frac{-t}{|a|_\infty \sqrt{128}}.$$
PROBLEM: MEDIAN OF MEANS

where the inequality follows from $32|a|^2_s s^2 \geq 1$ with $s = t/(128||a||^2)$. We then have proved that for any $t \geq 0$

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

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

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

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

1.3 Problem: Median of Means

1.3.1 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 $L \geq 0$

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

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 $X \in subG(\sigma^2)$, 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 $X \in subG(\sigma^2)$.

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 $\hat{\mu}$ of $\mu$ 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

$$\Pr\left[ \bar{X}_{B_l} - \mu \geq \frac{2\sigma}{\sqrt{m}} \right] \leq \frac{1}{4}.$$

2. Check that $\hat{\mu}_K = \text{median}(\bar{X}_{B_1}, \ldots, \bar{X}_{B_K})$ fulfills

$$\Pr\left[ \hat{\mu}_K - \mu \geq \frac{2\sigma}{\sqrt{m}} \right] \leq \Pr[\text{Binomial}(K, 1/4) \geq K/2].$$

3. Conclude that

$$\Pr\left[ \hat{\mu}_K - \mu \geq 2\sigma \sqrt{\frac{K}{n}} \right] \leq e^{-K/8}.$$
This last deviation bound is similar (up to constants) to (1.1) with \( L = K/8 \). Notice yet the two important features:

- the estimator \( \hat{\mu}_K \) 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 \( n/16 \).

### 1.3.2 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](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](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 \( \hat{\mu}_K \) 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 \( Q_0 \lor (Q_1 - 1.5(IQ)) \) and right value \( Q_4 \land (Q_3 + 1.5(IQ)) \). Finally, if some values fall outside the interval, they are represented as dots. We often refer to these values as "outliers".

![Figure 1.1: Description of the boxplot representation of a set of values.](image)

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, \( \bar{X}_n \sim N(\mu, \sigma^2/n) \), 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 $\hat{\mu}_K$ 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.
Part I

Sequential learning
2.1 Introduction

Let us describe informally the topic of this part of the lectures.

2.1.1 Batch statistical learning

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.

In statistical learning, we assume that the couple \((X, Y)\) is a random variable with distribution \( P_{(X, Y)} \).

For a specified function \( \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ \), usually called "loss function", and a predictor \( h : \mathcal{X} \to \mathcal{Y} \), we define the so-called risk of the predictor \( h \) by

\[
    r(h) = E_{(X, Y) \sim P_{(X, Y)}} [\ell(Y, h(X))].
\]

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

In practice, the distribution \( P_{(X, Y)} \) is unknown, so we can neither compute \( r(h) \) nor \( h^* \). Instead, we have access to a sample \( Z = (X_i, Y_i)_{i=1,...,n} \) gathering \( n \) observations i.i.d. with distribution \( P_{(X, Y)} \).

We then build a mapping \( F : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{Y} \) and predict \( Y \) with \( F(Z, X) \), resulting with a risk \( E_{(X, Y) \sim P_{(X, Y)}} [\ell(Y, F(Z, X))] \). The central question is then, how to choose and compute the mapping \( F \)?

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

\[
    f(\theta) = E_{(X, Y) \sim P_{(X, Y)}} [\ell(Y, h(\theta, X))], \quad \text{for } \theta \in \Theta,
\]

the goal is to design some \( \sigma(Z) \)-measurable variable \( \tilde{\theta} = g(Z) \), such that \( f(\tilde{\theta}) \) 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"
  \[
  f(\tilde{\theta}) - \min_{\theta \in \Theta} f(\theta),
  \]
  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 \( \tilde{\theta}(\omega) = g(Z(\omega)) \) and he/she uses
the function \( h(\hat{\theta}(\omega), \cdot) : X \rightarrow Y \) for prediction. Yet, when we investigate the statistical properties of the predictor, we consider \( Z \), and hence \( \hat{\theta} \), as a random variable. The goal is to understand the distribution of the risk \( f(\hat{\theta}) \) or the distribution of the parameter \( \hat{\theta} \) with respect to the randomness of \( Z \).

### 2.1.2 Sequential statistical learning

In batch statistical learning, we have access from the start to a whole data set \( Z = (X_i, Y_i)_{i=1,...,n} \). 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 next three chapters, 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 \( t = 1, 2, \ldots \), we will have to choose an "action" \( \hat{\theta}_t \) from available data at time \( t \). Each action \( \hat{\theta}_t \) has a risk or "cost" \( f_t(\hat{\theta}_t) \) and our goal will be to find strategies \( \hat{\theta} \) such that the cumulated cost

\[
\sum_{t=1}^{T} f_t(\hat{\theta}_t),
\]

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 \( X_t \in X \). The prediction is assumed to be of the form \( h(\hat{\theta}_t, X_t) \), where \( h(\theta, x) \) is a prescribed function. For a given loss function \( \ell : \mathbb{R}^2 \rightarrow \mathbb{R}_+ \), the "cost" associated to the choice of a parameter \( \theta \) is the integrated loss

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

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 \( h_1(t), \ldots, h_d(t) \) 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 \( \ell \), we will consider convex combinations of the predictions \( \sum_{k=1}^{d} \hat{\theta}_{k,t} h_k(t) \) and we will seek to minimise the cumulated loss

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

3) **Multi-armed bandit problems.**

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

\[
\sum_{t=1}^{T} \mathbb{E}[Z_t|\hat{\theta}_t] = \sum_{t=1}^{T} \mu(\hat{\theta}_t).
\]
or equivalently minimizing the regret
\[ \sum_{t=1}^{T} \left( \max_{\theta \in \Theta} \mu(\theta) - \mu(\hat{\theta}_t) \right). \]

2.2 Online learning with stochastic gradient descent

2.2.1 Stochastic gradient descent

Let us consider the sequential learning problem, where at time \( t \) we want to predict a real valued outcome \( Y_t \) from covariates \( X_t \) via a predictor \( h(\hat{\theta}_t, X_t) \). More precisely, we consider \( \Theta \subset \mathbb{R}^d \), a set \( X \) and a prescribed regression function \( h : \Theta \times X \rightarrow \mathbb{R} \), differentiable in the first variable \( \theta \). Typical examples of regression functions are
- linear predictor \( h(\theta, x) = \langle \theta, x \rangle \),
- logit predictor \( h(\theta, x) = \exp((\theta, x))/(1 + \exp((\theta, x))) \),
- kernel predictor \( h(\theta, x) = \sum_{i=1}^{d} \theta_i k(z_i, x) \),
- neural networks, etc.

For a parameter \( \theta \) and a differentiable loss function \( \ell : \mathbb{R}^2 \rightarrow \mathbb{R}_+ \), 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 across time \( t \), and hence we will compare ourselves to the best "constant" strategy
\[ \theta^* \in \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} f_t(\theta), \]
which we assume to exist.

We cannot use the parameter \( \theta^* \), as we do not have access to the functions \( f_t \). Actually, we have to choose the parameter \( \theta_t \) according to the information \( \mathcal{F}_{t-1} = \sigma((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})) \) available after the step \( t - 1 \).

Let us assume for simplicity that \( \Theta = \mathbb{R}^d \). A simple recipe for minimizing a differentiable function is to apply a gradient descent. Here we do not have access to the gradient of \( f_t \), but we can compute after step \( t \)
\[ g_t = \nabla_{\theta} \ell(Y_t, h(\hat{\theta}_t, X_t)). \]

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
\[ \mathbb{E} [\nabla_{\theta} \ell(Y_t, h(\theta, X_t))] = \nabla_{\theta} \mathbb{E} [\ell(Y_t, h(\theta, X_t))], \quad \text{for all } \theta \in \mathbb{R}^d, \]
the gradient \( g_t \) fulfills
\[ \mathbb{E} [g_t | \mathcal{F}_{t-1}] = \nabla f_t(\hat{\theta}_t). \]

A random vector \( g_t \) fulfilling the above property is called a stochastic gradient of \( f_t \) at \( \hat{\theta}_t \).

Motivated by the above example, we consider below the generic problem where, for a sequence of differentiable convex functions \( \{f_t : t = 1, \ldots, T\} \), and a filtration \( \mathcal{F}_t \), we seek to minimize the regret
\[ \sum_{t=1}^{T} (f_t(\theta_t) - f_t(\theta^*)), \]
under the constraint that \( \theta_{t+1} \) only has access to the past values \( \theta_t \), which is \( \mathcal{F}_{t-1} \)-measurable and to a random vector \( g_t \) fulfilling
\[ \mathbb{E} [g_t | \mathcal{F}_{t-1}] = \nabla f_t(\theta_t). \]
A simple algorithm for such a setting is a variant of the gradient descent algorithm, where the gradient \( \nabla f_t(\theta_t) \) is replaced by the stochastic gradient \( g_t \).

**Stochastic Gradient Descent (SGD)**

**Input:** \( \eta > 0 \), and \( \theta_1 \).

**Iterate:** for \( t = 1, \ldots, T - 1 \),
\[
\theta_{t+1} = \theta_t - \eta g_t
\]

**Remark.** In the special case where \( g_t = \nabla f_t(\theta_t) \), the SGD reduces to the classical gradient descent.

We have the following upper bound on the regret.

**Theorem 2.1 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 \( (\mathcal{F}_t)_{t \geq 1} \) be a filtration. Define the sequence \( (\theta_t)_{t \geq 1} \) by induction,
\[
\theta_{t+1} = \theta_t - \eta g_t, \quad t \geq 1, \quad \text{with} \quad \eta = \frac{R}{L \sqrt{T}}.
\]

where \( R = \|\theta_1 - \theta^*\| \) and where \( g_t \) is a random variable fulfilling
\[
\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f_t(\theta_t), \quad \text{and} \quad \mathbb{E}[\|g_t\|^2] \leq L^2, \quad \text{for} \ t = 1, \ldots, T.
\]

Then, the mean regret of the sequence of the SGD is upper-bounded by
\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta \in \mathbb{R}^d} \frac{1}{T} \sum_{t=1}^{T} f_t(\theta) \right] \leq \frac{LR}{\sqrt{T}}.
\]

**Remark.** Under the assumption of the theorem, the regret of SGD scales as \( 1/\sqrt{T} \) for the choice \( \eta = R/(L \sqrt{T}) \). Yet, we emphasize that choosing this \( \eta \) can be infeasible in practice, as the constant \( R \) and \( L \) are usually unknown. In Section 2.3 on Projected Gradient Descent, we will be able to overcome this issue.

**Proof of Theorem 2.1.** From the polarisation formula \( 2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|b - a\|^2 \), we get
\[
2\langle \eta g_t, \theta_t - \theta^* \rangle = \|\theta_t - \theta^*\|^2 + \eta^2\|g_t\|^2 - \|\theta_t - \theta^* - \eta g_t\|^2
\]
\[
= \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2 + \eta^2\|g_t\|^2.
\]

Summing over \( t \) gives
\[
\sum_{t=1}^{T} \langle g_t, \theta_t - \theta^* \rangle = \frac{\|\theta_1 - \theta^*\|^2 - \|\theta_{T+1} - \theta^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2.
\]
ONLINE LEARNING WITH STOCHASTIC GRADIENT DESCENT

As $f_t(\theta_t) - f_t(\theta^*) \leq \langle \nabla f_t(\theta_t), \theta_t - \theta^* \rangle$ and $\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f_t(\theta_t)$, we get

$$
\mathbb{E} \left[ \sum_{t=1}^{T} f_t(\theta_t) - f_t(\theta^*) \right] \leq \mathbb{E} \left[ \sum_{t=1}^{T} \langle \nabla f_t(\theta_t), \theta_t - \theta^* \rangle \right]
$$

$$
= \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{E}[(g_t, \theta_t - \theta^*)|\mathcal{F}_{t-1}] \right]
$$

$$
= \mathbb{E} \left[ \sum_{t=1}^{T} (g_t, \theta_t - \theta^*) \right]
$$

$$
\leq \frac{R^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E} \left[ \|g_t\|^2 \right] = \frac{R^2}{2\eta} + \frac{\eta}{2} L^2 T.
$$

The result of Theorem 2.1 follows by plugging the value $\eta = R/(L \sqrt{T})$. \hfill \Box

2.2.2 SGD for optimisation and batch learning

We underline in this section that SGD can be used for minimizing some convex function

$$\min_{\theta \in \mathbb{R}^d} f(\theta).$$

Minimisation of a convex function

We can notice that the above theorem only used the following assumption: $f_t$ is convex and differentiable, $g_t$ fulfills $\mathbb{E}[g_t|\mathcal{F}_{t-1}] = \nabla f_t(\theta_t)$, $\mathbb{E}[\|g_t\|^2] \leq L^2$.

In the case where $f_t = f$ for all $t$, we have $\theta^* \in \arg\min_{\theta} f(\theta)$ and by Jensen inequality

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

So under the hypotheses of Theorem 2.1, the mean $\bar{\theta}_T$ of the sequence $\{\theta_t : t = 1, \ldots, T\}$ output by the SGD fulfills

$$\mathbb{E} \left[ f(\bar{\theta}_T) - f(\theta^*) \right] \leq \frac{RL}{\sqrt{T}}.$$

Empirical loss minimisation in batch learning

Aside of the sequential context presented above, the SGD is widely used in the batch setting for minimizing (penalized) empirical risk. It was already implemented in the early 60’s to train a linear regression on the first generations of computers.

Let us assume that we observe some i.i.d. data $Z = (X_i, Y_i)_{i=1,\ldots,n}$ and we want to minimize the penalized empirical risk

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

where $\ell$ is some loss and $\Omega(\theta)$ is a convex penalty, for example $\Omega(\theta) = \lambda \|\theta\|^2$. Let us set $F_t(\theta) = \ell(Y_t, h(\theta, X_t)) + \Omega(\theta)$. For a random variable $I$ independent of the data and with a uniform distribution on $\{1, \ldots, n\}$, we have $f(\theta) = \mathbb{E}[F_I(\theta)|Z]$. If $f$ is differentiable, we also have $\nabla f(\theta) = \mathbb{E}[\nabla F_I(\theta)|Z]$, so the minimisation of $f$ is amenable to a SGD.

Actually, let us consider a sequence of i.i.d. uniform random variables $I_1, I_2, \ldots, I_n$ independent of the data $Z = (X_i, Y_i)_{i=1,\ldots,n}$. Setting $F_{t-1} = \sigma(Z, I_1, \ldots, I_{t-1})$, as $I_t$ is independent of...
\(\mathcal{F}_{t-1}\), we have \(\nabla f(\theta_t) = \mathbb{E}[\nabla F_t(\theta_t)|\mathcal{F}_{t-1}]\) for any \(\theta_t\) which is \(\mathcal{F}_{t-1}\)-measurable. So, we can apply the SGD scheme with \(g_t = \nabla F_t(\theta_t)\)

\[
\theta_{t+1} = \theta_t - \eta \nabla F_t(\theta_t), \quad t \geq 1,
\]
to minimize the penalized empirical loss \(f(\bar{\theta})\). Under convexity assumptions, we get a convergence of \(f(\bar{\theta}_T)\) to \(\min_\theta f(\theta)\) at a rate proportional to \(1/\sqrt{T}\).

An important advantage of the SGD compared to the vanilla gradient descent

\[
\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t), \quad t \geq 1,
\]
is that at each time step, we only need to compute \(\nabla F_t(\theta_t)\) instead of \(\nabla f(\theta_t) = \frac{1}{n} \sum_{i=1}^n \nabla F_i(\theta_t)\), reducing the computational complexity by a factor \(n\). When the dataset are large, with millions of samples, the speed-up is substantial.

**Remark.** At first sight, the application of SGD for empirical risk minimization looks similar to the application of SGD in online learning. We emphasize yet two important differences:

- For empirical risk minimisation, a data point \((X_i, Y_i)\) can be involved in multiple steps. The algorithm does “multiple passes” on the data. A contrario, in online learning, each data point \(Z_t = (Y_t, X_t)\) is used only once. The algorithm does a “single pass” on the data.
- The goal in empirical risk minimization and online learning are different. In the first case, we seek to minimize the (penalized) empirical risk

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

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

\[
f(\theta_1, \ldots, \theta_T) = \sum_{t=1}^T \mathbb{E}_{(X_t, Y_t)} [\ell(Y_t, h(\theta_t, X_t))],
\]
as small as possible.

**SGD with time-dependent step size**

We observe that the step size \(\eta = R/(L \sqrt{T})\) depends on the time horizon \(T\) which can be an issue in practice. We may wish to use instead a time-dependent step size, for example \(\eta_t = R/(L \sqrt{t})\). Let us investigate the SGD with updates \(\theta_{t+1} = \theta_t - \eta_{t+1} g_t\). Following the same lines as in the proof of Theorem 2.1, we get

\[
\mathbb{E} \left[ \sum_{t=1}^T \eta_{t+1} (f(\theta_t) - f(\theta^*)) \right] \leq \frac{R^2}{2} + \frac{L^2}{2} \sum_{t=1}^T \eta_{t+1}^2.
\]

Let us set

\[
\bar{\theta}_T = \frac{\sum_{t=1}^T \eta_{t+1} \theta_t}{\sum_{t=1}^T \eta_{t+1}}.
\]

According to Jensen inequality, we get

\[
\mathbb{E} \left[ f(\bar{\theta}_T) - f(\theta^*) \right] \leq \mathbb{E} \left[ \frac{\sum_{t=1}^T \eta_{t+1} (f(\theta_t) - f(\theta^*))}{\sum_{s=1}^T \eta_{s+1}} \right] \leq \frac{R^2 + L^2 \sum_{t=1}^T \eta_{t+1}^2}{2 \sum_{t=1}^T \eta_{t+1}}.
\]

Since for \(\eta_t = R/(L \sqrt{t})\) we have for some constant \(C > 0\)

\[
\sum_{i=1}^T \eta_{t+1}^2 \leq \frac{R^2 \log(T + 1)}{L^2} \quad \text{and} \quad \sum_{i=1}^T \eta_{t+1} \geq \frac{CR \sqrt{T}}{L},
\]

we get

\[
\mathbb{E} \left[ f(\bar{\theta}_T) - f(\theta^*) \right] \leq \frac{R^2 \log(T + 1)}{2L^2} + \frac{CR L \sqrt{\log(T + 1)}/L}{2}.
\]
so
\[ E \left[ f(\tilde{\theta}_T) - f(\theta^*) \right] \leq C' \frac{RL \log(T)}{\sqrt{T}}. \]

We obtain here an additional \( \log(T) \) term in the bound. This log term can be avoided yet as you can check in the exercise below.

**Exercise.** Consider the step size \( \eta_t = \frac{R}{(L \sqrt{t})} \) and define now
\[ \tilde{\theta}_T = \frac{\sum_{t=T/2}^T \eta_{t+1} \theta_t}{\sum_{s=T/2}^T \eta_{s+1}}. \]

Assuming that \( \|\theta_{T/2} - \theta^*\|^2 \leq R^2 \), check that we have
\[ E \left[ f(\tilde{\theta}_T) - f(\theta^*) \right] \leq C'' \frac{RL}{\sqrt{T}}. \]

### 2.3 Problem: Projected Gradient Descent

We consider a (non-empty) compact convex set \( C \subseteq \mathbb{R}^d \), and a convex function \( f: D \to \mathbb{R} \), with \( C \) included in the interior of \( D \). We consider in this section the constrained optimisation problem
\[ \min_{x \in C} f(x). \tag{2.1} \]

We assume below that there exists a unique \( x^* \in C \) solution to this minimisation problem.

#### 2.3.1 Projection over a convex set

As \( C \) is convex, the set of minimizers \( \arg\min_{u \in C} \|x - u\|^2 \) is non-empty. We denote by
\[ \pi_C x \in \arg\min_{u \in C} \|x - u\|^2 \]

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
\[ \langle u - \pi_C z, z - \pi_C z \rangle \leq 0 \quad \text{and} \quad \|\pi_C z - z\|^2 + \|u - \pi_C z\|^2 \leq \|u - z\|^2. \tag{2.2} \]
3. Prove that the projection \( \pi_C x \) of \( x \) onto the convex set \( C \) is uniquely defined.

#### 2.3.2 Rate for Lipschitz convex functions

We assume in the following that \( f \) is differentiable on the interior of \( D \). To solve (2.1), we can apply the Projected Gradient Descent algorithm (with \( \eta > 0 \)):

**Projected Gradient Descent (PGD)** Input: \( \eta > 0 \), and \( x_1 \in C \).
Iterate: for \( t = 1, \ldots, T - 1 \), \( x_{t+1} = \pi_C (x_t - \eta \nabla f(x_t)) \)
Output: \( \bar{x}_T \) or \( x_T \)

We will prove the following theorem.
Theorem 2.2 Rate for Lipschitz convex function.
Assume that \( \max_{x \in C} \| \nabla f(x) \| \leq L \) and that \( \max_{x, y \in C} \| x - y \| \leq R \). Then, for \( \eta = R/(L \sqrt{T}) \)
\[
f(\bar{x}_T) - f(x^*) \leq \frac{LR}{\sqrt{T}}.
\]

Remark. We get a bound similar to the bound in Theorem 2.1. Compared to Theorem 2.1, 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 \( f \).

Proof of Theorem 2.2.
For the analysis of the algorithm, we introduce the notation \( y_{t+1} = x_t - \eta \nabla f(x_t) \).

1. Similarly as in the unconstrained case, prove that
\[
f(x_t) - f(x^*) \leq \frac{1}{\eta} \langle x_t - y_{t+1}, x_t - x^* \rangle + \frac{\eta}{2} \left( \| x_t - x^* \|^2 - \| y_{t+1} - x^* \|^2 \right).
\]

2. With (2.2), prove that
\[
\frac{1}{T} \sum_{t=1}^T f(x_t) - f(x^*) \leq \frac{\eta L^2}{2} + \frac{\| x_1 - x_* \|^2}{2\eta T}.
\]

3. Conclude.
The proof is complete. \( \square \)

Exercise. Extend the analysis above to the case where, as in Theorem 2.1, we only have access to a stochastic gradient \( g_t \) instead of \( \nabla f(\theta_t) \).

2.3.3 Fast rates for strongly convex functions
When the function \( f \) is strongly convex, then the PGD converges at a much faster rate. In this part, we prove the following theorem.

Theorem 2.3 Rate for strongly convex function.
Let us assume that \( f \) is \( \alpha \)-strongly convex
\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\alpha}{2} \| y - x \|^2,
\]
and that \( \nabla f \) is \( \beta \)-Lipschitz. Then, for \( \eta = 1/\beta \),
\[
\| x_{T+1} - x^* \|^2 \leq \| x_1 - x^* \|^2 e^{-\rho T}
\]
with \( \rho = \alpha/\beta \).

We define \( g(x) = \beta \left( x - \pi_C(x - \frac{1}{\beta} \nabla f(x)) \right) \). The key of the proof is the next lemma, which evaluates the progress made by one step of the PGD algorithm.

Lemma 2.4 We have
\[
f \left( \pi_C(x - \frac{1}{\beta} \nabla f(x)) \right) - f(y) \leq \langle g(x), x - y \rangle - \frac{1}{2\beta} \| g(x) \|^2 - \frac{\alpha}{2} \| x - y \|^2, \quad \forall x, y \in C.
\]

1. Let us first assume that (2.4) holds. Prove the sequence of (in)equalities
\[
\| x_{t+1} - x^* \|^2 = \| x_t - x^* \|^2 - \frac{2}{\beta} \langle g(x_t), x_t - x^* \rangle + \frac{1}{\beta^2} \| g(x_t) \|^2
\leq (1 - \rho) \| x_t - x^* \|^2 \leq e^{-\eta t} \| x_1 - x^* \|^2.
\]
PROBLEM: PROJECTED GRADIENT DESCENT

2. It remains to prove (2.4). With the mean value theorem, prove that

\[ f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle \, dt \leq \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \| y - x \|^2. \]

3. We set \( x^+ = \pi_C(x - \frac{1}{\beta} \nabla f(x)) \). With the above question and (2.3), check that

\[ f(x^+) - f(y) \leq \langle \nabla f(x), x^+ - x \rangle + \frac{\beta}{2} \| x^+ - x \|^2 + \langle \nabla f(x), x - y \rangle - \frac{\alpha}{2} \| y - x \|^2. \]

4. With (2.2), prove that \( \langle \nabla f(x), x^+ - y \rangle \leq \langle g(x), x^+ - y \rangle \) for all \( y \in C \).

5. Conclude that

\[
\begin{align*}
  f(x^+) - f(y) &\leq \langle g(x), x^+ - y \rangle + \frac{1}{2\beta} \| g(x) \|^2 - \frac{\alpha}{2} \| y - x \|^2 \\
  &= \langle g(x), x - y \rangle - \frac{1}{2\beta} \| g(x) \|^2 - \frac{\alpha}{2} \| y - x \|^2.
\end{align*}
\]
Chapter 3

Prediction with experts

3.1 Introduction

3.1.1 The learning problem

We consider the problem where 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 \(h(t) = (h_1(t), \ldots, h_d(t)) \in \mathbb{R}^d\) 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 \(I_t = (y(1), \ldots, y(t-1), h(1), \ldots, h(t)) \in \mathbb{R}^{(d+1)t-1}\). An aggregation strategy \(\hat{\theta}\) is a sequence of mappings \(\{\hat{\theta}_t : t \geq 1\} \) with \(\hat{\theta}_t : \mathbb{R}^{(d+1)t-1} \rightarrow S_d\), where \(S_d\) is the simplex \(S_d = \{x \in [0, 1]^d : |x|_1 = 1\}\). For a given strategy \(\hat{\theta}\), the outcome \(y(t)\) is predicted by the convex aggregation

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

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

3.1.2 The regret of a strategy

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

\[
\sum_{t=1}^{T} \ell (\langle \theta, h(t) \rangle, y(t))
\]

is as small as possible.

We will compare a strategy to a best constant aggregation strategy \(\theta^*\)

\[
\theta^* \in \arg\min_{\theta \in S_d} \sum_{t=1}^{T} \ell (\langle \theta, h(t) \rangle, y(t)).
\]

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

The regret of a strategy \(\hat{\theta}\) is defined as

\[
\mathcal{R}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \ell (\langle \theta_t, h(t) \rangle, y(t)) - \frac{1}{T} \sum_{t=1}^{T} \ell (\langle \theta^*, h(t) \rangle, y(t)).
\]

In most of this chapter, we will derive upper-bounds on \(\mathcal{R}(\hat{\theta})\), 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 PGD

3.2.1 PGD for expert aggregation

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

\[
\nabla f_t(\theta) = \partial_1 \ell ((\theta, h(t)), y(t)) h(t),
\]

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

\[
\theta_{t+1} = \pi_{S_d}(\theta_t - \eta \nabla f_t(\theta_t)), \quad t = 1, 2, \ldots,
\]

where \( \pi_{S_d} \) is the projection onto the simplex \( S_d \) as defined in Section 2.3.1. The regret can then be upper-bounded in terms of \( R = \max_{\theta, \theta' \in S_d} ||\theta - \theta'|| \) and

\[
L = \max_{t=1, \ldots, T} \max_{\theta \in S_d} ||\nabla f_t(\theta)|| = \max_{t=1, \ldots, T} \max_{\theta \in S_d} ||\partial_1 \ell ((\theta, h(t)), y(t)) h(t)||.
\]

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

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

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 \( \partial_1 \ell ((\theta, h(t)), y(t)) \) is bounded in absolute value by some constant \( C \) for any \( \theta \in S_d \) and \( t \geq 1 \) and hence

\[
\max_{t=1, \ldots, T} \max_{\theta \in S_d} ||\nabla f_t(\theta)||_\infty = \max_{t=1, \ldots, T} \max_{\theta \in S_d} ||\partial_1 \ell ((\theta, h(t)), y(t)) h(t)||_\infty \leq CM.
\]

Therefore we have \( L \leq CM \sqrt{d} \) and the upper-bounds of Chapter 2 give

\[
\mathcal{R}(\tilde{\theta}^{PGD}) \leq CM \sqrt{\frac{2d}{T}}.
\]

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

1. The regret (3.3) for PGD holds for any sequences \((y(t))_{t \geq 1}\) and \((h_j(t))_{t \geq 1}\), for \( j = 1, \ldots, d \), 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

\[
\ell_t = [\ell_{j,t}]_{j=1, \ldots, d} = [\ell(h_j(t), y(t))]_{j=1, \ldots, d}.
\]
AGGREGATION WITH EXPONENTIAL UPDATES

Since \( \ell \) is convex on the first variable and since \( \theta_i \in S_d \), Jensen inequality ensures the upper-bound

\[
\ell \left( \langle \theta_t, h(t) \rangle, y(t) \right) = \ell \left( \sum_{j=1}^{d} \theta_j h_j(t), y(t) \right) \leq \sum_{j=1}^{d} \theta_j \ell(h_j(t), y(t)) = \langle \theta_t, \ell \rangle.
\] (3.4)

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

\[
\mathcal{R}_{lin}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \langle \theta, \ell_t \rangle - \min_{\theta \in S_d} \frac{1}{T} \sum_{t=1}^{T} \langle \theta, \ell_t \rangle = \frac{1}{T} \sum_{t=1}^{T} \langle \theta, \ell_t \rangle - \min_{\theta \in S_d} \sum_{t=1}^{T} \ell_j t,
\] (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 PGD. 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 \( \|\ell_t\|_\infty \leq \sqrt{d}B \) and the upper-bounds of Chapter 2 give

\[
\mathcal{R}_{lin}(\hat{\theta}_{PGD}) \leq B \sqrt{\frac{2d}{T}}.
\] (3.6)

3.3 Aggregation with exponential updates

3.3.1 Exponential updates in the linearized problem

In the linearized problem with \( f_i(\theta) = \langle \theta, \ell_i \rangle \), the iterate at time \( t + 1 \) amounts to update \( \theta_t \) into \( \theta_t - \eta \ell_t \) and then to project the result on \( S_d \) according to \( \pi_{S_d} \). A glimpse at (3.5) shows that the optimal \( \theta \) 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 \( \ell^1 \)-norm. This lead us to the exponential weights strategy

\[
\theta_1 = \frac{1}{d}, \quad \theta_{t+1} = \frac{\theta_t e^{-\eta \ell_t}}{\|\theta_t e^{-\eta \ell_t}\|_1}, \quad t = 1, 2, \ldots,
\] (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 \( j = 1, \ldots, d \).

By a simple induction, we get the following closed-form formula for the weights \( \theta_t \) of the exponential weights strategy

\[
\theta_{j,t} = \frac{\exp \left( -\eta \sum_{s=1}^{t-1} \ell_{j,s} \right)}{\sum_{k=1}^{d} \exp \left( -\eta \sum_{s=1}^{t-1} \ell_{k,s} \right)}, \quad j = 1, \ldots, d, \quad t = 1, 2, \ldots
\] (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 \( (\ell_t)_{t \geq 1} \).
Remark. Before proving Lemma 3.1 and Theorem 3.2, let us comment on these last results.

According to (3.4) we have

Proof of Theorem 3.2.

We observe first that

\[ E = \frac{1}{T} \sum_{t=1}^{T} \ell ((\theta_t, h(t)), y(t)) - \min_{j=1, \ldots, d} \frac{1}{T} \sum_{t=1}^{T} \ell (h_j(t), y(t)) \leq B \sqrt{\frac{\log(d)}{2T}}. \]

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 \( j = 1, \ldots, d \), with values in \( \mathcal{Y} \), the exponential weight strategy (3.7) with \( \eta = \sqrt{\frac{8 \log(d)}{B^2 T}} \) fulfills

\[ \frac{1}{T} \sum_{t=1}^{T} \ell ((\theta_t, h(t)), y(t)) - \min_{j=1, \ldots, d} \frac{1}{T} \sum_{t=1}^{T} \ell (h_j(t), y(t)) \leq B \sqrt{\frac{\log(d)}{2T}}. \]

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

- We have the same scaling \( 1 / \sqrt{T} \) with respect to \( T \);
- The \( \sqrt{d} \) for PGD 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 PGD.

Proof of Theorem 3.2.

According to (3.4) we have

\[ \sum_{t=1}^{T} \ell ((\theta_t, h(t)), y(t)) - \min_{j=1, \ldots, d} \sum_{t=1}^{T} \ell (h_j(t), y(t)) \leq \sum_{t=1}^{T} \langle \theta_t, \ell_t \rangle - \min_{j=1, \ldots, d} \sum_{t=1}^{T} \ell_{j,t}. \quad (3.9) \]

The proof of Theorem 3.2 then simply follows from Lemma 3.1 with \( a = 0 \) and \( b = B \). \( \square \)

Proof of Lemma 3.1.

For any \( t = 1, \ldots, T \), let \( Z_t : \Omega \to [a, b] \) be a random variable with distribution

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

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

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

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)

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

and hence

\[ \mathbb{E}[e^{-\eta Z_t}] = \sum_{j=1}^{d} \frac{e^{-\eta L_{j,t-1}} e^{-\eta \ell_{j,t}}}{|e^{-\eta L_{j,t-1}}|_1} = \frac{|e^{-\eta L_t}|_1}{|e^{-\eta L_{t-1}}|_1}, \quad \text{for any } t = 1, \ldots, T. \]
Summing the Inequality (3.10) over $t$ then gives

$$
\sum_{t=1}^{T} \langle \theta_t, \ell_t \rangle \leq \frac{(b-a)^2 \eta T}{8} - \frac{1}{\eta} \sum_{t=1}^{T} \left( \log(|e^{-\eta L_{\theta(|t-1|)}}|) - \log(|e^{-\eta L_{\theta(t-1)}}|) \right)
= \frac{(b-a)^2 \eta T}{8} - \frac{1}{\eta} \left( \log(|e^{-\eta L_{\theta(|t-1|)}}|) - \log(|e^{-\eta L_{\theta(t-1)}}|) \right).
$$

To conclude, we notice that $|e^{-\eta L_{\theta(|t-1|)}}| = d$ and $|e^{-\eta L_{\theta(t-1)}}| \geq \max_{j=1,...,d} e^{-\eta L_{j,T}}$, so that

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

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

### 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 $\theta_t$ depending only on the past losses $\ell_1, \ldots, \ell_{t-1}$, we have the minoration

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

Yet, the proof of Theorem 3.2 starts with Jensen inequality (3.9), and, similarly as in Section 2.3.3 for PGD, there is a room for improvement for strongly convex losses. In this section, we exhibit this phenomenon for the square loss $\ell((\theta, h(t)), y(t)) = (y(t) - \langle \theta, h(t) \rangle)^2$.

We assume that

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

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 $j = 1, \ldots, d$ fulfilling (3.11), the exponential weight strategy (3.7) with $\eta = 1/(2B)$ fulfills

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

**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 $\eta = \sqrt{\frac{8 \log(d)}{B^2 T}}$ 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 $\eta$. Indeed, inspecting the proof below, we observe that for the choice $\eta = \sqrt{\frac{8 \log(d)}{B^2 T}}$ as in Theorem 3.2, we would only a get the bound $B \frac{\log(d)}{8T}$ 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 \( Z_t : \Omega \to [-\sqrt{B}, \sqrt{B}] \) by

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

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

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

ensures that

\[
|\theta_t e^{-\eta t} |_1 = \sum_{j=1}^d \theta_j e^{-\eta (y(t) - h_j(t))^2} \\
\leq \exp \left( -\eta \left( \sum_{j=1}^d \theta_j (y(t) - h_j(t))^2 \right) \right) = \exp \left( -\eta (y(t) - \langle \theta_t, h(t) \rangle)^2 \right). \tag{3.12}
\]

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 \( t = 1, \ldots, T \)

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

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. \( \square \)

### 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 \( g_t := \nabla f_t(\theta_t) = \partial_t \ell((\theta_t, h(t)), y(t))h(t) \). We can replace the linear discount \( \theta_t - \eta g_t \) in the PGD (3.2), by an exponential discount \( \theta_t e^{-\eta g_t} \). Hence, we consider the exponential weight strategy relative to the gradients

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

As before, we have the closed form formula

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

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

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

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

**Remarks.** Let us compare this result to the result (3.3) for PGD and to Theorem 3.2.
1. Similarly as for Theorem 3.2, we observe that compared to PGD, 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 of the best aggregation strategy \(\theta^*\), 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 strategy (3.13) of the expert advice outperform the predictions of the best of the experts.

**Proof of Theorem 3.4.**
As the function \(\theta \mapsto \ell ((\theta, h(t)), y(t))\) is convex and \(g_t = \nabla_\theta \ell ((\theta_t, h(t)), y(t))\), we have

\[
\sum_{t=1}^{T} \ell ((\theta_t, h(t)), y(t)) - \min_{\theta \in S_d} \sum_{t=1}^{T} \ell ((\theta, h(t)), y(t)) = \max_{\theta \in S_d} \sum_{t=1}^{T} (\ell ((\theta_t, h(t)), y(t)) - \ell ((\theta, h(t)), y(t)))
\]

\[
\leq \max_{\theta \in S_d} \sum_{t=1}^{T} \langle g_t, \theta_t - \theta \rangle
\]

\[
= \sum_{t=1}^{T} \langle g_t, \theta_t \rangle - \min_{j=1, \ldots, d} \sum_{t=1}^{T} g_{j,t},
\]

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

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

The proof of Theorem 3.4 is complete. \(\square\)

### 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 \(\theta, \theta_t \in \mathbb{R}^d\),

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

MIRROR DESCENT 27
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 \( \theta \rightarrow f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle \), then the minimum is achieved for some \( \theta \) 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 \( \theta_t \). A simple recipe is then to add a quadratic term \( \| \theta - \theta_t \|^2 \) which prevents the minimizer from being far away from \( \theta_t \). Hence, we can replace the minimisation problem \( \min_{\theta \in \mathbb{R}^d} f(\theta) \) by

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

(3.16)

The solution \( \theta_{t+1} \) 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 \( \phi \), 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 \( \phi : \mathbb{R}^d \rightarrow \mathbb{R} \) be convex and differentiable. The Bregman divergence associated to \( \phi \) is

\[
D_\phi(\theta, \omega) = \phi(\theta) - \phi(\omega) - \langle \nabla \phi(\omega), \theta - \omega \rangle,
\]  

(3.17)

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

As \( \phi \) 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

\[
\theta_{t+1} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ f(\theta_t) + \langle \nabla f(\theta_t), \theta - \theta_t \rangle + \frac{1}{\eta} D_\phi(\theta, \theta_t) \right\}.
\]  

(3.18)

As \( \nabla D_\phi(\theta, \theta_t) = \nabla \phi(\theta) - \nabla \phi(\theta_t) \), differentiating the objective function, we get that \( \theta_{t+1} \) 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 \( \phi \).

More generally, when we have a sequence \( f_t \) of objective functions, we may consider the sequential mirror descent.
Sequential Mirror Descent (SMD)

Input: \( \theta_1 \in \mathbb{R}^d, \eta > 0 \).

Iterate: For \( t = 1, \ldots, T - 1 \),
\[
\nabla \phi(\theta_{t+1}) = \nabla \phi(\theta_t) - \eta \nabla f_t(\theta_t)
\]

When the minimisation is constrained to occur in a compact convex set \( C \), we will constrain the update (3.18) to occur in
\[
\theta_{t+1} \in \arg\min_{\theta \in C} \left\{ f_t(\theta) + \langle \nabla f_t(\theta), \theta - \theta_t \rangle + \frac{1}{\eta} D_{\phi}(\theta, \theta_t) \right\}
\]
\[
= \arg\min_{\theta \in C} \left\{ \langle \nabla f_t(\theta), \theta \rangle + \frac{1}{\eta} (\phi(\theta) - \langle \nabla \phi(\theta), \theta \rangle) \right\}
\]
\[
= \arg\min_{\theta \in C} \{ \phi(\theta) - \langle \nabla \phi(\theta_t), \theta - \eta \nabla f_t(\theta_t) \rangle \}, \quad \text{with } \nabla \phi(\theta_{t+1}) := \nabla \phi(\theta_t) - \eta \nabla f_t(\theta_t).
\]

Let us denote by \( \pi^\phi_C \) the projection relative to the Bregman divergence
\[
\pi^\phi_C(\omega) \in \arg\min_{\theta \in C} D_{\phi}(\theta, \omega), \quad (3.19)
\]
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 \( t = 1, \ldots, T - 1 \),
\[
\bullet \ \nabla \phi(\omega_{t+1}) = \nabla \phi(\theta_t) - \eta \nabla f_t(\theta_t)
\]
\[
\bullet \ \theta_{t+1} = \pi^\phi_C(\omega_{t+1})
\]

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 PGD established in Section 2.3.

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 $\arg\min_{\theta \in C} D_\phi(\theta, \omega)$ 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

$$\text{for any } \omega \in \hat{D}, \text{ we have } \hat{D} \cap \arg\min_{\theta \in C} D_\phi(\theta, \omega) \neq \emptyset. \quad (3.20)$$

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

The next lemma generalizes the results of Section 2.3.1 on the projection operator $\pi_C$.

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

1. the projection $\pi_C^\phi(y)$ is uniquely defined and belongs to $C \cap \hat{D}$,
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 $\theta_1 \in C \cap \hat{D}$, then the subsequent updates $\theta_2, \theta_3, \ldots$ of the Projected Mirror Descent are also in $C \cap \hat{D}$. Hence, the PMD algorithm can be run indefinitely.

**Proof of Lemma 3.5.** Let $\pi y$ be any element in $\hat{D} \cap \arg\min_{\theta \in C} D_\phi(\theta, y)$, 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 $\pi y$, 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

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

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

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

The result follows.

1- It remains to prove the first claim. Let $z \in \arg\min_{\theta \in C} D_\phi(\theta, y)$. 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 $\phi$. So, the projection $\pi_C^\phi y$ is uniquely defined and belongs to $C \cap \hat{D}$. \hfill \Box

**Regret bound for PMD**

In this section, we generalize the analysis of Section 2.3.2.

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

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

We notice that the $\alpha$-strong convexity property $(\alpha\text{-Cvx})$ implies the strict convexity of $\phi$ on $C$.

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

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

We assume below that the functions $f_t$ are uniformly Lipschitz with respect to $| \cdot |$:
• Lipschitz condition: $|\nabla f_i(\theta)| \leq L$ for all $\theta \in \hat{D} \cap C$. (Lip)

**Theorem 3.6** Regret bound for PMD

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

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

**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 PGD, with the two following differences:

- $R$ controls the divergence $D_\phi$ between $\theta^*$ and $\theta_1$,
- $L$ controls the dual norm of the gradients of $f_i$.

Hence, a good choice of $\phi$ for the PMD, is a choice fulfilling

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

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

**Proof of Theorem 3.6.** In the analysis of PGD, the starting point was the polarisation formula

\[
2\langle a, b \rangle = ||a||^2 + ||b||^2 - ||a - b||^2,
\]

applied with $a = \eta \nabla f_i(\theta_t)$ and $b = \theta_t - \theta^*$. We follow the same argument, but with the polarisation formula (3.21)

\[
\eta (f_i(\theta_t) - f_i(\theta^*)) \overset{convex}{\leq} \langle \eta \nabla f_i(\theta_t), \theta_t - \theta^* \rangle = \langle \nabla \phi(\theta_t) - \nabla \phi(\omega_{t+1}), \theta_t - \theta^* \rangle \overset{polar}{=} D_\phi(\theta_t, \omega_{t+1}) + D_\phi(\theta^*, \theta_t) - D_\phi(\theta^*, \omega_{t+1}) \overset{Lem.3.5}{\leq} D_\phi(\theta_t, \omega_{t+1}) - D_\phi(\omega_{t+1}, \omega_{t+1}) + D_\phi(\theta^*, \theta_t) - D_\phi(\theta^*, \theta_{t+1}).
\]

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

\[
D_\phi(\theta_t, \omega_{t+1}) - D_\phi(\omega_{t+1}, \omega_{t+1}) = \phi(\theta_t) - \phi(\omega_{t+1}) = \langle \nabla \phi(\omega_{t+1}), \theta_t - \theta_{t+1} \rangle = \eta \nabla f_i(\theta_t) \overset{\alpha}{\leq} \eta ||\nabla f_i(\theta_t)||_\alpha \leq \frac{\eta^2}{2\alpha} ||f_i(\theta_t)||_\alpha^2,
\]

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

\[
\sum_{t=1}^{T} (f_i(\theta_t) - f_i(\theta^*)) \leq \frac{\eta^2}{2\alpha} \sum_{t=1}^{T} ||f_i(\theta_t)||_\alpha^2 + \frac{1}{\eta} (D_\phi(\theta^*, \theta_1) - D_\phi(\theta^*, \theta_{T+1})) \overset{(Lip)}{\leq} \frac{\eta}{2\alpha} L^2 T + \frac{R^2}{\eta}.
\]

Setting $\eta = \sqrt{\frac{2\alpha R^2}{L^2 T}}$, we get the result. \(\square\)
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 $\hat{\theta}$ 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 $f_t(\theta) = \ell((\theta, h(t)), y(t))$. We denote by $g_t = \partial_1 \ell((\theta, h(t)), y(t)) h(t)$, the gradient of $f_t(\theta)$ and the constraint set is $C = S_d$.

Let us choose a function $\phi$. 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 $\ell^\infty$. Hence, we want to choose a function $\phi$ which is strongly convex with respect to the $\ell^1$-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 $\phi : \mathcal{D} \rightarrow \mathbb{R}$ by

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

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

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

and the Bregman divergence is given by

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

For $\theta \in S_d$, and $\omega \in S_d \cap \overset{-}{\mathcal{D}}$, 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 $\theta \rightarrow ||\theta - \omega||^2$ (left) and the geometry $\theta \rightarrow D_\phi(\theta, \omega)$ (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 \( \phi \). We have for \( \theta > 0 \)
\[
\log((\omega_{t+1})^j) = \log((\theta_t)^j) - \eta g_{j,t}
\]
so \( \omega_{t+1} = \theta_t e^{-\eta g_t} \). Let us compute now the projection operator \( \pi_{S_d}^{\phi} \) related to \( \phi \).

**Lemma 3.7** Let \( \phi : \mathcal{D} \to \mathbb{R}_+ \) be defined by (3.25). Then for any \( \omega \in \hat{\mathcal{D}} \), we have
\[
\pi_{S_d}^{\phi} (\omega) = \frac{\omega}{|\omega|^1} \in \hat{\mathcal{D}}.
\]

As a consequence, we have the immediate corollary.

**Corollary 3.8** If we set \( \theta_1 = \frac{1}{d} \), the sequence produced by the projected mirror descent with \( \phi \) 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 \( C = S_d \)
\[
L(\theta, \lambda) = D_{\phi}(\theta, \omega) + \lambda \sum_{j=1}^d \theta_j - 1, \quad \text{for } \theta \in \mathcal{D}, \lambda \in \mathbb{R}.
\]
1. Compute the partial derivative with respect to \( \theta_j \) of \( L(\theta, \lambda) \).
2. Check that the solution \( \theta_\lambda \) to \( \nabla_{\theta} L(\theta_\lambda, \lambda) = 0 \) is \( \theta_\lambda = e^{\lambda \omega} \).
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 \( D_{\phi} \) is strongly convex with respect to the \( \ell^1 \) norm on the simplex \( S_d \).

**Lemma 3.9** Pinsker inequality

*For the Bregman divergence (3.26), we have*
\[
D_{\phi}(\theta, \omega) \geq \frac{1}{2} |\theta - \omega|^2_1, \quad \text{for any } \theta \in S_d, \text{ and } \omega \in S_d \cap \hat{\mathcal{D}}.
\]

**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
\[
\psi(t) \geq \frac{t^2}{2(1 + t/3)}, \quad \text{for all } t \geq -1.
\]
2. Check that
\[
D_{\phi}(\theta, \omega) = \sum_{j=1}^d \omega_j \psi(r_j) \geq \frac{1}{2} \sum_{j=1}^d \omega_j \frac{r_j^2}{1 + r_j/3}.
\]
3. Prove the inequalities
\[
|\theta - \omega|^2_1 = \left( \sum_{j=1}^d \omega_j |r_j| \right)^2 \leq \sum_{j=1}^d \omega_j \frac{r_j^2}{1 + r_j/3} \times \sum_{k=1}^d \omega_k (1 + r_k/3) \leq 2D_{\phi}(\theta, \omega).
\]
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 $P, Q$, with $P \ll Q$, we have

$$\frac{1}{2} \mathbb{E}_Q \left[ \left| \frac{dP}{dQ} - 1 \right|^2 \right] \leq \mathbb{E}_P \left[ \log \left( \frac{dP}{dQ} \right) \right].$$

![Figure 3.2](image)

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 $D_\phi$ is strongly convex with respect to the $\ell^1$ norm on the simplex $S_d$. Let us now bound the gradients in terms of the dual norm. The dual norm of the $\ell^1$ norm is the $\ell^\infty$ 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 $\theta_1 = 1/d$. 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

$$\frac{1}{T} \sum_{t=1}^T (f_t(\theta_1) - f_t(\theta^*)) \leq L \sqrt{\frac{2\log(d)}{T}},$$

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

### 3.5 Regret bounds for stochastic sequences

Let us now consider the case where the sequence $y(t)$ is stochastic. We denote $y(t)$ by $Y_t$ in order to emphasize the randomness. Let $(\mathcal{F}_t)_{t \geq 1}$ be a filtration such that $\sigma(Y_1, \ldots, Y_t) \subset \mathcal{F}_t$ and let us define $S_t = \mathbb{E}[Y_t | \mathcal{F}_{t-1}]$ which is the best possible prediction of $Y_t$ from $\mathcal{F}_{t-1}$ for the square loss. We assume that the predictions $h_j(t)$ are $\mathcal{F}_{t-1}$ measurable.
ILLUSTRATION

Under the assumption that all the $Y_t, h_j(t)$ take values in $[-\sqrt{B}/2, \sqrt{B}/2]$, applying the exponential weight strategy (3.7) with $\eta = 1/(2B)$, we get from Theorem 3.3

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (Y_t - \langle \theta_t, h(t) \rangle)^2 \right] - \min_{j=1,\ldots,d} \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (Y_t - h_j(t))^2 \right] \leq \frac{2B \log(d)}{T}.$$ 

We remind the reader that for any $\mathcal{F}_{t-1}$-measurable random variable $Z$, we have

$$\mathbb{E} \left[ (Y_t - Z)^2 \right] = \mathbb{E} \left[ (S_t - Z)^2 \right] + \mathbb{E} \left[ (Y_t - S_t)^2 \right].$$

So, we get

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (S_t - \langle \theta_t, h(t) \rangle)^2 \right] + \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (Y_t - S_t)^2 \right] = \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (Y_t - \langle \theta_t, h(t) \rangle)^2 \right]$$

$$\leq \min_{j=1,\ldots,d} \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (Y_t - h_j(t))^2 \right] + \frac{2B \log(d)}{T}$$

$$= \min_{j=1,\ldots,d} \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (S_t - h_j(t))^2 \right] + \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (Y_t - S_t)^2 \right] + \frac{2B \log(d)}{T}.$$ 

As the variance terms $\mathbb{E} \left[ (Y_t - S_t)^2 \right]$ cancel out, we get the control on the risk of the aggregated prediction

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (S_t - \langle \theta_t, h(t) \rangle)^2 \right] \leq \min_{j=1,\ldots,d} \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (S_t - h_j(t))^2 \right] + \frac{2B \log(d)}{T}.$$ 

Let us illustrate the above result in the regression setting where we observe a sequence $(X_t, Y_t)_{t \geq 1}$ such that

$$Y_t = f(X_t) + \epsilon_t, \quad t = 1, \ldots, T,$$

where $\epsilon_t$ is a random variable independent of $(X_s, Y_s)_{s \leq t-1}$ and $X_t$, with finite variance $\sigma_t^2$. Then, setting $\mathcal{F}_{t-1} = \sigma((X_s, Y_s)_{s \leq t-1}, \text{ X}_t)$, we have $S_t = f(X_t)$ and we get

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (f(X_t) - \langle \theta_t, h(t) \rangle)^2 \right] \leq \min_{j=1,\ldots,d} \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} (f(X_t) - h_j(t))^2 \right] + \frac{2B \log(d)}{T}.$$ 

This kind of results can be extended to the unbounded case, by proving some variants of Theorem 3.3.

3.6 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)

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

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

<table>
<thead>
<tr>
<th></th>
<th>M. Heat</th>
<th>Ms. Sun</th>
<th>Local expert</th>
<th>EW1</th>
<th>EW2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>26348</td>
<td>29013</td>
<td>35930</td>
<td>26577</td>
<td>23381</td>
</tr>
</tbody>
</table>

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.

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 \(t = 1, 2, \ldots\), 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 \(Y(A_1), \ldots, Y(A_{t-1})\). 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: \(X_1(1), X_1(2), \ldots\)
- ...
- sample of arm \(K\): \(X_K(1), X_K(2), \ldots\)

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

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

**External randomness.** The choice of the arm \(A_t\) may depend on some auxiliary sequence of random numbers \(U(1), U(2), \ldots \in [0, 1]\). 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 observation. In mathematical words, the choice \(A_t\) is \(\sigma(U_1, Y_1, \ldots, Y_{t-1}, U_t)\)-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 \(\psi_t : \mathbb{R}^{2t-1} \rightarrow \{1, \ldots, K\}\). At time \(t = 1, 2, \ldots\), the arm \(A_t\) is then pulled according to \(A_t = \psi_t(U_1, Y_1, \ldots, Y_{t-1}, U_t)\).

---

\(^1\)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 $\nu_k$, and mean $\mathbb{E}[X_k(n)] = \mu_k$ for $n \geq 1$.

Next lemma connects the expected cumulated reward to the means $\{\mu_k : k = 1, \ldots, T\}$ 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

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

**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

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

**Lemma 4.2 Wald formula.**

Let $(\mathcal{G}_n)_{n \geq 0}$ be a filtration and let $N, X(1), X(2), \ldots$ be random variables such that for all $n \geq 1$,
- $N$ takes value in $\{0, \ldots, T\}$ and $\{N \geq n\} \in \mathcal{G}_{n-1}$;
- $X(n)$ is independent of $\mathcal{G}_{n-1}$ and $\mathbb{E}[X(n)] = \mu$.

Then, we have

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

**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

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

The proof of Wald formula is complete. □

Let us set $\mathcal{G}_n = \sigma(X_k(1), \ldots, X_k(n), (U(j))_{j \geq 1}, (X_k(j))_{j \geq 1}, \epsilon_{sk})$. Then $\{N_k(r) \geq n\} \in \mathcal{G}_{n-1}$ 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 $\psi$ 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 \( \mu_{k^*} = \max_{j=1, \ldots, K} \mu_j \), the regret at time \( T \) is

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

(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

\[
A_t \in \arg \max_{k=1, \ldots, K} \bar{X}_k(N_k(t-1)).
\]

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 \( \mu_{k^*} \), possibly much smaller than the expected reward \( \mu_k \) 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 too much from below from the expected reward \( \mu_k \), 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 \( \mu_k \). 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 \( \mu_k \) 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 \(A_t \in \arg\max_{k=1, \ldots, K} U_k(t)\), 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

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

Then

\[
R(T) \leq \sum_{k \neq k^*} \Delta_k \left( TP \left( \Omega_{k,T}^c \right) + \delta_T^{-1}(\Delta_k/2) \right),
\]

(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 \(TP(\Omega_{k,T}^c)\) and \(\delta_T^{-1}(\Delta_k/2)\). So \(\delta_T(n)\) must be large enough, so that \(\delta_T(\Omega_{k,T}^c) = O(1/T)\), 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 \(\Omega_{k,T}^c\) 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 \(t' \in [t, T]\)

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

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 \(1 \lor \delta_T^{-1}(\Delta_k/2) = \delta_T^{-1}(\Delta_k/2)\). \(\square\)
Corollary 4.5 Assume that the distributions of the rewards of each arm $k$ are in $\text{subG}(\mu_k, \sigma^2)$. Setting

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

we get

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

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

$$P(\bar{X}_n \geq \mu_k + \delta_T(n)) \leq e^{-n\delta^2_T(n)/2\sigma^2} = e^{-\frac{1}{T^2}}.$$ 

Hence, with a union bound

$$P(\Omega^c_{k}, T) \leq 2 \sum_{n=1}^{T} \frac{1}{T^2} = \frac{2}{T}.$$ 

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

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

4.2.3 Horizon free UCB

In the above section, the algorithm has the undesirable feature to depend on the time horizon $T$ via $\delta_T$. 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 $\delta_T$ by $\delta_t$ 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 $A_t \in \text{argmax}_{k=1,...,K} U_k(t)$, 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

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

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

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

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

$$R(T) \leq \sum_{k=1}^{K} \Delta_k \sum_{\ell=0}^{L} E[N_k(T_{\ell+1}) - N_k(T_{\ell})].$$
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 $t' \in [t, T_{\ell+1}]$

$$X_k(N_k(t)) + \delta_t(N_k(t)) \leq X_k(N_k(t)) + \delta_{T_{\ell+1}}(N_k(t)) \leq \mu_k + 2\delta_{T_{\ell+1}}(N_k(t)) \leq \mu_k - \tilde{X}_k(N_k(t')) + \delta_{T_{\ell+1}}(N_k(t')) + \delta_t(N_k(t')).$$

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

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

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

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

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

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 $L = 1 + \log_2 \log_2(T_L) \leq 1 + \log_2 \log_2(T)$. \hfill $\Box$

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

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

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

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

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

2. The Kullback-Leibler divergence between two Gaussian distributions $N(\mu, \sigma^2)$ and $N(\mu', \sigma^2)$ is

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

**Exercise. Pinsker inequality.**

Adapt the proof of Lemma 3.9 to prove that for two probability distributions $P$, $Q$, with $P \ll Q$, we have

$$|P - Q|_1 := E_Q \left[ \left| \frac{dP}{dQ} - 1 \right| \right] \leq \sqrt{2KL(P, Q)}.$$

Check also the following additive property.

**Exercise. Tensorization of the Kullback-Leibler divergence.**

For four probability distributions $P_1, P_2, Q_1, Q_2$, with $P_1 \ll Q_1$ and $P_2 \ll Q_2$, we have

$$KL(P_1 \otimes P_2, Q_1 \otimes Q_2) = KL(P_1, Q_1) + KL(P_2, Q_2).$$

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

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

$$kl(E_P[Z], E_Q[Z]) \leq KL(P, Q).$$

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

$$kl(P(A), Q(A)) \leq KL(P, Q), \quad (4.6)$$

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

**Lemma 4.10** Let $A$ be any event such that $Q(A) > 0$. Then for any $P \ll Q$

$$\log \frac{P(A)}{Q(A)} \leq E_P \left[ \log \left( \frac{dP}{dQ} \right) \bigg| A \right].$$

**Proof of Lemma 4.10.** Let us write $L = log(dP/dQ)$. We have

$$Q(A) = E_P \left[ e^{-L} \mathbf{1}_A \right] = E_P \left[ E_P \left[ e^{-L} \big| \mathbf{1}_A \right] \mathbf{1}_A \right] \geq E_P \left[ e^{-E_P[L] \mathbf{1}_A} \mathbf{1}_A \right],$$

where the last inequality is ensured by Jensen inequality. We observe that

$$E_P[L \big| \mathbf{1}_A] = E_P[L | A] \mathbf{1}_A + E_P[L | A^c] \mathbf{1}_{A^c}, \quad (4.7)$$

so

$$Q(A) \geq E_P \left[ e^{-E_P[L] \mathbf{1}_A} \mathbf{1}_A \right] = E_P \left[ e^{-E_P[L | A] \mathbf{1}_A} \right] = e^{-E_P[L] P(A),}$$

which conclude the proof of Lemma 4.10.

We first observe that if $Q(A) = 0$ (resp. $Q(A^c) = 0$) then $P(A) = 0$ (resp. $P(A^c) = 0$), hence (4.6) trivially holds for $Q(A) \in \{0, 1\}$. Hence, we assume below that $Q(A) \in (0, 1)$.
Keeping the notation $L = \log(d\mathbb{P}/d\mathbb{Q})$, we have from (4.7) and Lemma 4.10
\[
KL(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_p[L] = \mathbb{E}_p[\mathbb{E}_p[L | I_A]] = \mathbb{E}_p[L | A] \mathbb{P}(A) + \mathbb{E}_p[L | A^c] \mathbb{P}(A^c)
\geq \mathbb{P}(A) \log \left( \frac{\mathbb{P}(A)}{\mathbb{Q}(A)} \right) + \mathbb{P}(A^c) \log \left( \frac{\mathbb{P}(A^c)}{\mathbb{Q}(A^c)} \right) = kl(\mathbb{P}(A), \mathbb{Q}(A)),
\]
which proves (4.6).

To prove Proposition 4.9, we apply (4.6) to the event $A = \{(w, u) : Z(w) > u \} \subset \Omega \times [0, 1]$, and the probability distributions $\mathbb{P} \otimes \lambda$ and $\mathbb{Q} \otimes \lambda$, where $\lambda$ is the Lebesgue measure on $[0, 1]$
\[
k(l(\mathbb{P} \otimes \lambda(A)), \mathbb{Q} \otimes \lambda(A)) \leq KL(\mathbb{P} \otimes \lambda, \mathbb{Q} \otimes \lambda).
\]
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
\[
\mathbb{P} \otimes \lambda(A) = \int_{w \in \Omega} d\mathbb{P}(w) \int_{u=0}^{1} du 1_{Z(w) > u} = \mathbb{E}_{\mathbb{P}}[Z].
\]
The proof of Proposition 4.9 is complete. \qed

**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 $(X_i(n))_{n=1}^{\ldots, K}, (U_i(n))_{n=1}^{\ldots, K}$. To a policy $\psi$, we associate the random variable $I_i = I_i(\psi) = (U_1, Y_1, \ldots, Y_{i-1}, U_i)$ defined as in the introduction. We denote by $\mathbb{P}_\nu^I$ the distribution of $I_i$ under $\mathbb{P}_\nu$.

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

**Lemma 4.11** For two sets of distributions $\nu = (\nu_1, \ldots, \nu_K)$ and $\nu' = (\nu'_1, \ldots, \nu'_K)$ for the rewards, we have
\[
KL(\mathbb{P}_\nu^I, \mathbb{P}_\nu'^I) = \sum_{k=1}^{K} KL(\nu_k, \nu'_k) \mathbb{E}_\nu[N_k(T)].
\]

**Proof of Lemma 4.11.**
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 $\lambda$ 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
\[
d\mathbb{P}_\nu^{(I_T, Y_T)}(i_T, y_T) = d\mathbb{P}_\nu^I(i_T) \ d\mathbb{P}_\nu^Y(y_T | I_T = i_T).
\]
Hence, we get
\[
\log \left( \frac{d\mathbb{P}_\nu^Y(y_T | I_T = i_T)}{d\mathbb{P}_\nu'^Y(y_T | I_T = i_T)} \right) = \log \left( \frac{d\mathbb{P}_\nu^I(i_T)}{d\mathbb{P}_\nu'^I(i_T)} \right) + \log \left( \frac{d\mathbb{P}_\nu^Y(y_T | I_T = i_T)}{d\mathbb{P}_\nu'^Y(y_T | I_T = i_T)} \right).
\]
Integrating the first term in the right-hand side of (4.8), we get
\[
\int_{i_T} \int_{y_T} \log \left( \frac{d\mathbb{P}_\nu^I(i_T)}{d\mathbb{P}_\nu'^I(i_T)} \right) d\mathbb{P}_\nu^{(I_T, Y_T)}(i_T, y_T) = \int_{i_T} \int_{y_T} \log \left( \frac{d\mathbb{P}_\nu^I(i_T)}{d\mathbb{P}_\nu'^I(i_T)} \right) d\mathbb{P}_\nu^I(i_T) d\mathbb{P}_\nu^Y(y_T | I_T = i_T)
\]
\[
= \int_{i_T} \log \left( \frac{d\mathbb{P}_\nu^I(i_T)}{d\mathbb{P}_\nu'^I(i_T)} \right) d\mathbb{P}_\nu^I(i_T) = KL(\mathbb{P}_\nu^I, \mathbb{P}_\nu'^I).
\]
As for the second term of (4.8), we observe that under $\mathbb{P}_r$ the conditional distribution of $Y_T$ given $I_T = i_T$ is $\nu_{\phi_T(i_T)}$. Hence $d\mathbb{P}_r^Y(Y_T | I_T = i_T) = d\nu_{\phi_T(i_T)}(Y_T)$. So

$$
\int_{i_T} \int_{Y_T} \log \left( \frac{d\mathbb{P}_r^Y(Y_T | I_T = i_T)}{d\mathbb{P}_r^Y(Y_T)} \right) d\mathbb{P}_r^Y(i_T, Y_T)
$$

$$
= \int_{i_T} \int_{Y_T} \log \left( \frac{d\nu_{\phi_T(i_T)}(Y_T)}{d\nu_{\phi_T(i_T)}} \right) d\mathbb{P}_r^Y(i_T) d\nu_{\phi_T(i_T)}(Y_T)
$$

$$
= \int_{i_T} d\mathbb{P}_r^Y(i_T) \sum_{k=1}^K \mathbf{1}_{\phi_T(i_T) = k} \int_{Y_T} \log \left( \frac{d\nu_k(Y_T)}{d\nu'_k(Y_T)} \right) d\nu_k(Y_T)
$$

$$
= \sum_{k=1}^K KL(\nu_k, \nu'_k) \mathbb{E}_r \left[ \mathbf{1}_{A_T = k} \right], \quad (4.10)
$$

where $A_T = \phi_T(I_T)$. Hence combining (4.9) and (4.10), we get by induction that

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

The proof of Lemma 4.11 is complete.

We have the following immediate corollary of Proposition 4.9 and Lemma 4.11.

**Corollary 4.12** 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 distribution for the rewards. Then

$$
\text{kl}(\mathbb{E}_r[Z], \mathbb{E}_{\nu'}[Z]) \leq \sum_{k=1}^K KL(\nu_k, \nu'_k) \mathbb{E}_r \left[ N_k(T) \right]. \quad (4.11)
$$

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

### 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. Theorem 4.13 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 $\mathcal{D} = \{N(\mu, 1) : \mu \in \mathbb{R}\}$ denote the set of Gaussian distributions with variance 1;
- Let $\Psi_{\alpha\text{-fast}}$ denote the set of policies $\psi$ such that, for any $\nu_1, \ldots, \nu_K \in \mathcal{D}$, and any $k \neq k^*$, we have $\mathbb{E}_r \left[ N_k(T) \right] = o(T^\alpha)$ as $T \to \infty$.

**Theorem 4.13** For any $\nu_1, \ldots, \nu_K \in \mathcal{D}$, with $\nu_k = N(\mu_k, 1)$, for $k = 1, \ldots, K$ and for any $\alpha \in (0, 1)$, we have as $T \to \infty$

$$
\inf_{\psi \in \Psi_{\alpha\text{-fast}}} R(\psi, T) \geq \sum_{k \neq k^*} \frac{2(1 - \alpha)}{\Delta_k} \log(T) + o(1).
$$
Let us comment briefly on this result. We observe that a policy that has a regret $o(T^\alpha)$ uniformly over $(\nu_1, \ldots, \nu_K) \in D^K$, cannot have a regret smaller than $O(\log(T))$ anywhere in $D^K$. In addition, the lower bound on the regret is proportional to

$$\sum_{k \neq k^*} \frac{\log(T)}{\Delta_k} + o(1),$$

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 $(\nu_1, \ldots, \nu_K) \in D^K$.

**Proof of Theorem 4.13.**

According to (4.2), we only need to prove that

$$\mathbb{E}_\nu [N_k(T)] \geq \frac{2(1-\alpha)}{\Delta_k^2} \log(T) + o(1), \quad \text{for all } k \neq k^*.$$  \hspace{1cm} (4.12)

Let us fix $k \neq k^*$. We define the collection $(\nu'_1, \ldots, \nu'_K) \in D^K$ as follows. For $\varepsilon > 0$, we set $\nu'_1 = N(\mu_{k'}, \varepsilon, 1)$ 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 $\nu'_k$. According to (4.11) with $Z = N_k(T)/T$, we have

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

We also have

$$kl(p, q) = (1-p) \log \left( \frac{1}{1-q} \right) + p \log \left( \frac{1}{q} \right) + p \log(p) + (1-p) \log(1-p) \geq 0 \quad \text{for all } 0 \leq q \leq 1,$$

so

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

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

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

It follows that

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

$$\geq (1-\alpha) \log(T) + o(1).$$

Direct computations give

$$KL(\nu_k, \nu'_k) = KL(N(\mu_k, 1), N(\mu_{k'}, \varepsilon, 1)) = \frac{1}{2} (\mu_{k'} + \varepsilon - \mu_k)^2,$$

so taking $\varepsilon = o(1/\log(T))$, we get (4.12). The proof of Theorem 4.13 is complete. \hspace{1cm} \square
4.4 Problem: X-arme 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 \( \mu \) is \((\beta, L)\)-Hölder,

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

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

As \( \mu \) 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

\[
m_k = \frac{1}{K} \int_{J_k} \mu(x) \, dx.
\]

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

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

where \( \Delta_k = \max_{j=1,\ldots,K} m_j - m_{k^*} = m_{k^*} - m_k \).

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 4.14** When \( \mu \) is \((\beta, L)\)-Hölder, for a suitable choice of \( K \) (depending on \( \beta \)), 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 \((L, \beta)\).

**Proof of Theorem 4.14.** Prove the theorem by solving the five next questions.

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

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

Check that the first sum can be simply upper bounded by \( DT \).

3. Check that the second sum is upper bounded by

\[
\sum_{k : \Delta_k > D} \Delta_k \mathbb{E} [N_k(T)] \leq \frac{4K \log(T)}{D} + 3KL.
\]
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 the number \( K \) of blocks, conclude the proof of Theorem 4.14.

### 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](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 \( t \to R(t) \) and the arms sampled at each time step.

![Figure 4.2: Top: Sampled arms when time passes. Arm 1 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 \( A_t \).](image)

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 \( k = 1, \ldots, 4 \), 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 \( \mu_k \), and that the 4 upper confidence bounds \( U_1(t), \ldots, U_4(t) \) 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 $k = 1, \ldots, 4$. Arm 1 in black, Arm 2 in red, Arm 3 in green, Arm 4 in blue. Left: $t = 100$. Center: $t = 300$. Right: $t = 1000$. 
Chapter 5

Lower bounds

In the previous chapter, we have derived a specific lower bound for the $K$-armed bandit problem. This bound was mainly based on the fact that $KL(P^X, Q^X) \leq KL(P, Q)$ for any random variable $X$. While this simple inequality is powerful, it is sometimes too crude to capture the right dependencies in more complex situations.

In this chapter, we describe a general strategy for getting lower bounds in statistics, based on a more refined inequality issued from information theory.

5.1 Minimax risk

Let us consider a set $(\mathcal{P}_f)_{f \in \mathcal{F}}$ of distributions on a measurable space $(X, \mathcal{A})$. Let $d$ be a distance on $\mathcal{F}$. We assume that we only have access to an observation $X \in X$ distributed as $\mathbb{P}_f$ and our goal is to recover $f$ from $X$. Hence, we want to design an estimator $\hat{f} : X \rightarrow \mathcal{F}$ such that $d(\hat{f}(X), f)$ is as small as possible. For example, we seek for $\hat{f}$ such that, for some $p > 0$ the expected error $\mathbb{E}_f [d(\hat{f}(X), f)^p]$ is as small as possible.

It turns out that seeking for $\hat{f}$ such that $\mathbb{E}_f [d(\hat{f}(X), f)^p]$ is as small as possible is a degenerate problem. Indeed, we have for $f \in \mathcal{F}$

\[
\min_{\hat{f} : X \text{ measurable} \rightarrow \mathcal{F}} \mathbb{E}_f [d(\hat{f}(X), f)^p] = 0,
\]

where the minimum is taken over all the measurable applications $\hat{f} : X \rightarrow \mathcal{F}$, with the minimum reached for the constant application $\hat{f}(x) = f$. Hence, we will not consider pointwise optimality (i.e. for a single $f$) but optimality on the class $\mathcal{F}$. A popular notion of risk is the minimax risk which corresponds to best possible error uniformly over the class $\mathcal{F}$

\[
\mathcal{R}^*(\mathcal{F}) := \min_{\hat{f} : X \text{ measurable} \rightarrow \mathcal{F}} \max_{f \in \mathcal{F}} \mathbb{E}_f [d(\hat{f}(X), f)^p],
\]

where, again, the minimum is taken over all the measurable applications $\hat{f} : X \rightarrow \mathcal{F}$.

Our goal in this chapter is to derive a lower bound on $\mathcal{R}^*(\mathcal{F})$. Such a lower bound is useful in statistics, as, if we find an estimator $\hat{f}$ with a max-risk over $\mathcal{F}$ similar to the lower bound

\[
\max_{f \in \mathcal{F}} \mathbb{E}_f [d(\hat{f}(X), f)^p] \approx \text{lower bound, where lower bound} \leq \mathcal{R}^*(\mathcal{F}),
\]

then it means that the estimator $\hat{f}$ performs almost as well as the best possible estimator in terms on the max-risk over $\mathcal{F}$.

\[^1\text{throughout this chapter, we use the notation } \mathbb{E}_f [\phi(X)] = \int_X \phi(x) d\mathbb{P}_f(x) \text{ and } \mathbb{E}_Q [\phi(X)] = \int_X \phi(x) dQ(x).\]
5.2 A recipe for proving lower bounds

In probability theory, it is often delicate to handle suprema over an infinite, possibly uncountable, space $F$. When the objective function, here $f \rightarrow \mathbb{E}_f \left[ d(\hat{f}(X), f)^p \right]$, is regular, a standard recipe is to replace the maximum over $F$ by a maximum over a finite set $\{f_1, \ldots, f_N\}$. Indeed, if any point $f \in F$ can be well approximated (in terms of the distance $d$) by one of the $f_1, \ldots, f_N$, then the maximum over $F$ and the maximum over $\{f_1, \ldots, f_N\}$ will be close.

Once we have discretized the problem, then it is possible to use lower bounds lifted from information theory, in order to get a lower bound on the minimax risk $R^*(F)$.

5.2.1 Fano’s lemma

The next result is a central tool in order to derive lower bounds in statistics.

**Theorem 5.1 Fano’s lemma.**

Let $(P_j)_{j=1,\ldots,N}$ be a set of probability distributions on $X$. For any probability distribution $Q$ such that $P_j \ll Q$, for $j = 1, \ldots, N$, we have

$$\min_{j: X \rightarrow \{1,\ldots,N\}} \max_{j=1,\ldots,N} P_j \left[ j(X) \neq j \right] \geq 1 - \frac{1 + \frac{1}{N} \sum_{j=1}^{N} KL(P_j, Q)}{\log(N)},$$

(5.2)

where $KL(P, Q)$ denotes the Kullback-Leibler divergence between $P$ and $Q$ defined in (4.5).

A classical choice for $Q$ is

$$Q = \frac{1}{N} \sum_{j=1}^{N} P_j,$$

but some other choices are sometimes more handy, depending on the problem.

**Proof of Fano’s lemma.**

First, we observe that

$$\min_{j: X \rightarrow \{1,\ldots,N\}} \max_{j=1,\ldots,N} P_j \left[ j(X) \neq j \right] = 1 - \max_{j: X \rightarrow \{1,\ldots,N\}} \min_{j=1,\ldots,N} P_j \left[ j(X) = j \right].$$

It is usually uneasy to handle minimum of probabilities, so we replace it with an average

$$\min_{j=1,\ldots,N} P_j \left[ j(X) = j \right] \leq \frac{1}{N} \sum_{j=1}^{N} P_j \left[ j(X) = j \right].$$

This bound is often harmless, as both quantities are typically of the same size for situations of interest. Next lemma provides an explicit formula for the best average error.

**Lemma 5.2 Best average risk.**

We have

$$\max_{j: X \rightarrow \{1,\ldots,N\}} \sum_{j=1}^{N} P_j \left[ j(X) = j \right] = \mathbb{E}_Q \left[ \max_{j=1,\ldots,N} \frac{dP_j}{dQ}(X) \right].$$
Proof of Lemma 5.2. We have
\[ \sum_{j=1}^{N} P_j[j(X) = j] = \int_{X} \sum_{j=1}^{N} 1_{j(X) = j} \frac{dP_j}{dQ}(x) \, dQ(x) \]
\[ \leq \int_{X} \sum_{j=1}^{N} 1_{j(X) = j} \max_{j'=1,...,N} \frac{dP_{j'}}{dQ}(x) \, dQ(x) \]
\[ = \mathbb{E}_Q \left[ \max_{j'=1,...,N} \frac{dP_{j'}}{dQ}(X) \right]. \]

In addition, the inequality above is an equality for
\[ j(x) \in \text{argmax}_{j=1,...,N} \frac{dP_j}{dQ}(x). \]

The proof of Lemma 5.2 is complete. □

In the proof above, it is worth to notice that the best \( j \) corresponds to the maximum likelihood estimator for \( j \).

So far, we have obtained that
\[ \min_{j:X \to \{1,...,N\}} \max_{j'=1,...,N} P_j[j(X) \neq j] \geq 1 - \frac{1}{N} \mathbb{E}_Q \left[ \max_{j=1,...,N} \frac{dP_j}{dQ}(X) \right]. \tag{5.3} \]

It remains to bound \( \mathbb{E}_Q \left[ \max_{j=1,...,N} \frac{dP_j}{dQ}(X) \right] \) from above in terms of the \( KL(P_j, Q) \).

A (too) naive bound. When we have positive random variables \( Z_1, \ldots, Z_N \), a simple way to bound the deviations or the expectation of \( \max_{j=1,...,N} Z_j \) is to use an union bound, which amounts to replace the maximum \( \max \) by the sum \( \sum \) in the expectation
\[ \mathbb{E} \left[ \max_j Z_j \right] \leq \sum_j \mathbb{E} [Z_j]. \]

If we apply this simple bound in our case, we obtain
\[ \frac{1}{N} \mathbb{E}_Q \left[ \max_{j=1,...,N} \frac{dP_j}{dQ}(X) \right] \leq \frac{1}{N} \sum_{j=1,...,N} \mathbb{E}_Q \left[ \frac{dP_j}{dQ}(X) \right] = 1. \]

So, at the end, we have upper-bounded \( \min_j P_j[j(X) = j] \) by one, which could have been done more directly!
Hence, we need to bound more carefully the expectation of the maximum.

A powerful variant. A simple but powerful variant of the max/sum bound is to combine it with Jensen inequality as in the proof of Lemma 1.6. We state the result as a lemma in order to highlight the generality of the bound.

**Lemma 5.3 Bounding a maximum.**

Let \( Z_1, \ldots, Z_N \) be \( N \) random variables with values in an interval \( I \subset \mathbb{R} \).
Then, for any increasing and convex function \( \varphi : I \to \mathbb{R}^+ \) we have
\[ \varphi \left( \mathbb{E} \left[ \max_{j=1,...,N} Z_j \right] \right) \leq \sum_{j=1}^{N} \mathbb{E} [\varphi(Z_j)]. \tag{5.4} \]
Proof of Lemma 5.3. We have

\[
\varphi \left( \mathbb{E} \left[ \max_{j=1, \ldots, N} Z_j \right] \right) \leq \mathbb{E} \left[ \varphi \left( \max_{j=1, \ldots, N} Z_j \right) \right] \quad \text{(Jensen inequality)}
\]

\[
\leq \mathbb{E} \left[ \max_{j=1, \ldots, N} \varphi \left( Z_j \right) \right] \quad \text{(\(\varphi\) increasing)}
\]

\[
\leq \sum_{j=1}^{N} \mathbb{E} \left[ \varphi \left( Z_j \right) \right] \quad \text{(\(\varphi\) non-negative)},
\]

which gives (5.4).

In order to bound \( \mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \) from above with Lemma 5.3, it remains to choose the function \( \varphi \). Setting \( \psi(u) = u \log(u) - u + 1 \), we have already seen that

\[
\mathbb{E}_Q \left[ \psi \left( \frac{dP_j}{dQ}(X) \right) \right] = \int_X \log \left( \frac{dP_j}{dQ}(x) \right) \frac{dP_j}{dQ}(x) \, dQ(x) - \int_X \frac{dP_j}{dQ}(x) \, dQ(x) + 1 = K L(P_j, Q).
\]

The function \( \psi \) is non-negative and convex, but it is decreasing on \([0, 1]\), as can be visualized on the Figure 5.1.

![Figure 5.1: Plot of the function \( \psi(u) = u \log(u) - u + 1 \).](image)

Hence, to get a non-decreasing function, we will truncate \( \psi \) on \([0, 1]\) and set \( \varphi(u) = \psi(u)1_{u \geq 1} \) which is convex, non-decreasing and non-negative. Applying Lemma 5.3 and \( \varphi(u) \leq \psi(u) \), we get

\[
\varphi \left( \mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \right) \leq \sum_{j=1}^{N} \mathbb{E}_Q \left[ \varphi \left( \frac{dP_j}{dQ}(X) \right) \right] \leq \sum_{j=1}^{N} \mathbb{E}_Q \left[ \psi \left( \frac{dP_j}{dQ}(X) \right) \right] = \sum_{j=1}^{N} KL(P_j, Q).
\]
A RECIPE FOR PROVING LOWER BOUNDS

In addition, since
\[
\mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \geq \mathbb{E}_Q \left[ \frac{dP_1}{dQ}(X) \right] = 1
\]
we have
\[
\psi \left( \mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \right) = \varphi \left( \mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \right) \leq \sum_{j=1}^{N} KL(p_j, Q).
\]
Let us set
\[
u = \frac{1}{N} \mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right].
\]
We have
\[
\psi(\nu) = N(\log(N) + \log(u)) - Nu + 1
\]
\[
= N\log(N) + N(u \log(u) - u + 1) - (N - 1)
\]
\[
\geq Nu \log(N) - N,
\]
so replacing \( u \) by its value (5.5), we get
\[
\mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \times \log(N) \leq N + \psi \left( \mathbb{E}_Q \left[ \max_{j=1, \ldots, N} \frac{dP_j}{dQ}(X) \right] \right) \leq N + \sum_{j=1}^{N} KL(p_j, Q). \tag{5.6}
\]
Combining (5.3) and (5.6), we get Fano’s inequality (5.2).

5.2.2 From Fano’s lemma to a lower bound over a finite set

Let \( \{f_1, \ldots, f_N\} \subset \mathcal{F} \) be any discretization of \( \mathcal{F} \). Let us now explain how we can get a lower bound on
\[
\mathcal{R}^*(\mathcal{F}, f_1, \ldots, f_N) := \min_{f : \mathcal{X} \to \mathcal{F}} \max_{j=1, \ldots, N} \mathbb{E}_{f_j} \left[ d(\hat{f}(X), f_j)^p \right]
\]
from Fano’s inequality (5.2).

Fano’s inequality provides a lower bound on
\[
\min_{f : \mathcal{X} \to \mathcal{F}} \max_{j=1, \ldots, N} \mathbb{P}_{f_j} \left[ \hat{f}(X) \neq f_j \right].
\]
So, to get a lower bound on \( \mathcal{R}^*(\mathcal{F}, f_1, \ldots, f_N) \) from Fano’s inequality, we must reduce two problems. First, we must reduce the minimum over \( \hat{f} : \mathcal{X} \to \mathcal{F} \) to the minimum over \( f : \mathcal{X} \to \{f_1, \ldots, f_N\} \), and second we must lower bound \( \mathbb{E}_{f_j} \left[ d(\hat{f}(X), f_j)^p \right] \) in terms of \( \mathbb{P}_{f_j} \left[ \hat{f}(X) \neq f_j \right] \). As explained in the next paragraph, these two reductions can be easily obtained by “projecting” \( \hat{f}(x) \) over the finite set \( \{f_1(x), \ldots, f_N(x)\} \) for all \( x \in \mathcal{X} \).

For any measurable \( \hat{f} : \mathcal{X} \to \mathcal{F} \), let us define
\[
\hat{f}(x) \in \text{argmin}_{j=1, \ldots, N} d(\hat{f}(x), f_j).
\]
By triangular inequality and the definition of \( \hat{f}(x) \), we have for any \( j = 1, \ldots, N \)
\[
\min_{i \neq k} d(f_i, f_k) \mathbf{1}_{j(x) \neq j} \leq d(f_j, \hat{f}(x)) \\
\leq d(f_j, \hat{f}(x)) + d(\hat{f}(x), f_{j(x)}) \\
\leq 2d(f_j, \hat{f}(x)).
\]
Combining this last bound with Fano’s inequality (5.2), we have proved the following result.

Corollary 5.4 Lower bound for discrete problem.
For any \{f_1, \ldots, f_N\} \subset \mathcal{F} and for any probability distribution \(Q\) such that \(P_{f_j} \ll Q\), for \(j = 1, \ldots, N\), we have
\[
\min_{f, X} \max_{j = 1, \ldots, N} \mathbb{E}_{f_j} \left[ d(\hat{f}(X), f_j)^p \right] \geq 2^{-p} \left( 1 - \frac{1}{N} \sum_{j=1}^N KL(P_{f_j}, Q) \right) \min_{i \neq k} d(f_i, f_k)^p, \tag{5.7}
\]
where \(KL(P, Q)\) denotes the Kullback-Leibler divergence between \(P\) and \(Q\).

5.2.3 Back to the original problem: finding a good discretization
Since, for any \(\{f_1, \ldots, f_N\}\),
\[
\mathcal{R}^*(\mathcal{F}) = \min_{f, X} \max_{f \in \mathcal{F}} \mathbb{E}_{f_j} \left[ d(\hat{f}(X), f)^p \right] \geq \min_{f, X} \max_{j = 1, \ldots, N} \mathbb{E}_{f_j} \left[ d(\hat{f}(X), f_j)^p \right],
\]
the Corollary 5.4 provides a lower bound on \(\mathcal{R}^*(\mathcal{F})\). This lower bound is general, and in practice all the art is to find a good discretization of \(\mathcal{F}\) so that the lower bound (5.7) is as large as possible.

We observe that we must find a discrete set \(\{f_1, \ldots, f_N\} \subset \mathcal{F}\) with
1. \(\min_{i \neq k} d(f_i, f_k)\) as large as possible
2. and \(\log(N)^{-1} \left( 1 + \frac{1}{N} \sum_{j=1}^N KL(P_{f_j}, Q) \right)\) bounded away from above from 1.

Both conditions are antagonistic as the first condition requires the \(f_j\) to be as much spread as possible (with respect to \(d\)), as the second condition ask for the \(f_j\) to be close enough to each other (with respect to the KL-divergence). Hence, the size of the minimax risk \(\mathcal{R}^*(\mathcal{F})\), depends on how much we can separate points in \(\mathcal{F}\) (with respect to \(d\)), within a region of fixed diameter (with respect to the KL-divergence).

The best is to illustrate this step on a (simple) example.

5.3 Illustration
As an illustration, let us consider the simple case where \(\mathcal{F} = \mathbb{R}^d\) and \(P_f\) is the Gaussian \(N(f, \sigma^2 I_d)\) distribution on \(\mathbb{R}^d\). In the previous chapter, we have computed \(KL(P_f, P_g) = \|f - g\|^2/(2\sigma^2)\) and we consider the distance \(d(f, f') = \|f - f'\|\).

For \(\varepsilon, R > 0\), a set \(N_{\varepsilon}\) is called an \(\varepsilon\)-packing of \(B(0, R)\) if \(N_{\varepsilon} \subset B(0, R)\) and if \(\|f - f'\|^2 > \varepsilon\) for any \(f, f' \in N_{\varepsilon}\).

Let us consider a maximal \(\varepsilon\)-packing \(N_{\varepsilon}\) of \(B(0, R)\). We observe that
\[
B(0, R) \subset \bigcup_{f \in N_{\varepsilon}} B(f, \varepsilon), \tag{5.8}
\]
as otherwise
- there would exists \(f' \in B(0, R)\) fulfilling \(\|f - f'\| > \varepsilon\) for all \(f \in N_{\varepsilon}\),
ILLUSTRATION

• $N'_e = N_e \cup \{f'\}$ would be an $\varepsilon'$-packing of $B(0, R)$, which contradicts the fact that $N_e$ is a maximal $\varepsilon$-packing of $B(0, R)$.

Comparing the volume on both sides of (5.8), we get

$$R^d \leq |N_e|\varepsilon^d. \quad (5.9)$$

Let us set $R^2 = d\sigma^2$ and $\varepsilon = e^{-3}R$. We have

$$1 + \frac{R^2}{2\sigma^2} \leq \frac{3d}{2} = \frac{d}{2} \log(R/\varepsilon) \leq \frac{1}{2} \log(N_e).$$

Hence, Corollary 5.4 with $Q = P_0$ gives

$$R^*(F) \geq \min_{\hat{f}, \text{measurable}} \max_{f \in N_e} \mathbb{E}_f \left[ \|\hat{f}(X) - f\|^2 \right]$$

$$\geq \frac{1}{4} \left( 1 - \frac{1 + \frac{1}{|N_e|} \sum_{f \in N_e} \frac{\|f - 0\|^2}{2\sigma^2}}{\log(|N_e|)} \right) \min_{f \neq f' \in N_e} \|f - f'\|^2$$

$$\geq \frac{1}{4} \left( 1 - \frac{1 + \frac{R^2}{2\sigma^2}}{\log(|N_e|)} \right) \varepsilon^2$$

$$\geq \frac{1}{8} \varepsilon^2 = \frac{1}{8e^6} d\sigma^2.$$

Hence, we have proved that the minimax risk for estimating the mean of a Gaussian distribution can be lower-bounded by a constant times $d\sigma^2$. As the risk of the estimator $\hat{f}(x) = x$ is equal to $d\sigma^2$ (check it!), we obtain that the minimax risk $R^*(F)$ is proportional to $d\sigma^2$. 
Part II

Matrix analysis for Machine Learning
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.

6.1 Reminder on spectral decomposition of symmetric real matrices

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

**Theorem 6.1 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 \( A = U \text{diag}(\lambda_1, \ldots, \lambda_n) U^T \) with \( U = [u_1 u_2 \cdots u_n] \).

![Figure 6.1: Geometric representation of the spectral theorem.](image)

**Proof of Theorem 6.1.** Let us give an analytic proof of the spectral decomposition. Let \( F : \mathbb{R}^n \rightarrow \mathbb{R} \) be defined by \( F(u) = \frac{1}{2} u^T A u \). 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 \( \partial B_{\mathbb{R}^n}(0, 1) \)

\[
u_1 \in \arg\max_{u_1 \in \partial B_{\mathbb{R}^n}(0, 1)} F(u).
\]
The tangent plane to $\partial B_{\mathbb{R}^n}(0, 1)$ in $u_1$ is $u_1 + V_1$. 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 $x = \sum_k (u^T_k x) u_k$, we have

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

The proof of Theorem 6.1 is complete. $\square$

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

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

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

### 6.2 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 6.2 Singular value decomposition

Any \( n \times p \) matrix \( A \) of rank \( r \) can be decomposed as

\[
A = \sum_{j=1}^{r} \sigma_j u_j v_j^T, \quad \text{where}
\]

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

\[
AA^T u_j = \sigma_j^2 u_j \quad \text{and} \quad A^T A v_j = \sigma_j^2 v_j.
\]

Figure 6.3: Geometric representation of the SVD. In blue, \( x = 0.75v_1 + 0.3v_2 \) 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 (6.1) 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 6.3 For any matrix \( A \in \mathbb{R}^{n \times p} \), we have the orthogonal decomposition

\[
\mathbb{R}^p = \text{ker}(A) \bigoplus \text{range}(A^T).
\]

Proof of Lemma 6.3. First, we observe that \( \text{ker}(A) \perp \text{range}(A^T) \). Indeed, for any \( y = A^T x \in \mathbb{R}^p \), \( \langle A^T x, \text{ker}(A) \rangle = 0 \).
range\( (A^T) \) and \( x_0 \in \ker(A) \), we have
\[
\langle x_0, y \rangle = \langle x_0, A^T x \rangle = \langle A x_0, x \rangle = 0.
\]
Since \( \dim(\text{range}(A^T)) = \text{rank}(A^T) = p - \dim(\ker(A)) \), and since \( \ker(A) \perp \text{range}(A^T) \), the conclusion follows. \( \Box \)

According to Lemma 6.3, the range of \( A \) and the range of \( AA^T \) coincide, so \( \text{rank}(AA^T) = \text{rank}(A) = r \). Since \( AA^T \) is positive semidefinite with rank \( r \), we have a spectral decomposition
\[
AA^T = \sum_{j=1}^{r} \sigma_j u_j u_j^T,
\]
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
\[
\langle v_i, v_j \rangle = \lambda_i^{-1/2} \lambda_j^{-1/2} A^T u_i A^T u_j = \delta_{i,j},
\]
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
\[
\sum_{j=1}^{r} \sigma_j u_j v_j^T = \sum_{j=1}^{r} \lambda_j^{1/2} \lambda_j^{-1/2} u_j u_j^T A
\]
\[
= \left( \sum_{j=1}^{r} u_j u_j^T \right) A.
\]
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 6.3, the range of \( A \) and the range of \( AA^T \) coincide so \( P \) is also the projection onto the range of \( A \). Hence
\[
\sum_{j=1}^{r} \sigma_j u_j v_j^T = \left( \sum_{j=1}^{r} u_j u_j^T \right) A = PA = \text{Proj}_{\text{range}(A)} A = A.
\]
The proof of Lemma 6.2 is complete. \( \Box \)

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 6.4  Min–Max / Max–Min formulas

For any $n \times p$ matrix $A$ and $k \leq r = \text{rank}(A)$, we have

$$\sigma_k(A) = \max_{S: \dim(S) = k} \min_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|}, \quad (6.2)$$

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

Symmetrically, we have

$$\sigma_k(A) = \min_{S: \dim(S) = k-1} \max_{x \in S \setminus \{0\}} \frac{\|Ax\|}{\|x\|}, \quad (6.3)$$

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 $\{v_{r+1}, \ldots, v_p\}$, such that $\{v_1, \ldots, v_p\}$ is an orthonormal basis of $\mathbb{R}^p$. We define $S_k = \text{span} \{v_1, \ldots, v_k\}$ and $W_k = \text{span} \{v_k, \ldots, v_p\}$. 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

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

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

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

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 (6.2). The min–max formula (6.3) is proved similarly. \(\square\)

6.3  Matrix analysis

6.3.1  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

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

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 6.5  For any matrices $A, B, C$ with compatible dimensions, we have

$$\langle AB, C \rangle_F = \langle A, CB^T \rangle_F = \langle B, A^T C \rangle_F \quad \text{and} \quad \langle A, I \rangle_F = \text{Tr}(A).$$
**Operator norm.** The $\ell^2 \to \ell^2$ operator norm is defined by

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

The last equality has been proved in the exercise page 66.

**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

$$\|A\|_{(p, q)} = \left( \sum_{k=1}^q \sigma_k(A)^p \right)^{1/p}.$$  \hspace{1cm} (6.4)

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

The three following inequalities are very useful.

**Lemma 6.6** We have

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

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

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

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

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

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

$$\|AB\|_F^2 = \sum_j \|(AB)j\|^2 = \sum_j \|AB_j\|^2 \leq \sum_j |A|_{\text{op}}^2 \|B_j\|^2 = |A|_{\text{op}}^2 \|B\|_F^2.$$  \hspace{1cm} (6.5)

The proof of Lemma 6.6 is complete.

**6.3.2 Low rank projection**

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

**Lemma 6.7** For an $n \times p$ matrix $A$ and $k \leq \min(n, p)$, we have for any $q \times n$ matrix $B$

$$\sigma_k(BA) \leq |B|_{\text{op}} \sigma_k(A).$$ \hspace{1cm} (6.6)

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

$$\sigma_k(AB) \leq |B|_{\text{op}} \sigma_k(A).$$ \hspace{1cm} (6.7)
Proof. From the definition of the operator norm, we have \( \|BAx\| \leq |B|_{op} \|Ax\| \). The inequality (6.5) then follows from (6.2). Furthermore, we have \( \sigma_k(AB) = \sigma_k(B^T A^T) \leq |B|_{op} \sigma_k(A^T) = |B|_{op} \sigma_k(A) \), which gives (6.6). \( \square \)

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

**Lemma 6.8** For any matrices \( A, B \in \mathbb{R}^{n \times p} \), we set \( q = \text{rank}(A) \wedge \text{rank}(B) \). 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 (6.4).

**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 \( q = \text{rank}(B) \), we have \( \|B\|_F = \|B\|_{(2,q)} \), and the lemma is proved. \( \square \)

The last result characterizes the “projection” on the set of matrices of rank \( r \).

**Theorem 6.9** For \( A = \sum_{k=1}^r \sigma_k(A) u_k v_k^T \) and \( q < r \), we have

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

where the minimum is achieved for

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

**Proof.** According to Lemma 6.8, 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 \( B = \sum_{k=1}^q \sigma_k(A) u_k v_k^T \). \( \square \)
6.4 Explicit computations with SVD decomposition

6.4.1 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 $D_{jj} > 0$ and $[D^+]_{jj} = 0$ otherwise.

**Symmetric matrices**

Write $A = UDU^T$ for a spectral decomposition of $A$ with $D$ diagonal and $U$ unitary. 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

$$A^+A = \sum_{j=1}^{r} v_jv_j^T = \text{Proj}_{\text{range}(A^r)} \quad \text{and} \quad AA^+ = \sum_{j=1}^{r} u_ju_j^T = \text{Proj}_{\text{range}(A^r)}.$$  

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 $A^+ = A^{-1}$.

6.4.2 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

$$\mathbb{E} \left[ \|Z\|^2 \right] = \|\mathbb{E} [Z] \|^2 + \text{Tr} (\text{Cov}(Z)),$$

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 $Y = X\beta + \varepsilon$, with $Y, \varepsilon \in \mathbb{R}^n$ et $\beta \in \mathbb{R}^p$. The matrix $X$ and the vector $\beta$ are non-random. We assume that $\mathbb{E}[\varepsilon] = 0$ and $\text{Cov}(\varepsilon) = \sigma^2I_n$.

We only observe the vector $Y \in \mathbb{R}^n$ and the matrix $X \in \mathbb{R}^{n\times p}$. Our goal is to estimate the vector $\beta$. In the following, the dimension $p$ can be larger than the dimension $n$.

For $\lambda > 0$, the ridge estimator $\hat{\beta}_\lambda$ is defined by

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} L(\beta) \quad \text{with} \quad L(\beta) = \|Y - X\beta\|^2 + \lambda\|\beta\|^2. \quad (6.7)$$

---

$^1$U unitary if $U^TU = UU^T = I$. 

### Explicit Computations with SVD Decomposition

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 \( \hat{\beta}_\lambda = A_\lambda Y \) with \( A_\lambda = (X^T X + \lambda I_p)^{-1} X^T \).
3. Let \( \sum_{k=1}^r \sigma_k u_k v_k^T \) be a singular value decomposition of \( X \). Prove that
   \[
   A_\lambda = \sum_{k=1}^r \frac{\sigma_k}{\sigma_k^2 + \lambda} v_k u_k^T \xrightarrow{\lambda \to 0^+} X^+,
   \]
   where \( A^+ \) is the Moore-Penrose pseudo-inverse of \( A \).
4. Check that we have
   \[
   X \hat{\beta}_\lambda = \sum_{k=1}^r \frac{\sigma_k^2}{\sigma_k^2 + \lambda} (u_k, Y) u_k. \tag{6.8}
   \]
5. Check that we have
   \[
   \mathbb{E} \left[ \hat{\beta}_\lambda \right] = \sum_{k=1}^r \frac{\sigma_k^2}{\sigma_k^2 + \lambda} (v_k, \beta) v_k.
   \]
6. Prove that the mean square error \( \mathbb{E} \left[ \| \hat{\beta}_\lambda - \beta \|^2 \right] \) of the Ridge estimator can be decomposed as
   \[
   \mathbb{E} \left[ \| \hat{\beta}_\lambda - \beta \|^2 \right] = \| \beta - \mathbb{E} \left[ \hat{\beta}_\lambda \right] \|^2 + \text{Tr}(\text{Cov}(\hat{\beta}_\lambda)).
   \]
   The first term is the norm of the bias of the Ridge estimator and the second term is the variance \( \mathbb{E} \left[ \| \hat{\beta}_\lambda - \mathbb{E} [\hat{\beta}_\lambda] \|^2 \right] \) 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 \( X^T \). Prove the following formula for the bias term
   \[
   \| \beta - \mathbb{E} \left[ \hat{\beta}_\lambda \right] \|^2 = \| \beta - P \beta \|^2 + \sum_{k=1}^r \left( \frac{\lambda}{\lambda + \sigma_k^2} \right)^2 (v_k, \beta)^2.
   \]
8. Check that the variance of the ridge estimator is given by
   \[
   \text{Tr}(\text{Cov}(\hat{\beta}_\lambda)) = \sigma^2 \text{Tr}(A_\lambda A_\lambda^T) = \sigma^2 \sum_{k=1}^r \left( \frac{\sigma_k}{\sigma_k^2 + \lambda} \right)^2.
   \]
9. How do the bias and the variance of \( \hat{\beta}_\lambda \) vary when \( \lambda \) increases?

**Remark.** We notice from (6.8) that \( X \hat{\beta}_\lambda \) 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 7

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.

7.1 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 7.1 Weyl inequality**

For two $n \times p$ matrices $A$ and $B$, we have for any $k \leq \min(n, p)$

$$|\sigma_k(A) - \sigma_k(B)| \leq \sigma_1(A - B) = |A - B|_{op}.$$  \hspace{1cm} (7.1)

**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 (7.1) follows by applying the Max–Min formula (6.2). \hspace{1cm} \Box

The Inequality (7.1) can be generalized as follows.

**Theorem 7.2 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 7.4.1 page 81 for a proof of these inequalities.

7.2 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 \( \{u_1, \ldots, u_r\} \) and \( \{v_1, \ldots, v_r\} \) are two orthonormal bases of \( \mathbb{R}^n \). We want to compare the eigenspaces span \( \{u_1, \ldots, u_r\} \) and span \( \{v_1, \ldots, v_r\} \), spanned by the \( r \) leading eigenvectors of \( A \) and \( B \).

A first idea could be to compare the two matrices \( U_r = [u_1, \ldots, u_r] \) and \( V_r = [v_1, \ldots, v_r] \). Yet, for any orthogonal transformation \( R \), we have span \( \{Ru_1, \ldots, Ru_r\} = \text{span} \{u_1, \ldots, u_r\} \), 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 \( \{u_1, \ldots, u_r\} \) and span \( \{v_1, \ldots, v_r\} \) 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 7.3** Let \( U_r = [u_{r+1}, \ldots, u_n] \) and \( V_r = [v_{r+1}, \ldots, v_n] \). Then, we have

\[
\|U_r U_r^T - V_r V_r^T\|^2_F = 2\|V_r^T U_r\|^2_F = 2\|U_r^T V_r\|^2_F.
\]

**Proof of Lemma 7.3.** We first expand the squares and use that the Frobenius norm of a projector is equal to its rank

\[
\|U_r U_r^T - V_r V_r^T\|^2_F = \|U_r U_r^T\|^2_F + \|V_r V_r^T\|^2_F - 2(U_r U_r^T, V_r V_r^T)_F
\]

\[
= 2r - 2\text{Tr}(U_r^T V_r U_r - V_r^T U_r).
\]

Then, since span \( \{v_{r+1}, \ldots, v_n\} \) is the orthogonal complement of span \( \{v_1, \ldots, v_r\} \), we have \( V_r V_r^T = I_n - V_r V_r^T \). So, as \( U_r^T U_r = I_r \)

\[
\|U_r U_r^T - V_r V_r^T\|^2_F = 2r - 2\text{Tr}(I_r - U_r^T V_r V_r^T U_r)
\]

\[
= 2\text{Tr}(U_r^T V_r V_r^T U_r) = 2\|V_r^T U_r\|^2_F.
\]

The second equality of Lemma 7.3 follows by symmetry. \( \square \)

Next result is the main theorem of this chapter. It provides a (classical) upper-bound on the norm \( \|U_r^T V_r\|^2_F \).

**Theorem 7.4** 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 \( U_r = [u_1, \ldots, u_r] \), \( U_{r-1} = [u_{r+1}, \ldots, u_n] \) and similarly \( V_r = [v_1, \ldots, v_r] \), \( V_{r-1} = [v_{r+1}, \ldots, v_n] \). Then, we have

\[
\|U_r^T V_r\|_F \leq \left( \sqrt{r} |A - B|_{\text{op}} \right) \wedge \|A - B\|_F
\]

\[
\leq 2 \frac{(\sqrt{r} |A - B|_{\text{op}}) \wedge \|A - B\|_F}{\lambda_r - \lambda_{r+1}}.
\]

In many cases, we only wish to compare the two leading eigenvectors of \( A \) and \( B \), which corresponds to the case \( r = 1 \).

**Corollary 7.5** Comparing leading eigenvectors.

\[
\sqrt{1 - \langle u_1, v_1 \rangle^2} \leq \frac{2 \inf_{\lambda \in \mathbb{R}} |A + \lambda I - B|_{\text{op}}}{\lambda_1 - \lambda_2}.
\]
Proof of Corollary 7.5.
We first observe that
\[
\|UU^T v_1\|^2 = v_1^T U_{-1} U^T v_1 = v_1^T (I - u_1 u_1^T) v_1 = 1 - (u_1^T v_1)^2.
\]
In addition, the eigenvectors of $A$ and $A + M$ are the same, while the eigenvalues are all translated by $\lambda$, preserving the eingap between the two largest eigenvalues. So, for any $\lambda \in \mathbb{R}$, the Inequality (7.3) applied to $A + M$ and $B$ gives
\[
\sqrt{1 - \langle u_1, v_1 \rangle^2} \leq \frac{2 |A + M - B|_{\text{op}}}{\lambda_1 - \lambda_2}.
\]
The proof Corollary 7.5 is complete. \hfill \Box

Proof of Theorem 7.4.
We first observe that the Bound (7.3) directly follows from (7.2) and the inequalities
\[
\lambda_r - \lambda_{r+1} = \lambda_r - \rho_{r+1} - (\rho_r - \rho_{r+1}) + \rho_r - \lambda_{r+1}
\leq (\lambda_r - \rho_{r+1}) + (\rho_r - \lambda_{r+1}) \leq 2((\rho_r - \lambda_{r+1}) \vee (\lambda_r - \rho_{r+1})).
\]
Let us prove (7.2). As a starting point, we notice that either $\rho_r > \lambda_{r+1}$ or $\lambda_r \geq \rho_{r+1}$, so
\[
(\rho_r - \lambda_{r+1}) \vee (\lambda_r - \rho_{r+1}) = (\rho_r - \lambda_{r+1})_+ \vee (\lambda_r - \rho_{r+1})_+.
\]
In addition, we observe from Lemma 7.3 that the role of $A$ and $B$ are symmetric. Hence, we only need to prove
\[
\|UU^T V_r\|_F \leq \frac{(\sqrt{r} |A - B|_{\text{op}}) \wedge \|A - B\|_F}{(\rho_r - \lambda_{r+1})_+}.
\]
When $\rho_r \leq \lambda_{r+1}$ 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
\[
\|UU^T V_r\|_F^2 = \sum_{k=1}^r \|UU^T v_k\|_F^2,
\]
so we will start by bounding the square norms $\|UU^T v_k\|_F^2$. Since
\[
A = \sum_{k=1}^n \lambda_k u_k u_k^T = U_r \text{diag}(\lambda_1, \ldots, \lambda_r) U_r^T + U_{-r} \text{diag}(\lambda_{r+1}, \ldots, \lambda_n) U_{-r}^T,
\]
we have $U_{-r}^T A = \text{diag}(\lambda_{r+1}, \ldots, \lambda_n) U_{-r}^T$. Hence, with $B v_k = \rho_k v_k$, we have for $k = 1, \ldots, r$
\[
\rho_k U_{-r}^T v_k = U_{-r}^T B v_k = U_{-r}^T (A + B - A) v_k
= \text{diag}(\lambda_{r+1}, \ldots, \lambda_n) U_{-r}^T v_k + U_{-r}^T (B - A) v_k.
\]
Hence
\[
U_{-r}^T v_k = \text{diag}(\rho_k - \lambda_{r+1}, \ldots, \rho_k - \lambda_n)^{-1} U_{-r}^T (B - A) v_k,
\]
and then, since $\rho_k \geq \rho_r > \lambda_{r+1}$,
\[
\|U_{-r}^T v_k\|^2 \leq | \text{diag}(\rho_k - \lambda_{r+1}, \ldots, \rho_k - \lambda_n)^{-1} U_{-r}^T v_k |^2 \| (B - A) v_k \|^2 \leq \frac{\|(B - A) v_k \|^2}{(\rho_k - \lambda_{r+1})^2} \leq \frac{\|(B - A) v_k \|^2}{(\rho_r - \lambda_{r+1})^2},
\]
where the last inequality follows from the Inequality (7.3) applied to $A + M$ and $B$.
for \( k = 1, \ldots, r \). To conclude, we observe that
\[
\sum_{k=1}^{r} \|(B - A)v_k\|_2^2 \leq |B - A|_{\text{op}}^2 \sum_{k=1}^{r} \|v_k\|_2^2 = r |B - A|_{\text{op}},
\]
since \( \|v_k\| = 1 \). So, with (7.6) we get
\[
\|U^T V_r\|_F^2 \leq \frac{r |A - B|_{\text{op}}^2}{(\rho_r - \lambda_{r+1})^2}.
\]
(7.7)

In addition, since \( I_n = V_r V_r^T + V_{r+1} V_{r+1}^T \), we have
\[
\|A - B\|_F^2 = \langle (A - B)(V_r V_r^T + V_{r+1} V_{r+1}^T), A - B \rangle_F = \|(A - B)V_r\|_F^2 + \|(A - B)V_{r+1}\|_F^2.
\]
so
\[
\sum_{k=1}^{r} \|(B - A)v_k\|_2^2 = \|(B - A)V_r\|_F^2 \leq \|(B - A)V_r\|_F^2 + \|(B - A)V_{r+1}\|_F^2 = \|B - A\|_F^2.
\]
Combining this bound with (7.6), we get
\[
\|U^T V_r\|_F^2 \leq \frac{\|A - B\|_F^2}{(\rho_r - \lambda_{r+1})^2}.
\]
(7.8)

Combining (7.7) and (7.8) we get (7.5), so the proof of Theorem 7.4 is complete. \( \square \)

### 7.3 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 \( N(0, \sigma^2) \) distribution. All the results of this section are valid for i.i.d. entries with \( \text{subG}(\sigma^2) \) distribution, at the price of larger numerical constants.

#### 7.3.1 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 7.6 Hanson-Wright inequality for Gaussian vectors

Symmetric forms. Let $\varepsilon$ 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 $L \geq 0$

$$\Pr \left[ \varepsilon^T S \varepsilon - \operatorname{Tr}(S) > \sqrt{8\|S\|_F^2 L} \vee (8 \|S\|_{\text{op}} L) \right] \leq e^{-L}. \quad (7.9)$$

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 $L \geq 0$

$$\Pr \left[ \varepsilon^T A \varepsilon' > \sqrt{4\|A\|_F^2 L} \vee (4 \|A\|_{\text{op}} L) \right] \leq e^{-L}. \quad (7.10)$$

Remarks:
1. The Inequality (7.9) is equivalent to the following statement: there exists an exponential random variable $\xi$ with parameter $1$ such that

$$\varepsilon^T S \varepsilon - \operatorname{Tr}(S) \leq \sqrt{8\|S\|_F^2 \xi} \vee (8 \|S\|_{\text{op}} \xi).$$

Using this compact formulation can be useful in complex proofs.
2. The Hanson-Wright inequalities remain valid (with worst constants) when $\varepsilon$ 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 $S = \text{diag}(a_1, \ldots, a_n)$.

Proof of Theorem 7.6.

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 7.7 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 7.7. Since $-\log(1 - x) \leq x + x^2$ for $|x| \leq 1/2$, we have

$$\mathbb{E} \left[ \exp(s(Z^2 - 1)) \right] = \frac{e^{-s}}{(1 - 2s)^{1/2}} \leq e^{2s^2}.$$

The proof of Lemma 7.7 is complete. $\square$

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 $\{v_1, \ldots, v_p\}$ form an orthonormal basis of $\mathbb{R}^p$, the matrix $V = [v_1, \ldots, v_p]$ 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 $k = 1, \ldots, p$ are i.i.d. $\mathcal{N}(0, 1)$-random variables.

Applying Markov inequality, we get for $t \geq 0$ and $|s| \leq (4 \|S\|_{\text{op}})^{-1}$

$$\Pr \left[ \varepsilon^T S \varepsilon - \operatorname{Tr}(S) > t \right] \leq e^{-st} \mathbb{E} \left[ e^{s(e^T S \varepsilon - \operatorname{Tr}(S))} \right]$$

$$\leq e^{-st} \prod_{k=1}^p \mathbb{E} \left[ \exp \left( s \lambda_k (Z_k^2 - 1) \right) \right]$$

$$\leq \exp \left( -st + 2s^2 \sum_{k=1}^p \lambda_k^2 \right) = \exp \left( -st + 2\|S\|_F^2 s^2 \right)$$
The minimum of \( s \to -st + 2\|S\|_F^2 s^2 \) over \( |s| \leq (4\|S\|_{op})^{-1} \) is achieved for \( s = \frac{1}{4}(t/\|S\|_F^2) \wedge (1/\|S\|_{op}) \) and hence

\[
\min_{|s| \leq (4\|S\|_{op})^{-1}} \left(-st + 2\|S\|_F^2 s^2\right) = \frac{-t^2}{8\|S\|_F^2} \mathbf{1}_{\{t \leq \|S\|_F^2/\|S\|_{op}\}} + \left(\frac{\|S\|_F^2}{8\|S\|_{op}} - \frac{t}{4\|S\|_{op}}\right) \mathbf{1}_{\{t > \|S\|_F^2/\|S\|_{op}\}}
\]

\[
\leq -\frac{1}{8} \left( \frac{t^2}{\|S\|_F^2} \wedge \frac{t}{\|S\|_{op}} \right).
\]

The Bound (7.9) follows.

**Cross products.** The trick for (7.10) is to notice that

\[
e^T A e' = \begin{bmatrix} e \\ e' \end{bmatrix}^T S \begin{bmatrix} e \\ e' \end{bmatrix}, \quad \text{with} \quad S = \frac{1}{2} \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}.
\]

Since \( S \) is symmetric we can apply (7.9). The conclusion follows by checking that \( \text{Tr}(S) = 0 \), \( |S|_{op} = |A|_{op}/2 \) and \( \|S\|_F^2 = \|A\|_F^2/2 \).

### 7.3.2 Concentration of random Gram matrices

In this section, we derive some bounds on the operator norm \( |E|_{op} = \sigma_1(E) \) of a random matrix with entries \( E_{ij} \) following an i.i.d. \( N(0, \sigma^2) \) distribution. As \( |E|_{op} = \sqrt{|EE^T|_{op}} \), we focus on the random Gram matrix \( EE^T \).

We first observe that

\[
\mathbb{E} \left[ E_{ik} E_{jk} \right] = \begin{cases} \mathbb{E}[E_{ik}] \mathbb{E}[E_{jk}] = 0 & \text{for } i \neq j \\ \mathbb{E} \left[ E_{ik}^2 \right] = \sigma^2 & \text{for } i = j. 
\end{cases}
\]

Hence, we have \( \mathbb{E}[EE^T] = p \sigma^2 I_n \). Instead of simply upper-bounding \( |EE^T|_{op} \), 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 7.8 Concentration of \( EE^T \).**

Assume that \( E \in \mathbb{R}^{n \times p} \) has i.i.d. entries following a \( N(0, \sigma^2) \) distribution. Then, there exists a random variable \( \xi \) with exponential distribution with parameter 1 such that

\[
|EE^T - p \sigma^2 I_n|_{op} \leq 4\sigma^2 \sqrt{p(6n + 2\xi) + (48n + 16\xi)\sigma^2}.
\]

(7.11)

We can derive from this theorem the following control on \( |E|_{op} \).

**Corollary 7.9** Under the hypotheses of Theorem 7.8, there exists a random variable \( \xi \) with exponential distribution with parameter 1 such that

\[
|E|_{op} \leq \sigma \left( \sqrt{p} + 7\sqrt{n + \xi} \right).
\]
OPERATOR NORM OF RANDOM MATRICES

Proof of Corollary 7.9.
By the triangular inequality, we have
\[ \sigma_1(E)^2 = \|EE^T\|_{op} \leq \|p\sigma^2 I_n\|_{op} + \|EE^T - p\sigma^2 I_n\|_{op} \]
\[ \leq \sigma^2 \left(p + 4\sqrt{p(6n + 2\xi)} + 48n + 16\xi\right) \]
\[ \leq \sigma^2 \left(\sqrt{p} + 7\sqrt{n + \xi}\right)^2. \]
The Corollary follows. □

Proof of Theorem 7.8.
Sketch of the proof. Before starting the proof of Theorem 7.8, let us sketch the main lines.
First of all, dividing both sides of (7.11) by \(\sigma^2\), we can assume with no loss of generality that \(\sigma^2 = 1\).
As \(EE^T - pI_n\) is a symmetric matrix, we have
\[ \|EE^T - pI_n\|_{op} = \sup_{x \in \partial B_{\mathbb{R}^n}(0, 1)} |\langle(EE^T - pI_n)x, x\rangle|. \]
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 (7.9).
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
\[ \mathbb{P}\left[\max_{i=1,\ldots,n} Z_i > t\right] \leq \sum_{i=1}^n \mathbb{P}[Z_i > t]. \]
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
\[ \|A\|_{op} = \sup_{x \in \partial B_{\mathbb{R}^n}(0, 1)} |\langle Ax, x\rangle|. \]
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 \(\varepsilon\)-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 \(|x - y| \leq \varepsilon\). Next lemma links the operator norm of a matrix to a supremum over an \(\varepsilon\)-net.

Lemma 7.10 For any symmetric matrix \(A \in \mathbb{R}^{n \times n}\) and any \(\varepsilon\)-net of \(\partial B_{\mathbb{R}^n}(0, 1)\), we have
\[ \|A\|_{op} \leq \frac{1}{1 - 2\varepsilon} \sup_{x \in \mathcal{N}_\varepsilon} |\langle Ax, x\rangle|. \quad (7.12) \]
Proof of Lemma 7.10.
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 \( \|x^* - y\| \leq \varepsilon \).
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
\[
|A|_{\text{op}} \leq |\langle Ax^*, x^* \rangle| \leq |\langle Ay, y \rangle| + |\langle A(x^* - y), y \rangle| + |\langle Ax^*, x^* - y \rangle| \\
\leq \sup_{y \in \mathcal{N}_\varepsilon} \|\langle Ay, y \rangle\| + 2 |A|_{\text{op}} \varepsilon.
\]
The Bound (7.12) then follows.

Next lemma provides an upper bound on the cardinality of a minimal \( \varepsilon \)-net of \( \partial B_{\mathbb{R}^n}(0, 1) \).

Lemma 7.11
For any \( n \in \mathbb{N} \) and \( \varepsilon > 0 \), there exists an \( \varepsilon \)-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 7.4.2 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
\[
|EE^T - pI_n|_{\text{op}} \leq 2 \max_{x \in \mathcal{N}_{1/4}} |\langle (EE^T - pI_n)x, x \rangle| = 2 \max_{x \in \mathcal{N}_{1/4}} \|E^T x\|^2 - p. \tag{7.13}
\]

Step 2: concentration of the quadratic forms. Let \( E_i \) denotes the \( i \)th column of \( E \). We observe that \( x^T E_i \sim \mathcal{N}(0, x^T x) \) and that the \( (x^T E_i)_{i=1,\ldots,p} \) 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 \( E_i x \) are i.i.d. \( \mathcal{N}(0, 1) \), and the random vector \( \xi_x = E^T x \) follows a standard Gaussian distribution \( \mathcal{N}(0, I_p) \) in \( \mathbb{R}^p \). Hanson-Wright inequality (7.9) with \( S = I_p \) and \( S = -I_p \) ensures that there exist two exponential random variables \( \xi_x, \xi'_x \), such that
\[
\|E^T x\|^2 - p \leq \sqrt{8p} \xi_x \vee 8\xi_x \quad \text{and} \quad p - \|E^T x\|^2 \leq \sqrt{8p} \xi'_x \vee 8\xi'_x.
\]
Therefore, combining with (7.13), we obtain the concentration bound
\[
|EE^T - pI_n|_{\text{op}} \leq 2 \left( \sqrt{8p} \max_{x \in \mathcal{N}_{1/4}} (\xi_x \vee \xi'_x) + 8 \max_{x \in \mathcal{N}_{1/4}} (\xi_x \vee \xi'_x) \right). \tag{7.14}
\]

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 \( \xi \) 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 (7.14), we obtain (7.11). The proof of Theorem 7.8 is complete. \( \square \)
7.4 Exercices

7.4.1 Generalized Weyl inequalities

In this exercise, you will prove the Theorem 7.2.

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

$$\max_{x \in S_i \setminus \{0\}} \frac{\|Ax\|}{\|x\|} \leq \sigma_i(A)$$

and

$$\max_{x \in S_j \setminus \{0\}} \frac{\|(B - A)x\|}{\|x\|} \leq \sigma_j(B - A).$$

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 (6.2).

7.4.2 Cardinality of an $\varepsilon$-net

In this exercise, we prove the Lemma 7.11. Let us define $N_\varepsilon$ as follows. Start from any $x_1 \in \partial B_{\mathbb{R}^n}(0, 1)$, and for $k = 2, 3, \ldots$, choose recursively any $x_k \in \partial B_{\mathbb{R}^n}(x_j, \varepsilon)$ such that $x_k \not\in \bigcup_{j=1}^{k-1} B_{\mathbb{R}^n}(x_j, \varepsilon)$. When no such $x_k$ remains, stop and define $N_\varepsilon = \{x_1, x_2, \ldots\}$.

1. Why is the cardinality of $N_\varepsilon$ finite?

2. Observe that $N_\varepsilon$ is an $\varepsilon$-net of $\partial B_{\mathbb{R}^n}(0, 1)$ and that $\|x - y\| > \varepsilon$ for any $x, y \in N_\varepsilon$, with $x \neq y$.

3. Check that

   i) the balls $\{B_{\mathbb{R}^n}(x, \varepsilon/2) : x \in N_\varepsilon\}$ are disjoint;

   ii) $\bigcup_{x \in N_\varepsilon} B_{\mathbb{R}^n}(x, \varepsilon/2) \subset B_{\mathbb{R}^n}(0, 1 + \varepsilon/2)$.

4. Conclude by comparing the volume of the balls $B_{\mathbb{R}^n}(x, \varepsilon/2)$ and $B_{\mathbb{R}^n}(0, 1 + \varepsilon/2)$.

7.4.3 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

$$\frac{1}{n} \sum_{k=1}^{n} \delta_{\rho^{-1/2} \sigma_k(E)}(x)$$

of the singular values of the matrix $\rho^{-1/2} E \in \mathbb{R}^{n \times p}$ converges almost surely to the Marchenko-Pastur distribution $[\beta]$, which has a density on $[1 - \sqrt{\beta}, 1 + \sqrt{\beta}]$ given by

$$f_\beta(x) = \frac{1}{\pi \beta x} \sqrt{(x^2 - (1 - \sqrt{\beta})^2)(1 + \sqrt{\beta})^2 - x^2}. \quad (7.15)$$

This classical result of random matrix theory is illustrated in Figure 7.1.
Figure 7.1: Plot of the histogram of the singular values of $p^{-1/2}E$ and of the Marchenko-Pastur distribution (7.15) 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 8

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 distortion. 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 methods in data analysis. As we will see, the Principal Component Analysis (PCA) is tightly linked to the Singular Value Decomposition (SVD) introduced in Chapter 6.

8.1 Principal Component Analysis

8.1.1 Finding the best low dimensional linear representation of data

The principle of Principal Component Analysis (PCA) is to seek for a linear span \(\hat{V}_d\) 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 \(\hat{V}_d\).

For data points \(X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p\) and any dimension \(d \leq p\), the PCA computes the linear span \(\hat{V}_d\) in \(\mathbb{R}^p\) minimizing

\[
\hat{V}_d \in \text{argmin}_{V \text{ dim}(V) \leq d} \sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|_2, \tag{8.1}
\]

where the minimum is over all the subspaces \(V \subset \mathbb{R}^p\) with dimension at most \(d\) and \(\text{Proj}_V\) is the orthogonal projection matrix onto the linear span \(V\).

Let us stack the data \(X^{(1)}, \ldots, X^{(n)}\) into a \(n \times p\) matrix

\[
X = \begin{pmatrix}
(X^{(1)})^T \\
\vdots \\
(X^{(n)})^T
\end{pmatrix} \in \mathbb{R}^{n \times p},
\]

and let us denote by \(X = \sum_{k=1}^{r} \sigma_k u_k v_k^T\) a SVD of \(X\) with the usual convention \(\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0\).
Theorem 8.1 PCA algorithm.

The solution to (8.1) is the linear span \( \hat{V}_d = \text{span} \{ v_1, \ldots, v_d \} \).

In addition, the coordinates of \( \text{Proj}_{\hat{V}_d} X^{(i)} \) in the orthonormal basis \( (v_1, \ldots, v_d) \) of \( \hat{V}_d \) 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 8.1: The data point \( X^{(i)} \) projected on \( \hat{V}_2 \), 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 \( X \).
2. The projection \( \text{Proj}_{\hat{V}_d} X^{(i)} \in \mathbb{R}^p \) lies in \( \hat{V}_d \), 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 \( \hat{V}_d \) is a linear span and not an affine span, it is highly recommended to first center the data points

\[
\overline{X}^{(i)} = X^{(i)} - \frac{1}{n} \sum_{j=1}^{n} X^{(j)}
\]

and then proceed with a PCA on the \( \overline{X}^{(1)}, \ldots, \overline{X}^{(n)} \).

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 \( X v_k = \sigma_k u_k \), the principal component \( c_k \) is obtained as the image of \( v_k \) by the matrix \( X \). We emphasize that the principal axes are orthonormal and the principal components are orthogonal.

Proof of Theorem 8.1. To start with, we observe that

\[
\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2 = \|X - X \text{Proj}_V\|_F^2.
\]

For any linear span \( V \) of dimension \( d \), the rank of the matrix \( X \text{Proj}_V \) is not larger than \( d \), so according to Theorem 6.9 in Chapter 6, page 69,

\[
\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_V X^{(i)}\|^2 = \|X - X \text{Proj}_V\|_F^2 \geq \min_{\text{rank}(B) \leq d} \|X - B\|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2. \quad (8.2)
\]
PRINCIPAL COMPONENT ANALYSIS

Furthermore, for $\tilde{V}_d = \text{span} \{v_1, \ldots, v_d\}$, we have

$$
X \text{Proj}_{\tilde{V}_d} = \sum_{k=1}^r \sigma_k u_k v_k^T \sum_{j=1}^d v_j v_j^T = \sum_{k=1}^d \sigma_k u_k v_k^T.
$$

So

$$
\|X - X \text{Proj}_{\tilde{V}_d}\|_F^2 = \sum_{k=d+1}^r \sigma_k u_k v_k^T \sum_{k=d+1}^r \sigma_k u_k v_k^T.
$$

Comparing (8.2) and (8.3), we find that $\tilde{V}_d = \text{span} \{v_1, \ldots, v_d\}$ is solution to (8.1).

In addition, the coordinate of $\text{Proj}_{\tilde{V}_d} X^{(i)}$ over $v_k$

$$
\langle X^{(i)}, v_k \rangle = \langle X^T e_i, v_k \rangle = \langle e_i, X v_k \rangle = \sigma_k \langle e_i, u_k \rangle,
$$

where we used for the last equality that $X v_k = \sigma_k u_k$. Hence, the coordinates of $\text{Proj}_{\tilde{V}_d} X^{(i)}$ in the orthonormal basis $(v_1, \ldots, v_d)$ of $\tilde{V}_d$ are given by $(c_1(i), \ldots, c_d(i))$, where $c_k := \sigma_k u_k$.

8.1.2 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 [7], which gathers 1100 scans of each digit. Each scan is a $16 \times 256$ image, which can be encoded as a vector in $\mathbb{R}^{256}$. The Figure 8.2 illustrates the compressed images, when they are projected on the linear span $\tilde{V}_{10}$ 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

$$
\bar{X}^{(i)} = X^{(i)} - \frac{1}{n} \sum_{j=1}^n X^{(j)}.
$$

Then, we proceed with a PCA on the matrix

$$
\bar{X} = \begin{pmatrix}
(\bar{X}^{(1)})^T \\
\vdots \\
(\bar{X}^{(n)})^T
\end{pmatrix} = \sum_{k=1}^r \sigma_k u_k v_k^T,
$$

by computing $\tilde{V}_{10} = \text{span} \{\bar{v}_1, \ldots, \bar{v}_{10}\}$. The vectors $\text{Proj}_{\tilde{V}_{10}}(\bar{X}^{(1)}), \ldots, \text{Proj}_{\tilde{V}_{10}}(\bar{X}^{(n)}) \in \mathbb{R}^{256}$ give the best approximation of $\bar{X}^{(1)}, \ldots, \bar{X}^{(n)}$ by a projection on a linear span of dimension 10. To obtain the final compressed images, we de-center the images

$$
\text{compressed}(X^{(i)}) = \text{Proj}_{\tilde{V}_{10}}(\bar{X}^{(i)}) + \frac{1}{n} \sum_{j=1}^n X^{(j)}, \quad i = 1, \ldots, n.
$$

These compressed images, together with the original images are plotted in Figure 8.2. 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).
8.2 Interpreting PCA

Formulas for projection

Let us denote by $Z$ the projected data: $Z^{(i)} = \text{Proj}_{\hat{U}_d} X^{(i)}$, 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 $X$ by its transpose

$$X^T = \sum_k \sigma_k v_k u_k^T,$$

and apply PCA to $X^T$. Theorem 8.1 ensures that the best possible approximation space is $\hat{U}_d = \text{span} \{u_1, \ldots, u_d\}$ and

$$\text{Proj}_{\hat{U}_d} X_a = \sum_{k=1}^d \sigma_k v_k(a) u_k = \sum_{k=1}^d v_k(a) c_k.$$
INTERPRETING PCA

Hence, \( \nu(a) \) represents the coordinates of \( \text{Proj}_{\hat{U}_d}X_a \) in the orthogonal (but not orthonormal!) basis \( \{c_1, \ldots, c_d\} \) of \( \hat{U}_d \).

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 8.1, we get the following measure of the quality of approximation

\[
\sum_{i=1}^{n} \|X^{(i)} - \text{Proj}_{\hat{V}_d}X^{(i)}\|^2 = \|X - X \text{Proj}_{\hat{V}_d}\|^2 = \sum_{k=d+1}^{r} \sigma_k^2.
\]

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

\[
\frac{\|X - X \text{Proj}_{\hat{V}_d}\|^2}{\|X\|^2} = \frac{\sum_{k=d+1}^{r} \sigma_k^2}{\sum_{k=1}^{r} \sigma_k^2} = 1 - \frac{\sum_{k=1}^{d} \sigma_k^2}{\sum_{k=1}^{r} \sigma_k^2}.
\]

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 8.3, 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 \( \hat{V}_2 \). More precisely, for \( i = 1, \ldots, n \), we plot the vector \( c(i) \) of coordinates of \( Z^{(i)} = \text{Proj}_{\hat{V}_2}X^{(i)} \) on the orthonormal basis \( \{v_1, v_2\} \) of \( \hat{V}_2 \). 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 \( \hat{V}_2 \). 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 \( \hat{U}_2 \). It is interesting to note that

\[
\langle \text{Proj}_{\hat{U}_2}X_a, \text{Proj}_{\hat{U}_2}X_b \rangle = \sum_{k=1}^{2} \sigma_k^2 \nu_k(a)\nu_k(b),
\]

so we can plot the vectors \( \{\sigma_k \nu_k(a)\}_{k=1,2} \) and \( \{\sigma_k \nu_k(b)\}_{k=1,2} \) in order to visualize the correlations between the variables \( a \) and \( b \). Again, it is good to check if the ratio \( \|\text{Proj}_{\hat{U}_2}X_a\|/\|X_a\| \) 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 \( \nu \), where we plot simultaneously the \( c(i) \) and \( \nu(a) \). In this case \( c(i) \) represents the projection of the data point \( i \) and \( \nu(a) \) represents the projection of the variable \( a \). We emphasize that there is a distortion in the representation of the variable \( a \), as \( \nu(a) \) corresponds to the coordinates of \( \text{Proj}_{\hat{U}_2}X_a \) in the orthogonal, but not orthonormal basis \( \{c_1, c_2\} \) of \( \hat{U}_2 \). We can observe on a biplot the correlation between individuals and variables. As the projection \( Z^{(i)} = \text{Proj}_{\hat{V}_2}X^{(i)} \) on \( \hat{V}_2 \) is given by \( Z_a^{(i)} = \langle c(i), \nu(a) \rangle \), we can visualize on the biplot the size of the entry \( Z_a^{(i)} \) by looking at the scalar product \( \langle c(i), \nu(a) \rangle \). If \( c(i) \) and \( \nu(a) \) are well aligned, then the entry \( Z_a^{(i)} \) will be large, while if \( c(i) \) and \( \nu(a) \) 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 8.4.2.

The results of the heptathlon event at the 1988 Seoul Olympic Games are displayed in the table below.

<table>
<thead>
<tr>
<th></th>
<th>hurdles</th>
<th>highjump</th>
<th>shot</th>
<th>run200m</th>
<th>longjump</th>
<th>javelin</th>
<th>run800m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joyner-Kersee (USA)</td>
<td>12.69</td>
<td>1.86</td>
<td>15.80</td>
<td>22.56</td>
<td>7.27</td>
<td>45.66</td>
<td>128.51</td>
</tr>
<tr>
<td>John (GDR)</td>
<td>12.85</td>
<td>1.80</td>
<td>16.23</td>
<td>23.65</td>
<td>6.71</td>
<td>42.56</td>
<td>126.12</td>
</tr>
<tr>
<td>Behmer (GDR)</td>
<td>13.20</td>
<td>1.83</td>
<td>14.20</td>
<td>23.10</td>
<td>6.68</td>
<td>44.54</td>
<td>124.20</td>
</tr>
<tr>
<td>Sablovskaitaite (URS)</td>
<td>13.61</td>
<td>1.80</td>
<td>15.23</td>
<td>23.92</td>
<td>6.25</td>
<td>42.78</td>
<td>132.24</td>
</tr>
<tr>
<td>Choubenkova (URS)</td>
<td>13.51</td>
<td>1.74</td>
<td>14.76</td>
<td>23.93</td>
<td>6.32</td>
<td>47.46</td>
<td>127.90</td>
</tr>
<tr>
<td>Schulz (GDR)</td>
<td>13.75</td>
<td>1.83</td>
<td>13.50</td>
<td>24.65</td>
<td>6.33</td>
<td>42.82</td>
<td>125.79</td>
</tr>
<tr>
<td>Fleming (AUS)</td>
<td>13.38</td>
<td>1.80</td>
<td>12.88</td>
<td>23.59</td>
<td>6.37</td>
<td>40.28</td>
<td>132.54</td>
</tr>
<tr>
<td>Greiner (USA)</td>
<td>13.55</td>
<td>1.80</td>
<td>14.13</td>
<td>24.48</td>
<td>6.47</td>
<td>38.00</td>
<td>133.65</td>
</tr>
<tr>
<td>Lapinberova (CZE)</td>
<td>13.63</td>
<td>1.83</td>
<td>14.28</td>
<td>24.86</td>
<td>6.11</td>
<td>42.20</td>
<td>136.05</td>
</tr>
<tr>
<td>Bouraga (URS)</td>
<td>13.25</td>
<td>1.77</td>
<td>12.62</td>
<td>23.59</td>
<td>6.28</td>
<td>39.06</td>
<td>134.74</td>
</tr>
<tr>
<td>Wijnsma (HOL)</td>
<td>13.75</td>
<td>1.86</td>
<td>13.01</td>
<td>25.03</td>
<td>6.34</td>
<td>37.86</td>
<td>131.49</td>
</tr>
<tr>
<td>Dimitrova (BUL)</td>
<td>13.24</td>
<td>1.80</td>
<td>12.88</td>
<td>23.59</td>
<td>6.37</td>
<td>40.28</td>
<td>132.54</td>
</tr>
<tr>
<td>Scheider (SWI)</td>
<td>13.85</td>
<td>1.86</td>
<td>11.58</td>
<td>24.87</td>
<td>6.05</td>
<td>47.50</td>
<td>134.93</td>
</tr>
<tr>
<td>Braun (FRG)</td>
<td>13.71</td>
<td>1.83</td>
<td>13.16</td>
<td>24.78</td>
<td>6.12</td>
<td>44.58</td>
<td>142.82</td>
</tr>
<tr>
<td>Ruotsalainen (FIN)</td>
<td>13.79</td>
<td>1.80</td>
<td>12.32</td>
<td>24.61</td>
<td>6.08</td>
<td>45.44</td>
<td>137.06</td>
</tr>
<tr>
<td>Yuping (CHN)</td>
<td>13.93</td>
<td>1.86</td>
<td>12.41</td>
<td>25.00</td>
<td>6.40</td>
<td>38.60</td>
<td>146.67</td>
</tr>
<tr>
<td>Hagger (GB)</td>
<td>13.47</td>
<td>1.80</td>
<td>12.75</td>
<td>25.47</td>
<td>6.34</td>
<td>35.76</td>
<td>138.48</td>
</tr>
<tr>
<td>Brown (USA)</td>
<td>14.07</td>
<td>1.83</td>
<td>12.69</td>
<td>24.83</td>
<td>6.13</td>
<td>44.34</td>
<td>146.43</td>
</tr>
<tr>
<td>Mulliner (GB)</td>
<td>14.39</td>
<td>1.71</td>
<td>12.68</td>
<td>24.92</td>
<td>6.10</td>
<td>37.76</td>
<td>138.02</td>
</tr>
<tr>
<td>Hautenauve (BEL)</td>
<td>14.04</td>
<td>1.77</td>
<td>11.81</td>
<td>25.61</td>
<td>5.99</td>
<td>35.68</td>
<td>133.90</td>
</tr>
<tr>
<td>Kytola (FIN)</td>
<td>14.31</td>
<td>1.77</td>
<td>11.66</td>
<td>25.69</td>
<td>5.75</td>
<td>39.48</td>
<td>133.35</td>
</tr>
<tr>
<td>Geremias (BRA)</td>
<td>14.23</td>
<td>1.71</td>
<td>12.95</td>
<td>25.50</td>
<td>5.50</td>
<td>39.64</td>
<td>144.02</td>
</tr>
<tr>
<td>Hui-Ing (TAI)</td>
<td>14.85</td>
<td>1.68</td>
<td>10.00</td>
<td>25.23</td>
<td>5.47</td>
<td>39.14</td>
<td>137.30</td>
</tr>
<tr>
<td>Jeong-Mi (KOR)</td>
<td>14.53</td>
<td>1.71</td>
<td>10.83</td>
<td>26.61</td>
<td>5.50</td>
<td>39.26</td>
<td>139.17</td>
</tr>
<tr>
<td>Launa (PNG)</td>
<td>16.42</td>
<td>1.50</td>
<td>11.78</td>
<td>26.16</td>
<td>4.88</td>
<td>46.38</td>
<td>163.43</td>
</tr>
</tbody>
</table>

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 $d = 2$.

![Plot of square singular values](image1)

Figure 8.3: 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 $\text{Proj}_{\hat{U}_2} X_a/\|X_a\|$. These vectors are close to the unit circle, so each discipline is well represented by the projection on $\hat{U}_2$.

![Plot of vectors](image2)

Figure 8.4: Plot of the vectors $\text{Proj}_{\hat{U}_2} X_a/\|X_a\|$ relative to the unit circle.

Finally, we draw the biplot of the athletes (in black) and the disciplines (in red).
8.3 Theory for PCA

8.3.1 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 $X = B + E$, 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 $X$ and $B$

$$X = \sum_{k=1}^{r} \hat{\sigma}_k \hat{u}_k \hat{v}_k^T, \quad B = \sum_{k=1}^{r} \sigma_k \hat{u}_k \hat{v}_k^T.$$

We have set some "hats" on the SVD of $X$ in order to emphasize that these quantities can be computed from the data $X$, while the matrix $B$ and its SVD are not observed. We estimate $B$ by the
THEORY FOR PCA

projection of the data \( X \) on the \( d \) first principal axes

\[
\hat{B}_d = X \text{Proj}_{\hat{V}_d} = \sum_{k=1}^{d} \hat{\sigma}_k \hat{u}_k \hat{v}_k^T.
\]  

(8.5)

Next theorem provides a bound on the estimation error \( \| \hat{B}_d - B \|_F^2 \) in terms of the operator norm \( |E|_{\text{op}} \) of the fluctuations and in terms of the best possible approximation error of \( B \) by a matrix of rank \( d \) (see Theorem 6.9, page 69)

\[
\min_{M: \text{rank}(M) \leq d} \| B - M \|_F^2 = \sum_{k=d+1}^{r} \sigma_k^2.
\]

Theorem 8.2 Recovering low rank component.
The estimator \( \hat{B}_d \) fulfills the error bound

\[
\| \hat{B}_d - B \|_F^2 \leq 9 \left( \| E \|_{\text{op}}^2 + \sum_{k=d+1}^{r} \sigma_k^2 \right).
\]  

(8.6)

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 \( N(0, \sigma^2) \) distribution. Then, we have the next result which directly follows from Theorem 8.2 and Corollary 7.9, page 78.

Corollary 8.3 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 \( N(0, \sigma^2) \) distribution, we have with probability at least \( 1 - e^{-L} \)

\[
\| \hat{B}_d - B \|_F^2 \leq 9d \left( \sqrt{p} + 7 \sqrt{n + L} \right)^2 \sigma^2 + 9 \sum_{k=d+1}^{r} \sigma_k^2.
\]

We observe that in the setting of Corollary 8.3, we have

\[
\mathbb{E} \left[ \| X - B \|_F^2 \right] = \mathbb{E} \left[ \| E \|_F^2 \right] = np\sigma^2,
\]

which is much larger than \( d(p + n)\sigma^2 \) if \( d \ll n \land p \). So, when \( B \) is approximately of rank \( d \) with \( d \ll n \land p \), there is a substantial gain in using \( \hat{B}_d \) instead of \( X \) in order to estimate \( B \).

Proof of Theorem 8.2. Before starting the proof, we remind the reader two useful inequalities.

Lemma 8.4 For any \( a > 0 \), and \( x, y \in \mathbb{R}^n \), we have

\[
2 \langle x, y \rangle \leq a \| x \|^2 + a^{-1} \| y \|^2,
\]

(8.7)

\[
\| x + y \|^2 \leq (1 + a) \| x \|^2 + (1 + a^{-1}) \| y \|^2.
\]

(8.8)

The Inequality (8.8) immediately follows from (8.7) and the Inequality (8.7) 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 (8.6). 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 6.9, page 69). As $\hat{B}_d$ is the best approximation of $X$ by a matrix of rank $d$, we have

$$\|X - \hat{B}_d\|^2 \leq \|X - B_d\|^2.$$ 

Using the decomposition $X = B + E$ and expanding the squares, the previous inequality is equivalent to

$$\|B - \hat{B}_d\|^2 \leq \|B - B_d\|^2 + 2(E, \hat{B}_d - B_d).$$

(8.9)

We observe that rank($\hat{B}_d - B_d$) $\leq 2d$, so according to Lemma 6.8, we can upper-bound the scalar product $(E, \hat{B}_d - B_d)$ in terms of the $(2, 2d)$-Ky-Fan norm

$$2(E, \hat{B}_d - B_d) \leq 2\|E\|_{(2, 2d)}\|\hat{B}_d - B_d\|_{(2, 2d)} = 2\|E\|_{(2, 2d)}\|\hat{B}_d - B_d\|_F.$$ 

Applying Inequality (8.7) with $a = 5/2$, then Inequality (8.8) with $a = 1/9$, and finally $\|E\|_{(2, 2d)} \leq 2d \|E\|_{\text{op}}$, we get

$$2(E, \hat{B}_d - B_d) \leq \frac{5}{2}\|E\|_{(2, 2d)}^2 + \frac{2}{5}\|\hat{B}_d - B_d\|_F^2,$$

$$\leq \frac{5}{2}\|E\|_{(2, 2d)}^2 + \frac{2}{5}\left((10/9)\|\hat{B}_d - B\|_F^2 + 10\|B - B_d\|_F^2\right),$$

$$\leq 5d \|E\|_{\text{op}}^2 + 4\|\hat{B}_d - B\|_F^2 + 4\|B - B_d\|_F^2.$$ 

Combining this last inequality with (8.9) and

$$\|B - B_d\|_F^2 = \left\| \sum_{k=d+1}^r \sigma_k u_k v_k^T \right\|_F^2 = \sum_{k=d+1}^r \sigma_k^2,$$

we get (8.6).

**8.3.2 Dimension selection**

The Bound (8.6) can be used in order to propose a choice for $d$ with theoretical garanties. Indeed, let us notice that the Bound (8.6) can be written as

$$\|\hat{B}_d - B\|_F^2 \leq 9 \left( \sum_{k=1}^d (\|E\|_{\text{op}}^2 - \sigma_k^2) + \sum_{k=1}^r \sigma_k^2 \right).$$

(8.10)

As the second term is independent of $d$, the integer $d$ minimizing the right-hand side of (8.10) 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 $\sigma_j^2 \geq |E|_{\text{op}}^2$ and then it increases. Hence in order to minimize the right-hand side of (8.6), the best is to choose the dimension

$$d^* := \max \{ k : \sigma_k \geq |E|_{\text{op}} \}.$$ 

(8.11)

While the operator norm $|E|_{\text{op}}$ can be evaluated in some cases, for example with Corollary 7.9 page 78, the singular values $\sigma_1 \geq \sigma_2 \geq \ldots$ of $B$ are not observed, so we cannot directly use (8.11). Yet, combining Weyl Inequality (7.1), page 73, with the Bound (8.10), we get

$$\|\hat{B}_d - B\|_F^2 \leq 9 \left( \sum_{k=1}^d (|E|_{\text{op}}^2 - (\hat{\sigma}_k - |E|_{\text{op}}^2) + \sum_{k=1}^r \sigma_k^2 \right),$$

with $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \ldots$ the singular values of $X$. Following the same reasoning as before, with $\sigma_k$ replaced by $(\hat{\sigma}_k - |E|_{\text{op}})_+$, we get the next selection rule for $d$. 

Corollary 8.5 Dimension selection.

For \( \hat{d} = \max \{ k : \hat{\sigma}_k \geq 2 |E|_{\text{op}} \} \), we have

\[
\| \hat{B}_{\hat{d}} - B \|_F^2 \leq 9^2 \min_{d \geq 0} \left( d |E|_{\text{op}}^2 + \sum_{k=d+1}^{r} \sigma_k^2 \right).
\]

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 \( \tilde{d} = \max \{ k : \hat{\sigma}_k \geq 2 \left( \sqrt{p} + 7 \sqrt{2n} \right) \sigma \} \) and then get with probability at least \( 1 - e^{-n} \)

\[
\| \hat{B}_{\tilde{d}} - B \|_F^2 \leq 9^2 \min_{d \geq 0} \left( d \left( \sqrt{p} + 7 \sqrt{2n} \right)^2 \sigma^2 + \sum_{k=d+1}^{r} \sigma_k^2 \right).
\]

Proof of Corollary 8.5. According to Weyl Inequality (7.1) and the definition of \( \hat{d} \), we have

\[
\sigma_{\hat{d}} \geq \hat{\sigma}_{\hat{d}} - |E|_{\text{op}} \geq |E|_{\text{op}},
\]

so \( \tilde{d} \leq d^* \), with \( d^* \) defined by (8.11). In addition, Weyl Inequality (7.1) and the definition of \( \hat{d} \) also ensure that for any \( k \in [\hat{d} + 1, d^*] \), we have

\[
\sigma_k \leq \tilde{\sigma}_k + |E|_{\text{op}} \leq 3 |E|_{\text{op}},
\]

where the second inequality comes from \( k \geq \hat{d} + 1 \) and the definition of \( \hat{d} \). So, the Bound (8.6) gives

\[
\| \hat{B}_{\tilde{d}} - B \|_F^2 \leq 9 \left( \tilde{d} |E|_{\text{op}}^2 + \sum_{k=\tilde{d}+1}^{d^*} \sigma_k^2 + \sum_{k=d^*+1}^{r} \sigma_k^2 \right)
\leq 9 \left( \tilde{d} |E|_{\text{op}}^2 + 9(d^* - \hat{d}) |E|_{\text{op}}^2 + \sum_{k=d^*+1}^{r} \sigma_k^2 \right)
\leq 9^2 \left( d^* |E|_{\text{op}}^2 + \sum_{k=d^*+1}^{r} \sigma_k^2 \right) = 9^2 \min_{d \geq 0} \left( d |E|_{\text{op}}^2 + \sum_{k=d+1}^{r} \sigma_k^2 \right),
\]

where the last equality follows from the definition (8.11) of \( d^* \). \( \square \)

8.4 Exercises

8.4.1 Rank recovery

Let us consider the setting of Section 8.3.1 and let us assume that the rank of \( B \) is \( d \). As in Corollary 8.5, let us set for some \( \lambda > 0 \)

\[
\hat{d} = \max \{ k : \hat{\sigma}_k \geq \lambda \}.
\]

(8.12)

In this exercise, we will give some conditions ensuring that \( \hat{d} = d \) with large probability.

1. Check that

\[
P \left[ \hat{d} \neq d \right] = P \left[ \hat{\sigma}_{d+1} \geq \lambda \text{ or } \hat{\sigma}_d < \lambda \right].
\]

2. With Weyl Inequality (7.1), page 73, prove that

\[
P \left[ \hat{d} \neq d \right] \leq P \left[ |E|_{\text{op}} \geq \lambda \wedge (\sigma_d - \lambda) \right].
\]
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 7.9, page 78, that the integer $\hat{d}$ defined by (8.12) fulfills

$$\Pr \left[ \hat{d} = d \right] \geq 1 - e^{-n}.$$ 

**Remark.** We have recovered the rank under a condition ensuring that the singular value $\sigma_d$ 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 $\sigma_d$ is small, much smaller than the fluctuations of $\hat{\sigma}_d - \sigma_d$, then no procedure can detect if $\sigma_d$ is or not 0.

### 8.4.2 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](https://cran.r-project.org). Then, you can enter the following code to download and analyze the data.

```R
# 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 9

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 in 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.

9.1 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 \( \Sigma_i = \text{Cov}(X_i) \) 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 \( G^* = \{G^*_1, \ldots, G^*_K\} \) 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 \( G^* = \{G^*_1, \ldots, G^*_K\} \) 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

\[
\text{for all } i \in G^*_k : \quad \mathbb{E}[X_i] = \theta_k \text{ and } \text{Cov}(X_i) = \Lambda_k. \quad (9.1)
\]

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 \( X \in \mathbb{R}^{n \times p} \) the matrix whose \( i \)-th row is given by \( X_i \). We define similarly the matrices

- \( E \in \mathbb{R}^{n \times p} \) 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 \( \theta_k \);
- and \( A \in \mathbb{R}^{n \times K} \) the membership matrix defined by \( A_{ik} = 1_{i \in G^*_k} \), for \( i = 1, \ldots, n \), and \( k = 1, \ldots, K \).

Then we have the compact formula

\[
X = A\Theta + E. \quad (9.2)
\]

**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 \( \pi \) on \( \{1, \ldots, K\} \).
9.2 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.

![Dendrogram](image)

**Figure 9.1:** Left: data points in \( \mathbb{R}^2 \). Right: dendrogram of hierarchical clustering with Euclidean distance \( d \) and complete linkage \( \ell \). 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 \( d(x, y) = |x - y| \). Some classical examples of linkage are:

- **Single linkage**: single linkage corresponds to the smallest distance between the points of the two clusters
  \[
  \ell_{\text{single}}(G, G') = \min \left\{ d(x_i, x_j) : i \in G, \ j \in G' \right\}.
  \]
  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
  \[
  \ell_{\text{complete}}(G, G') = \max \left\{ d(x_i, x_j) : i \in G, \ j \in G' \right\}.
  \]
  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 $G, G'$

$$\ell_{\text{average}}(G, G') = \frac{1}{|G||G'|} \sum_{i \in G, j \in G'} d(x_i, x_j).$$

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 9.4.4, 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 $\ell$
- **Initialization:** $G^{(n)} = \{\{1\}, \ldots, \{n\}\}$
- **Iterations:** for $t = n, \ldots, 2$
  - find $(\hat{a}, \hat{b}) \in \arg\min_{(a,b)} \ell(G^{(t)}_a, G^{(t)}_b)$
  - build $G^{(t-1)}$ from $G^{(t)}$ by merging $G^{(t)}_{\hat{a}}$ and $G^{(t)}_{\hat{b}}$. 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 $G^{(1)}_1, \ldots, G^{(n)}_n$ 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 9.1:

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 9.4.2. 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 9.4.4 for an illustration. Next section presents another recipe for clustering, based on more global informations, carried by the singular vectors of $X$. 
9.3 Spectral clustering

9.3.1 Spectral clustering recipe

The recipe of spectral clustering algorithms is to compute the $K$ first principal components of $X$ 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 (9.2), 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 $\theta_k$, 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 $X$, we apply a clustering algorithm on the rows of a good estimator $\hat{B}$ 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 $\hat{V}_K$ spanned by the $K$ first right singular vectors (principal axes) of $X$. The coordinates of the projection $\text{Proj}_{\hat{V}_K}(X_i)$ are given by the $i$-th coordinates of the $K$ first principal components $\hat{c}_k = \hat{\sigma}_k \hat{u}_k$ of $X$. This line of reasoning leads to the spectral clustering algorithm described below.

### Spectral clustering algorithm

1. Compute the singular value decomposition $X = \sum_{k \geq 1} \hat{\sigma}_k \hat{u}_k \hat{v}_k^T$;

2. Extract the $K$ first principal components

   $$\hat{C}_K = [\hat{c}_1, \ldots, \hat{c}_K] := [\hat{\sigma}_1 \hat{u}_1, \ldots, \hat{\sigma}_K \hat{u}_K]$$

3. Apply a clustering procedure on the rows of $\hat{C}_K$ in order to get a partition $\hat{G}$ 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 9.3.2, 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 $X$ involves the whole matrix $X$, 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 $\hat{C}_K = [\hat{c}_1, \ldots, \hat{c}_K]$, is the clustering of the rows of the left-singular vectors matrix $\hat{U}_K = [\hat{u}_1, \ldots, \hat{u}_K]$. Next lemma shows that this clustering also makes sense.
Lemma 9.1 Let
\[ A\Theta = \sum_{k=1}^{K} \sigma_k u_k v_k^T \]
be a singular decomposition of \( A\Theta \), and set \( U = [u_1, \ldots, u_K] \in \mathbb{R}^{n \times K} \). Then, there exist \( Z_1, \ldots, Z_K \in \mathbb{R}^K \), such that
\[ U_i: = Z_k \text{ for all } i \in G_k^*, \text{ and } \|Z_k - Z_\ell\|^2 = \frac{1}{|G_k^*|} + \frac{1}{|G_\ell^*|}, \text{ for all } k \neq \ell. \]

Proof of Lemma 9.1. Let us define \( \Delta = \text{diag} \left( |G_1^*|^{-1/2}, \ldots, |G_K^*|^{-1/2} \right) \) and \( B := A\Delta \). We notice that the columns of \( B \) are
\[ b_k = [A\Delta]_k = \left[ \frac{1_{i \in G_k^*}}{|G_k^*|^{1/2}} \right] \text{ for } i = 1, \ldots, n. \]
In particular, the columns \( b_1, \ldots, b_K \) of \( B \) are orthonormal. As
\[ \text{span} \{b_1, \ldots, b_K\} = \text{range}(A) \supset \text{range}(A\Theta) = \text{span} \{u_1, \ldots, u_K\}, \]
with \( b_1, \ldots, b_K \) orthonormal, the projection on \( \text{range}(A) \) is given by \( BB^T \) and \( BB^T u_k = u_k \) for \( k = 1, \ldots, K \). Hence,
\[ U = [u_1, \ldots, u_K] = B B^T U, \]
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
\[ U_{ij} = \sum_{\ell=1}^{K} 1_{i \in G_\ell^*} (\Delta R)_{\ell,j} = (\Delta R)_{kj} \text{ for } i \in G_k^*. \]
Hence, for \( i \in G_k^* \), we have \( U_i: = Z_k \), where the vectors
\[ Z_k := [\Delta R]_k = \frac{|G_k^*|^{-1/2} R_k}{}, \text{ for } k = 1, \ldots, K, \]
are orthogonal with square norm \( \|Z_k\|^2 = 1/|G_k^*| \), as \( ||R_k||^2 = 1 \). Hence,
\[ \|Z_k - Z_\ell\|^2 = \frac{1}{|G_k^*|} + \frac{1}{|G_\ell^*|}, \text{ for } \ell \neq k. \]
The proof of Lemma 9.1 is complete. \( \square \)

Debiasing spectral clustering
The left-singular vectors \( \tilde{u}_1, \tilde{u}_2, \ldots \) of \( X \) correspond to the eigenvectors of \( XX^T \). Computing the expectation of an entry
\[ (XX^T)_{ij} = (A\Theta \Theta^T A^T)_{ij} + E_i \Theta^T E_j^T \Theta A^T + \Theta^T A^T E_j, \]
we get
\[ \mathbb{E} \left[ (XX^T)_{ij} \right] = (A\Theta \Theta^T A^T)_{ij} + \mathbb{E} \left[ E_i \Theta^T E_j \right] \]
\[ = (A\Theta \Theta^T A^T)_{ij} + 1_{i=j} \text{Tr} (\text{Cov}(E_i)). \]
Let us denote by $\Gamma$ the diagonal matrix, with entries $\Gamma_{ii} = \text{Tr}(\text{Cov}(E_i))$. We observe that we have in expectation

$$\mathbb{E} [XX^T] = A\Theta\Theta^T A^T + \Gamma.$$ 

When $\Gamma$ 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 $XX^T$ are not biased. When $\Gamma$ is not proportional to the identity matrix, it is wise to reduce the bias of $XX^T$ by considering the eigenvalue decomposition of $XX^T - \hat{\Gamma}$, for some estimator $\hat{\Gamma}$ of $\Gamma$. Yet, unless the matrix $\Gamma$ is known in advance, it is not straightforward to design a (good) estimator $\hat{\Gamma}$ of $\Gamma$. We refer to the Exercise 9.4.3 for an example of such an estimator.

Let us sum-up the debiased spectral algorithm.

**Debiased spectral clustering algorithm**

1. Compute the eigenvalue decomposition $XX^T - \hat{\Gamma} = \sum_{k \geq 1} \hat{d}_k \hat{u}_k \hat{u}_k^T$, with eigenvalues ranked in decreasing order;
2. Apply a clustering procedure either on the rows of $\hat{U}_K = [\hat{u}_1, \ldots, \hat{u}_K]$ or on the rows of $\hat{C}_K = [\hat{d}_1^{1/2} \hat{u}_1, \ldots, \hat{d}_K^{1/2} \hat{u}_K]$, in order to get a partition $\hat{G}$ of $\{1, \ldots, n\}$.

### 9.3.2 Recovery bounds

In this section, we investigate the ability of spectral clustering to recover the partition $G^*$ from $X$. 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 $N(z_i \theta, \sigma^2 I_p)$ for $i = 1, \ldots, n$. Stacking as before the observations $X_1, \ldots, X_n$ into a $n \times p$ matrix $X$, we then observe

$$X = \varepsilon \theta^T + E,$$

where $\varepsilon \in \{-1, +1\}^n$ and the $E_{ij}$ are i.i.d. with a $N(0, \sigma^2)$ distribution. The underlying partition is $G^* = \{\{i : z_i = 1\}, \{i : z_i = -1\}\}$.

Let us define for $x \in \mathbb{R}^n$

$$|x|_0 = \sum_{i=1}^n 1_{x_i \neq 0}.$$ 

A good clustering algorithm, is an algorithm that recovers the vector $\varepsilon$, up to a sign change. Hence,
if $\tilde{z} \in \{-1, +1\}^n$ encodes the clustering output by this algorithm, ($\tilde{z}_i = 1$ if $i \in \tilde{G}_1$ and $\tilde{z}_i = -1$ if $i \in \tilde{G}_2$), we measure the quality of the clustering by the metric

$$\text{recov}(\tilde{z}) := \frac{1}{n} \min_{\delta \in \{-1, +1\}} |z - \delta \tilde{z}|_0,$$

which counts the proportion of mismatches between $\tilde{G}$ and $G^*$.

When $X$ follows the Model (9.3), we have

$$\mathbb{E}[XX^T] = \|\theta\|^2 zz^T + \Gamma,$$

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 $XX^T$ in the spectral clustering algorithm. Hence, we set $\hat{\Gamma} = 0$. Since $zz^T$ is of rank one, we only focus on the first eigenvector $\hat{u}_1$ of $XX^T$.

The first eigenvector $\hat{u}_1$ of $XX^T$ 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 (9.3) is that we can choose a very simple clustering procedure. Actually, as, hopefully, $\hat{u}_1 \approx \pm \frac{z}{\|z\|}$, we can simply take the sign of the entries of $\hat{u}_1$ in order to get a partition of $\{1, \ldots, n\}$ into two groups, corresponding to positive and negative entries of $\hat{u}_1$. We consider then the following spectral clustering algorithm

$$\tilde{z} = \text{sign}(\hat{u}_1), \text{ with } \hat{u}_1 \text{ a leading eigenvector of } \frac{1}{n}XX^T. \tag{9.5}$$

**Theorem 9.2** Assume that $X$ follows the model (9.3). There exists a numerical constant $C \geq 1$ such that, with probability at least $1 - 2e^{-n/2}$, the spectral clustering (9.5) fulfills the recovery bound

$$\text{recov}(\tilde{z}) \leq 1 \wedge \frac{C}{s^2}, \tag{9.6}$$

with $s^2$ defined by

$$s^2 = \frac{\|\theta\|^4}{\|\theta\|^2 \sigma^2 + \frac{p}{n} \sigma^4}. \tag{9.7}$$

We observe that the upper bound (9.6) 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 9.2.

**Proof of Theorem 9.2.**

Let us first connect the Hamming distance $|z - \delta \tilde{z}|_0$ to the square norm $\|z - \delta \sqrt{n} \hat{u}_1\|^2$.

**Lemma 9.3** For any $x \in \{-1, +1\}^n$ and $y \in \mathbb{R}^n$, we have

$$|x - \text{sign}(y)|_0 \leq 2 \min_{\alpha > 0} \|x - \alpha y\|^2.$$

This lemma simply follows from the inequality

$$|x_i - \text{sign}(y_i)| = |x_i - \text{sign}(\alpha y_i)| = 2 I_{x_i \neq \text{sign}(\alpha y_i)} \leq 2 |x_i - \alpha y_i|^2,$$

for any $\alpha > 0$ and $i = 1, \ldots, n$. 

From Lemma 9.3 with $\alpha = \sqrt{n}$ and $\|z\|^2 = n$, we get

$$
\frac{1}{n} \min_{\delta = -1, +1} |z - \delta \text{sign}(\hat{u}_1)|_0 \leq \frac{2}{n} \min_{\delta = -1, +1} \|z - \delta \sqrt{n} \hat{u}_1\|^2
$$

$$
= \frac{2}{n} \min_{\delta = -1, +1} (2n - 2\delta \sqrt{n}(z, \hat{u}_1))
$$

$$
= (4 - (z / \sqrt{n}, \hat{u}_1))^2
$$

where we have used that $|(z / \sqrt{n}, \hat{u}_1)| \leq \|z / \sqrt{n}||\|\hat{u}_1\| = 1$ 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 (7.4) with $A = \frac{1}{n} \|\theta\|^2 zz^T$ and $B = \frac{1}{n} XX^T$, we get

$$
\min_{\delta = -1, +1} \frac{1}{n} |z - \delta \hat{c}|_0 \leq 4(1 - (z / \sqrt{n}, \hat{u}_1)^2) \leq 16 \inf_{\theta \in \mathbb{R}} \frac{|A_{\theta}^n + \frac{1}{n} XX^T - \frac{1}{n} \|\theta\|^2 zz^T_{\|\theta\|^2}|_{\text{op}}}{\|\theta\|^4}
$$

$$
\leq 16 \frac{\|\theta\|^2 zz^T - \frac{1}{n} \|\theta\|^2 zz^T - \frac{p \|\theta\|^2}{n} I_n}_{\text{op}}^2.
$$

It remains to bound from above $\frac{1}{n} XX^T - \frac{1}{n} \|\theta\|^2 zz^T - \frac{p \|\theta\|^2}{n} I_n$.

**Lemma 9.4** There exists two exponential random variables $\xi, \xi'$ with parameter 1, such that the operator norm of

$$
W = \frac{1}{n} XX^T - \frac{1}{n} \|\theta\|^2 zz^T - \frac{p \|\theta\|^2}{n} I_n
$$

is upper-bounded by

$$
|W|_{\text{op}} \leq 4\sigma^2 \sqrt{\frac{p}{n}} \left(6 + 2\frac{\xi}{n}\right) + \left(48 + \frac{16\xi}{n}\right) \sigma^2 + 2\|\theta\| \sigma \left(1 + \sqrt{\frac{8\xi'}{n}}\right). \tag{9.9}
$$

Let us explain how Theorem 9.2 follows from the Bound (9.9). According to (9.8) and (9.9), we have with probability at least $1 - 2e^{-n/2}$, the upper bound

$$
\min_{\delta = -1, +1} \frac{1}{n} |z - \delta \hat{c}|_0 \leq 1 \wedge \left(\frac{16 \sqrt{\frac{7p}{n}} + 15^2 + 24\|\theta\|/\sigma}{\|\theta\|^2/\sigma^2}\right)^2.
$$

The right-hand side is smaller than 1 only if $15 \leq \|\theta\|/\sigma$, so $15^2 \leq 15\|\theta\|/\sigma$, from which follows

$$
\min_{\delta = -1, +1} \frac{1}{n} |z - \delta \hat{c}|_0 \leq 1 \wedge \left(\frac{16 \sqrt{7p/n} + 39\|\theta\|/\sigma}{\|\theta\|^2/\sigma^2}\right)^2
$$

$$
\leq 3584 \frac{p/n + \|\theta\|^2/\sigma^2}{\|\theta\|^4/\sigma^4} = \frac{3584}{s^2},
$$

which gives (9.6). It remains to prove Lemma 9.4.

**Proof of Lemma 9.4.**

We have $nW = (EE^T - p\sigma^2 I_n) + E\theta z^T + z\theta^T E^T$. 

---
EXERCISES

The quadratic term $E E^T - p \sigma^2 I_n$ is controlled by the bound (7.11) of Theorem 7.8: There exists an exponential random variable $\xi$ with parameter 1 such that

$$|E E^T - p \sigma^2 I_n|_{op} \leq 4 \sigma^2 \sqrt{p(6n + 2\xi)} + (48n + 16\xi)\sigma^2.$$  

Let us now control the cross terms in $W$.

**Lemma 9.5** There exists an exponential random variable $\xi'$ with parameter 1, such that

$$\frac{1}{n} |z \theta^T E^T|_{op} = \frac{1}{n} |E \theta z^T|_{op} \leq ||\theta|| \sigma \left(1 + \sqrt{\frac{8\xi'}{n}}\right). \quad (9.10)$$

**Proof of Lemma 9.5.** Dividing left and right-hand side of (9.10) by $\sigma$, 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,

$$||E u v^T x|| = |v^T x|||E u|| \leq ||E u||,$$

with equality for $x = v$. Hence $|E u v^T|_{op} = ||E u||$.

For the same reasons as in Step 2 of the proof of Theorem 7.8, the random variable $E u$ follows a standard Gaussian $\mathcal{N}(0, I_n)$ distribution. Hence, according to Hanson-Wright inequality (7.9) with $S = I_n$, there exists an exponential random variable $\xi'$ with parameter 1, such that

$$|E u v^T|_{op} = ||E u||^2 \leq n + \sqrt{8n \xi'} + 8 \xi' \leq \left(\sqrt{n} + \sqrt{8 \xi'}\right)^2.$$  

Since $\frac{1}{n} |E \theta z^T|_{op} = \frac{||\theta||}{\sqrt{n}} |E u v^T|_{op}$, the Bound (9.10) follows. \[ \square \]

Combining the Theorem 7.8, the Lemma 9.5, and the decomposition

$$n W = (E E^T - p \sigma^2 I_n) + E \theta z^T + z \theta^T E^T,$$

we get (9.9). The proof of Lemma 9.4 is complete. \[ \square \]

9.4 Exercises

9.4.1 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

$$S(n, k) = k S(n - 1, k) + S(n - 1, k - 1), \quad \text{for } 2 \leq k \leq n - 1.$$  

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 $n$. 
9.4.2 Exact recovery with hierarchical clustering

In this exercise, we provide some conditions ensuring that hierarchical clustering exactly recovers the hidden partition in the setting (9.3) of Section 9.3.2. We denote by \( \mathcal{W} = \{(i, j) : z_i = z_j, \ i < j\} \) 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 \( \mathcal{B} = \{(i, j) : z_i \neq z_j, \ i < j\} \) 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
\[
\mathbb{P} \left( \max_{(i, j) \in \mathcal{W}} \|X_i - X_j\|^2 \geq 2p\sigma^2 + (\sigma^2 \sqrt{96 \log(n)} + (48\sigma^2 \log(n)) \right) \leq \frac{1}{2n}.
\]

3. Similarly, prove that
\[
\mathbb{P} \left( \min_{(i, j) \in \mathcal{B}} \|X_i - X_j\|^2 \leq 2p\sigma^2 + 4\|\theta\|^2 - (\sigma^2 \sqrt{24p \log(n)} + (48\sigma^2 \log(n)) \right) \leq \frac{1}{2n}.
\]

4. Conclude that, when \( \|\theta\|^2 \geq \sigma^2 \left( \sqrt{24p \log(n)} + (24 \log(n)) \right) \), the hierarchical clustering algorithm with Euclidean distance and single or complete linkage recovers the clusters \( G_1 \) and \( G_2 \) with probability at least \( 1/n \).

9.4.3 Estimating \( \Gamma \)

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 9.3.1,
\[
(XX^T)_{ij} = (A\Theta\Theta^T A^T)_{ij} + E_i^T E_j + E_i^T (\Theta^T A^T) j + (A\Theta)_{ij} E_j,
\]
with \( \mathbb{E} \left[ E_i^T (\Theta^T A^T) j \right] = 0 = \mathbb{E} \left[ (A\Theta)_{ij} E_j \right] \), and
\[
\mathbb{E} \left[ E_i^T E_j \right] = \mathbf{1}_{i \neq j} \mathbb{E} \left[ \|E_i\|^2 \right].
\]

Let us denote by \( \overline{\Gamma} \) the diagonal matrix with \( \overline{\Gamma}_{ii} := \|E_i\|^2 \). In order to avoid the systematic bias induced by \( \overline{\Gamma} \) in the spectral decomposition, we would like to work on the matrix \( XX^T - \overline{\Gamma} \). 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 \( XX^T - \overline{\Gamma} \), where \( \overline{\Gamma} \) is built from data and is a good evaluation of \( \Gamma \).

If we knew the partition \( G^* \), “estimating\(^1\)” the random quantity \( \overline{\Gamma}_{ii} \) (or the parameter \( \Gamma_{ii} \)) would be a simple task. Actually, if \( i \neq i' \) belongs to the same group as \( i \), then
\[
\langle X_i - X_{i'}, X_i \rangle = \|E_i\|^2 - \langle E_i, E_{i'} \rangle + \langle \mu_i, E_i - E_{i'} \rangle,
\]
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 \( \overline{\Gamma}_{ii} \) (and \( \Gamma_{ii} \)).

The difficulty is that we do not know \( G^* \), and we need to estimate \( \Gamma_{ii} \) to estimate \( G^* \). To break this vicious spiral, we can build yet on (9.11), by replacing \( i' \) by a data-driven choice \( i^\ast \). We observe that we have the decomposition
\[
\langle X_i - X_{i^\ast}, X_i \rangle - \langle X_i - X_{i^\ast}, X_i \rangle = \langle X_i - X_{i^\ast}, X_{i^\ast} \rangle - \langle X_i - X_{i^\ast}, X_{i^\ast} \rangle,
\]

\(^1\)with some abuse of language, we use the word “estimation” even if \( \overline{\Gamma}_{ii} \) is a random quantity, not a parameter.
and we will be able to control the size of the right hand side, if we are able to control \( \max_k |\langle X_i - X_j, X_k \rangle| \). This observation motivates the definition of the following estimator

\[
\hat{\Gamma}_{ii} := \langle X_i - X_j, X_i \rangle, \quad \text{with} \quad \hat{\ell} \in \arg\min_j \max_{k: k \neq i,j} |\langle X_i - X_j, X_k \rangle|.
\]

(9.12)

In this exercise, you will prove the following bound on the error \( |\hat{\Gamma} - \Gamma|_\infty \).

**Proposition 9.6** Assume that the observations \( X_1, \ldots, X_n \) are independent, with \( X_i \) following a \( N(\mu_i, \Sigma_i) \) Gaussian distribution.

Assume also that each group \( G_k \) has a cardinality at least 2.

Then, with probability at least \( 1 - 2e^{-L} \), the diagonal matrix \( \hat{\Gamma} \) fulfills

\[
|\hat{\Gamma} - \Gamma|_\infty \leq 6\|\Theta\|_{2\infty} \sqrt{\Sigma_{Op}(3 \log(n) + L) + 10(\|\Sigma\|_F \sqrt{2 \log(n) + L}) \lor (\|\Sigma\|_{Op}(4 \log(n) + 2L)),
\]

where

\[
\|\Theta\|_{2\infty} := \max_k \|\theta_k\|, \quad \|\Sigma\|_{Op} := \max_i \|\Sigma_i\|_{Op} \quad \text{and} \quad \|\Sigma\|_F := \max_i \|\Sigma_i\|_F.
\]

While this estimator \( \hat{\Gamma} \) 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 9.6, answer to the following questions.

1. Take \( i' \neq i \) in the same group as \( i \). Starting from the decomposition

\[
\langle X_i - X_j, X_i \rangle = \langle X_i - X_i', X_i \rangle + \langle X_i - X_j, X_i' \rangle - \langle X_i - X_i', X_i' \rangle
\]

and using the definition of \( \hat{\ell} \), prove the inequalities

\[
|\hat{\Gamma}_{ii} - \hat{\Gamma}_{ii'}| \leq |\langle \mu_i, E_i - E_i' \rangle| + |\langle E_i', E_i \rangle| + \max_{k \neq i,i'} |\langle X_i - X_j, X_k \rangle| + \max_{k \neq i,i'} |\langle X_i - X_i', X_k \rangle|
\]

\[
\leq |\langle \mu_i, E_i - E_i' \rangle| + |\langle E_i', E_i \rangle| + 2 \max_{k \neq i,i'} |\langle X_i - X_i', X_k \rangle|
\]

\[
\leq 3 \max_{k,i,i'} |\langle \mu_k, E_i - E_i' \rangle| + 3 \max_{k,i} |\langle E_i, E_k \rangle|.
\]

2. Check that \( \text{Var}(\langle \mu_k, E_i - E_i' \rangle) = \mu_k^T \Sigma_i + \Sigma_{i'} \mu_k \leq 2 \|\Sigma\|_{Op} \|\Theta\|_{2\infty} \). What is the distribution of \( \langle \mu_k, E_i - E_i' \rangle \)?

3. Prove the bound

\[
P \left[ \max_{k,i,i'} |\langle \mu_k, E_i - E_i' \rangle| > 2 \|\Theta\|_{2\infty} \sqrt{\Sigma_{Op}(3 \log(n) + L)} \right] \leq e^{-L}.
\]

4. We can write the scalar product \( \langle E_i, E_k \rangle \) as \( \varepsilon_i^T \Sigma_{1/2} \Sigma_{1/2}^T \varepsilon_k \), with \( \varepsilon_i, \varepsilon_k \), two independent standard Gaussian random variables in \( \mathbb{R}^P \). With Hanson-Wright inequality (7.10) page 77, proves that

\[
P \left[ |\langle E_i, E_k \rangle| > (2 \|\Sigma_i^{1/2} \Sigma_k^{1/2} \|_F \sqrt{L}) \lor (4 \|\Sigma_i^{1/2} \|_{Op} \Sigma_k^{1/2} \|_F \sqrt{L}) \right] \leq e^{-L}.
\]

5. Deduce from the previous question that

\[
P \left[ \max_{i \neq k} |\langle E_i, E_k \rangle| > (2 \|\Sigma\|_F \sqrt{2 \log(n) + L}) \lor (4 \|\Sigma\|_{Op} (2 \log(n) + L)) \right] \leq e^{-L},
\]

and conclude the proof of Proposition 9.6.
9.4.4 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 $N(\theta, I_2/2)$ distribution and the other half are i.i.d. with a Gaussian $N(-\theta, I_2)$ 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 these lectures notes). The results are displayed in Figure 9.2.

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 9.3.
Figure 9.3: Dendrograms built from complete linkage (left), single linkage (middle) and average linkage (right).

You can reproduce these results by running the following R-code.

```r
# 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 9.4.

![Figure 9.4: Results for well separated clusters (\( \mu = [1.5, 1.5] \)).](image-url)
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

\[
\mathbb{E} \left[ e^{i\langle \lambda, Y \rangle} \right] = \exp \left( i\langle \lambda, m \rangle - \frac{1}{2} \lambda^T \Sigma \lambda \right), \quad \text{for all } \lambda \in \mathbb{R}^d. \tag{A.1}
\]

When matrix \( \Sigma \) 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 \( AY + b \)

\[
\mathbb{E} \left[ e^{i\langle \lambda, AY + b \rangle} \right] = \mathbb{E} \left[ e^{i(A^T \lambda, Y) + i\langle \lambda, b \rangle} \right] = \exp \left( i(A^T \lambda, m) - \frac{1}{2} (A^T \lambda)^T \Sigma A^T \lambda \right)e^{i\langle \lambda, b \rangle}
\]

\[
= \exp \left( i\langle \lambda, Am + b \rangle - \frac{1}{2} \lambda^T A\Sigma A^T \lambda \right).
\]

The second identity is obtained with \( A = a^T \) and \( b = 0 \). \( \square \)

**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 \( \Sigma \). Then the variables \( \text{Proj}_S Y \) and \( \text{Proj}_V Y \) are independent and follow, respectively, the \( \mathcal{N}(\text{Proj}_S m, \text{Proj}_S \Sigma \text{Proj}_S) \) and \( \mathcal{N}(\text{Proj}_V m, \text{Proj}_V \Sigma \text{Proj}_V) \) Gaussian distribution.
Proof. Since the projection matrices $\text{Proj}_S$ and $\text{Proj}_V$ are symmetric, we obtain that the joint characteristic function of $\text{Proj}_S Y$ and $\text{Proj}_V Y$ is
\[
\mathbb{E} \left[ e^{i\langle \lambda, \text{Proj}_S Y \rangle + i\langle \gamma, \text{Proj}_V Y \rangle} \right] = \mathbb{E} \left[ e^{i(\text{Proj}_S \lambda + \text{Proj}_V \gamma) \cdot Y} \right]
\]
\[
= \exp \left( i(\langle \lambda, \text{Proj}_S \rangle + \langle \gamma, \text{Proj}_V \rangle) - \frac{1}{2} (\text{Proj}_S \lambda + \text{Proj}_V \gamma)^T \Sigma (\text{Proj}_S \lambda + \text{Proj}_V \gamma) \right)
\]
\[
= \exp \left( i\langle \lambda, \text{Proj}_S m \rangle - \frac{1}{2} \lambda^T \text{Proj}_S \Sigma \text{Proj}_S \lambda \right)
\times \exp \left( i\langle \gamma, \text{Proj}_V m \rangle - \frac{1}{2} \gamma^T \text{Proj}_V \Sigma \text{Proj}_V \gamma \right)
\]
\[
= \mathbb{E} \left[ e^{i\langle \lambda, \text{Proj}_S Y \rangle} \right] \mathbb{E} \left[ e^{i\langle \gamma, \text{Proj}_V Y \rangle} \right].
\]
We conclude with Lemma A.1. □

A.2 Chi-Square Distribution

Let $Y \in \mathbb{R}^n$ be a random vector with $N(0, I_n)$ Gaussian distribution. The $\chi^2$ 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 $N(0, I_n)$ 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 $N(0, \text{Proj}_S)$ Gaussian distribution and the square-norm $\|\text{Proj}_S Y\|^2$ follows a $\chi^2$-distribution of degree $d$.

In particular, $\mathbb{E} \left[ \|\text{Proj}_S Y\|^2 \right] = \dim(S)$.

Proof. The projection $\text{Proj}_S$ is symmetric, so $\text{Proj}_S \text{Proj}_S^T = \text{Proj}_S$ and $\text{Proj}_S Y$ follows a $N(0, \text{Proj}_S)$ Gaussian distribution according to Lemma A.1.

Let $u_1, \ldots, u_d$ be an orthonormal basis of $S$ and set $U = [u_1, \ldots, u_d]$. Since $U^T U = I_d$, the vector $U^T Y$ follows a $N(0, I_d)$-distribution and
\[
\|\text{Proj}_S Y\|^2 = \sum_{k=1}^{d} (u_k^T Y)^2 = \|U^T Y\|^2
\]
follows a $\chi^2$ 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 $A = \{1, \ldots, k\}$ and $B = \{1, \ldots, p\} \setminus A$, and a Gaussian random vector $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \in \mathbb{R}^p$ with $N(0, \Sigma)$ distribution. We assume that $\Sigma$ is nonsingular and write $K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix}$ for its inverse.

In the next formulas, $K^{-1}_{AA}$ 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 $N(-K^{-1}_{AA}K_{AB}X_B, K^{-1}_{AA})$ distribution. In other words, we have the decomposition

$$X_A = -K^{-1}_{AA}K_{AB}X_B + \epsilon_A, \quad \text{where } \epsilon_A \sim N(0, K^{-1}_{AA}) \text{ is independent of } X_B.$$  \hfill (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

$$g(x_A|x_B) = \frac{1}{(2\pi)^{k/2}} \exp \left( -\frac{1}{2} x_A^T K_{AA} x_A - x_A^T K_{AB} x_B - \frac{1}{2} x_B^T (K_{BB} - \Sigma^{-1}_{BB}) x_B \right),$$

with $\Sigma_{BB}$ the covariance matrix of $X_B$. Since $\Sigma^{-1}_{BB} = K_{BB} - K_{BA}K_{AA}^{-1}K_{AB}$, 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 $N(-K^{-1}_{AA}K_{AB}x_B, K^{-1}_{AA})$ distribution. \hfill \Box

**Corollary A.5** For any $a \in \{1, \ldots, p\}$, we have

$$X_a = -\sum_{b \cdot b \neq a} \frac{K_{ab}}{K_{aa}} X_b + \epsilon_a, \quad \text{where } \epsilon_a \sim N(0, K^{-1}_{aa}) \text{ is independent of } \{X_b : b \neq a\}. \hfill (A.3)$$

**Proof.** We apply the previous lemma with $A = \{a\}$ and $B = A^c$. \hfill \Box

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

$$\text{cor}(X_a, X_b|X_c : c \neq a, b) = \frac{\text{cov}(X_a, X_b|X_c : c \neq a, b)}{\sqrt{\text{var}(X_a|X_c : c \neq a, b) \text{var}(X_b|X_c : c \neq a, b)}}.$$  \hfill (A.4)

**Corollary A.6** For any $a, b \in \{1, \ldots, p\}$, we have

$$\text{cor}(X_a, X_b|X_c : c \neq a, b) = \frac{-K_{ab}}{\sqrt{K_{aa}K_{bb}}}. \hfill (A.4)$$

**Proof.** The previous lemma with $A = \{a, b\}$ and $B = A^c$ gives

$$\text{cov}(X_A|X_B) = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ab} & K_{bb} \end{pmatrix}^{-1} = \frac{1}{K_{aa}K_{bb} - K_{ab}^2} \begin{pmatrix} K_{bb} & -K_{ab} \\ -K_{ab} & K_{aa} \end{pmatrix}.$$  \hfill (A.4)

Plugging this formula in the definition of the conditional correlation, we obtain Formula (A.4). \hfill \Box
Appendix B

Constrained optimization

Let \( f, g_1, \ldots, g_N \) be \( N + 1 \) functions from \( \mathbb{R}^d \) to \( \mathbb{R} \). In this appendix, we recall some basic results related to the minimization problem

\[
\min_{x \in \mathbb{R}^d : g_1(x) \leq 0, \ldots, g_N(x) \leq 0} f(x).
\] (B.1)

B.1 Dual problem

B.1.1 Lagrangian and dual functions

Two functions play an important role in the investigation of the optimization problem (B.1): The Lagrangian function

\[
L(x, \lambda) = f(x) + \sum_{j=1}^{N} \lambda_j g_j(x), \quad \text{for} \ (x, \lambda) \in \mathbb{R}^d \times \mathbb{R}^N,
\] (B.2)

and the dual function

\[
q(\lambda) = \inf_{x \in \mathbb{R}^d} L(x, \lambda), \quad \text{for} \ \lambda \in \mathbb{R}^N.
\] (B.3)

Since \( q(\lambda) \) is an infimum of affine functions, the dual function \( q \) is concave.

B.1.2 Weak duality

Let us abbreviate the \( N \) conditions \( \lambda_j \geq 0, j = 1, \ldots, N \) by \( \lambda \geq 0 \). For any \( \lambda \geq 0 \) and \( x \in \mathbb{R}^d \) such that \( g_j(x) \leq 0 \) for \( j = 1, \ldots, N \), we have

\[
q(\lambda) \leq L(x, \lambda) = f(x) + \sum_{j=1}^{N} \lambda_j g_j(x) \leq f(x).
\]

So, for any \( \lambda \geq 0 \), we have

\[
q(\lambda) \leq \min_{g_j(x) \leq 0} f(x).
\]

We then obtain the following lower bound on the minimization problem (B.1)

\[
\sup_{\lambda \geq 0} q(\lambda) \leq \min_{g_j(x) \leq 0} f(x) \quad \text{(weak duality)}. \] (B.4)

This inequality is called the weak-duality condition. We observe that while the primal problem (B.1) can be hard to solve numerically in general, when the function \( q \) has an explicit expression, the dual problem

\[
\sup_{\lambda \geq 0} q(\lambda) \quad \text{(dual problem)} \] (B.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 (B.1) and the value of the dual problem (B.5) is called the duality gap. When the inequality in (B.4) is strict, this gap is positive and otherwise it is zero.

**B.1.3 Finding a solution**

When the functions are convex and differentiable, we can give some simple necessary conditions for $x^*$ to be a minimizer of the primal problem (B.1).

**Lemma B.1 Karush-Kuhn-Tucker (KKT) conditions**

Assume that $f, g_1, \ldots, g_N$ are convex and differentiable. If $(x^*, \lambda^*) \in \mathbb{R}^d \times \mathbb{R}^N$ are such that

\[
\nabla f(x^*) + \sum_{j=1}^{N} \lambda_j^* \nabla g_j(x^*) = 0 \quad \text{(first order condition)}
\]

\[
\min(\lambda_j^*, -g_j(x^*)) = 0 \quad \text{(slackness condition)}
\]

then we have

\[
q(\lambda^*) = \max_{\lambda \geq 0} q(\lambda) = \inf_{g_j(x) \leq 0} f(x) = f(x^*).
\]

The conclusion of Lemma B.1 is twofold. First, if we find $(x^*, \lambda^*)$ fulfilling the first order and the slackness conditions, then $x^*$ minimizes the primal problem (B.1). Second, there is no duality gap, and $\lambda^*$ is solution of the dual problem (B.5).

We underline that the slackness condition combines two conditions:
1. both $\lambda_j^*$ and $g_j(x^*)$ must be non-negative;
2. at least one of these two quantities is zero.

The first condition garantees that $x^*$ and $\lambda^*$ fulfill the conditions of the primal and the dual problem respectively. 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 B.1.** Since $L$ is convex in $x$ and since $x^*$ is a critical point for $x \to L(x, \lambda^*)$ we have

\[
L(x^*, \lambda^*) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*) = q(\lambda^*).
\]

By the slackness condition, we then have

\[
q(\lambda^*) = L(x^*, \lambda^*) = f(x^*) + \sum_{j=1}^{N} \lambda_j^* g_j(x^*) = f(x^*).
\]

Furthermore, since $\lambda^* \geq 0$ and since $x^*$ fulfills the conditions $g_j(x^*) \leq 0$ for $j = 1, \ldots, N$, the weak-duality ensures that

\[
q(\lambda^*) \leq \sup_{\lambda \geq 0} q(\lambda) \leq \min_{g_j(x) \leq 0} f(x) \leq f(x^*).
\]

Combining this inequality with (B.6), we get the conclusion. □
Index

Bregman divergence, 28
Bregman projection, 30
chi-square distribution, 110
conditional correlation, 111
convex function, 71
Davis-Kahan inequality, 74
Dendrogram, 97
Dual function, 113
Dual problem, 113
Duality, 113
Exponential weight aggregation, 23, 26
Fano’s lemma, 54
Frobenius norm, 67
Hanson-Wright inequality, 76
Hierarchical clustering, 96, 105
Hoeffding concentration inequality, 3
KKT conditions, 114
Kullback-Leibler divergence, 32, 45
Ky–Fan norm, 68
Lagrangian function, 113
Low rank approximation, 68
Marchenko-Pastur distribution, 81
Max–Min formula, 67, 73
Median of means, 5
Minimax risk, 53
Mirror Descent, 29
Moore–Penrose pseudo-inverse, 70, 71
Multi-armed bandits, 40
nuclear norm, 68
operator norm, 68, 73
PCA: Principal Component Analysis, 83
Pinsker Inequality, 33, 46
Projected gradient descent, 17
ridge regression, 70
singular value, 65, 66, 67, 73
singular vector, 65
Spectral clustering, 98
Stochastic gradient descent, 14
SubGaussian distribution, 1
SVD: Singular Value Decomposition, 64, 67, 71, 83
UCB policy, 44
Wald formula, 41
Weyl inequality, 73
X-armed bandits, 50