Selected topics on robust statistical learning theory

Lecture Notes

Matthieu Lerasle
CNRS-CREST-Ensae

Contents

Chapter 1

Introduction

1.1 Statistical learning

These notes gather some results dealing with robustness issues in statistical learning. Most of the results lie within the framework introduced by Vapnik [58], see also [44]. Given a dataset $\mathcal{D}_N = (Z_1, \dots, Z_N)$, where each Z_i belongs to a measurable space $\mathcal{Z}$, the goal is to infer from $\mathcal{D}_N$ relevant informations regarding the stochastic mechanism that generated $\mathcal{D}_N$. To proceed, assume first that all data have the same (unknown) distribution P and let Z denote a random variable with distribution P independent of $\mathcal{D}_N$. Choose a set of parameters F and a real valued function $\ell : F \times \mathcal{Z} \to \mathbb{R}$, $(f, z) \mapsto \ell_f(z)$, ℓ is called the *loss*. Based on this loss, the *risk* of any parameter $f \in F$ is defined as the integral of ℓ_f with respect to the distribution P:

$$\forall f \in F, \qquad P\ell_f := \mathbb{E}_{Z\sim P}[\ell_f(Z)] \ .$$

The goal is to infer from $\mathcal{D}_N$ the "best" parameter f^* in F which is the one minimizing the risk:

$$f^* \in \operatorname{argmin}_{f\in F} P\ell_f \ .$$

Hereafter, such a minimizer is assumed to exist to simplify notations. The interested reader can check that all results pertain if $P\ell_{f^*}$ is replaced by $\inf_{f\in F} P\ell_f$ in the following. f^* is unknown as it depends on P, it is usually called the *oracle* as it is the parameter that would have chosen someone knowing the distribution P. It cannot be used as an estimator, it is rather an ideal that any procedure tries to mimic. Indeed, most of the material presented here aims at bounding the *excess risk* of any estimator $\widehat{f} \in F$ defined by

$$\mathcal{E}(\widehat{f}) = P[\ell_{\widehat{f}} - \ell_{f^*}] = \mathbb{E}[\ell_{\widehat{f}}(Z) - \ell_{f^*}(Z)|\mathcal{D}_N] \ .$$

For any $f \in F$, $\mathcal{E}(f) = P[\ell_f - \ell_{f^*}]$ measures by how much f fails to minimize $P\ell_f$. It is worth noticing that $\mathcal{E}(\widehat{f})$ is a random variable, the integral defining the risk being with respect to the random variable $Z \sim P$ that is *independent* of $\mathcal{D}_N$. Bounding $\mathcal{E}(\widehat{f})$ from above means here finding $\Delta_{N,\delta}(F)$ such that

$$\mathbb{P}(\mathcal{E}(\widehat{f}) \leqslant \Delta_{N,\delta}(F)) \geqslant 1 - \delta \ .$$

This type of result will be refered to as *oracle inequality* as it compares the risk of the estimator $P\ell_{\widehat{f}}$ with the one of an oracle $P\ell_{f^*} = \inf_{f\in F} P\ell_f$. This problem covers many classical problems in statistics and learning, we present here some basic examples, other will follow in the notes.

Univariate mean estimation In this example, given real valued random variables $Z_1, \dots, Z_N$ with common distribution P, the goal is to infer the expectation $P[Z] = \mathbb{E}_{Z\sim P}[Z]$.

Set $F = \mathbb{R}$ and let $\ell_f(z) = (z-f)^2$ so, if $P[Z^2] < \infty$, then for any $f \in \mathbb{R}$, the expectation $f^* = P[Z]$ satisfies

$$P\ell_f = \mathbb{E}[(Z-f)^2] = (f-f^*)^2 + \mathbb{E}[(Z-f^*)^2] = (f-f^*)^2 + P\ell_{f^*} \ .$$

It follows that f^* is the unique minimizer of $P\ell_f$ over F. This example is simultaneously the simplest one can imagine and a natural building block for any learning procedure. Chapter 2 is therefore dedicated to this elementary problem.

Multivariate mean estimation Assume now that data $Z \in \mathbb{R}^d$ and, setting $F = \mathbb{R}^d$, the goal is to estimate $f^* = P[Z]$. Let $\|\cdot\|$ denote the Euclidean norm and let $\ell_f(z) = \|f-z\|^2$. For any $f \in F$, it holds

$$\begin{aligned} P\ell_f &= \mathbb{E}[\|Z-f\|^2] \\ &= \mathbb{E}[\|Z-f^*\|^2] + 2\mathbb{E}[(Z-f^*)^T(f-f^*)] + \|f-f^*\|^2 \\ &= P\ell_{f^*} + \|f-f^*\|^2 \ . \end{aligned}$$

Here the second equality follows by linearity of the expectation and $\mathbb{E}[Z - f^*] = 0$. It follows that f^* is the unique minimizer of $P\ell_f$ over F. This example allows to understand the central role of uniform concentration inequalities to bound the excess risk of estimators. Chapter 4 is dedicated to this problem.

Regression While previous problems are typical examples of *unsupervised* learning tasks where data are not labeled, regression is one of the most classical example of *supervised* learning task where data are labeled: $Z = (X, Y)$ with X the input or feature taking values in a measurable space $\mathcal{X}$ and Y is the output or label takes value in a subset $\mathcal{Y} \subset \mathbb{R}$. The goal is to *predict* Y from X. The purpose of regression is to estimate the regression function defined as any function f^* such that, for any bounded measurable function $\varphi : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}[Y\varphi(X)] = \mathbb{E}[f^*(X)\varphi(X)] \ .$$

Assume that $P[Y^2] < \infty$, let $F = L^2(P_X)$ and $\ell_f(x,y) = (y-f(x))^2$. For any $f \in F$,

$$\begin{aligned} P\ell_f &= \mathbb{E}[(Y-f^*(X))^2] + 2\mathbb{E}[(Y-f^*(X))(f^*(X)-f(X))] + \mathbb{E}[(f^*(X)-f(X))^2] \\ &= P\ell_{f^*} + \mathbb{E}[(f^*(X)-f(X))^2] \ . \end{aligned}$$

It follows that f^* is P_X-almost surely the unique minimizer of $P\ell_f$. An important difference with the previous examples is that the "natural" set of parameters F is here infinite dimensional. To bound properly the risk of the estimators,

it is necessary to consider strict subsets $F_0 \subset F$ and consider only estimators taking values in F_0. This implies that, rather than the regression function f^*, the estimators are more natural estimators of the "local" oracle

$$f_0^* \in \operatorname{argmin}_{f \in F_0} P\ell_f \ ,$$

provided that such function exists. Chapter 7 is dedicated to the least-squares regression problem.

Empirical risk minimisation

One of the most classical algorithm in statistical learning is empirical risk minimization, see [58], which considers the estimator $\widehat{f}_{\mathrm{erm}}$ of f^* defined by

$$\widehat{f}_{\mathrm{erm}} \in \operatorname{argmin}_{f \in F} P_N \ell_f, \qquad \text{where} \qquad P_N \ell_f := \frac{1}{N} \sum_{i=1}^{N} \ell_f(Z_i) \ .$$

One of the reasons explaining the success of this estimator is that it is minimax optimal in many problems. Minimax optimal rates can usually be proved for the ERM in problems where data are assumed *independent, identically distributed and sub-Gaussian*. In the univariate and multivariate mean estimation problems, the empirical risk minimizer $\widehat{f}_{\mathrm{erm}}$ is simply the empirical mean $N^{-1}\sum_{i=1}^{N} Z_i$. In these examples, data are called sub-Gaussian if the Laplace transform of P is bounded from above by the one of Gaussian random variable. In the univariate mean estimation problem, this means that there exists $\sigma^2 > 0$ such that

$$\forall s > 0, \qquad \log \mathbb{E}[e^{s(Z - \mathbb{E}[Z])}] \leqslant \frac{s^2\sigma^2}{2} \ . \tag{1.1}$$

Under this assumption, Markov's inequality ensures that, for any $t > 0$, and $s > 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} Z_i > \mathbb{E}[Z] + t \right) &= \mathbb{P}(e^{sN^{-1}\sum_{i=1}^{N}(Z_i - \mathbb{E}[Z])} > e^{st}) \\
&\leqslant e^{-st + \log \mathbb{E}[e^{sN^{-1}\sum_{i=1}^{N}(Z_i - \mathbb{E}[Z])}]} \\
&= e^{-st + \sum_{i=1}^{N} \log \mathbb{E}[e^{sN^{-1}(Z_i - \mathbb{E}[Z])}]} \\
&= e^{-st + \frac{s^2\sigma^2}{2N}} \ .
\end{aligned}$$

Optimizing over s yields

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} Z_i > \mathbb{E}[Z] + t \right) \leqslant e^{-Nt^2/2\sigma^2} \ ,$$

or, equivalently

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} Z_i > \mathbb{E}[Z] + \sigma\sqrt{\frac{2t}{N}} \right) \leqslant e^{-t} \ .$$

Sub-Gaussian deviations of the empirical mean are central in the analysis of ERM. They are somehow optimal as shown in [14, Proposition 6.1], the result is

recalled in Proposition 1 of Chapter 2. The sub-Gaussian deviation inequality only involves moments of order 1 and 2 of the Z_i, and an interesting question is whether this inequality remains valid if the Z_i are only assumed to have finite moments of order 2. As explained in Chapter 2, sub-Gaussian deviations of the empirical mean only holds under the sub-Gaussian assumption (1.1). When Z is only assumed to have 2 moments, one cannot essentially do better than Chebishev's inequality (see [14, Proposition 6.2] that is recalled in Proposition 8) which states that

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^{N} Z_i > \mathbb{E}[Z] + \sigma \sqrt{\frac{2t}{N}} \right) \leqslant \frac{1}{t} \ .$$

Providing estimators of the mean with sub-Gaussian deviations in a relaxed setting where P is only assumed to have a finite second moment is one of the guidelines in these notes. In this example, this phenomenon holds because estimators are evaluated through their *deviation* properties rather than in expectation. Actually, in expectation

$$\mathbb{E}[\mathcal{E}(\widehat{f}_{\mathrm{erm}})] = \mathbb{E}[P\ell_{\widehat{f}_{\mathrm{erm}}} - P\ell_{f^*}] = \mathbb{E}[(\widehat{f}_{\mathrm{erm}} - f^*)^2] = \frac{\sigma^2}{N} \ .$$

This result is true if Z has a finite moment of order 2 and does not improve if Z is Gaussian. This is why these notes focus on Catoni's point of view, see [14], evaluating estimators by their deviation properties and proving oracle inequalities.

1.2 Robustness

Robustness is a classical topic in statistics that has been around since the seminal works of Hampel [22, 23, 24, 25, 26], Huber [29, 28] and Tukey [54, 55, 57, 56], see the classical textbook [30] for an overview. Informally, an estimator is called robust if it behaves nicely even when data are not i.i.d. and sub-Gaussian. This holds for example, when data are i.i.d. but satisfy only weak moment assumptions like the existence of a second moment only as in the example of univariate mean estimation. A large part of these notes deals with this issue. An extensive literature has also been studied the case where the dataset is "close" to the ideal setup, but may have been corrupted. This includes the following well known examples.

Model misspecification In statistics, this means that the distribution P of Z does not lie into the statistical model $\mathcal{P}$ where the estimator $\hat{P}$ of P lies. A classical example of Birgé [8] is the following: assume that P is the mixture $\mathrm{d}P(x) = (1 - 1/N)\mathbf{1}_{x \in [0,1]} + (1/N)\delta_{x=N^2}$ and that the statistical model is the set of uniform distributions $\mathcal{P} = \{\mathcal{U}[0,t], t > 0\}$. The distribution P is "close" to the model $\mathcal{P}$ since, for example, the Hellinger distance between P and the uniform distribution $\mathcal{U}([0,1])$ is bounded from above as follows:

$$h^2(P, \mathcal{U}([0,1])) \leqslant \frac{1}{N} \ .$$

However, one of the most classical ERM in statistics, the maximum likelihood estimator, has positive probability to be the distribution $\mathcal{U}[0, N^2]$ which is a very poor estimator of P.

Huber's contamination model. In this model, see [30], it is assumed that data are i.i.d. with common distribution

$$\mathrm{d}P = (1-\epsilon)\mathrm{d}P_I + \epsilon \mathrm{d}P_O \ .$$

P_I is the distribution of *inliers*, most of the sample is drawn from this distribution. P_I is the distribution on which one wants to make assumptions. P_O is the distribution of *outliers*. These are data corrupting the dataset that may have nothing to do with the learning task. Birgé's example is a particular instance of the Huber contamination problem where most of the data is drawn from the uniform distribution on $[0,1]$ but some data may be equal to N^2. Usually, very few assumptions are granted on the outliers distribution. However, in this model, these data are always independent and independent from the inliers.

The $O \cup I$ frameworks In this setting introduced in [33], there exists a partition (unknown to the statistician) of $\{1,\dots,N\}$ in two blocks O and I. Data $(Z_i)_{i\in O}$ are the *outliers*, nothing is assumed on these data. Data $(Z_i)_{i\in I}$ are the inliers on which one may grant some assumptions. This model is closely related to the ϵ-contamination model while being slightly different:

- Outliers may not be independent, nor independent from the other data $(Z_i)_{i\in I}$. This allows "aggressive" outliers which can look at the dataset to corrupt it.
- The proportion of outliers is fixed in the $O \cup I$ setting, it is random in the Huber contamination model (although concentrated around ϵ).

The main challenges in robust statistics are to *resist* and *detect* outliers. Resist means looking for procedures that behave in the ϵ-contamination model as well as "good" estimators such that MLE do when $P = P_I$. Detect means identifying outliers (think about fraud detection for example).

Of course, in both Huber's contamination model and $O \cup I$ frameworks, it is possible to consider situations where, besides being contaminated, the "inliers" (those distributed as P_I in Huber's contamination's model and data $(Z_i)_{i\in I}$ in the $O \cup I$ frameworks) only satisfy moment assumptions. In these notes, I will mostly consider the situation where data are i.i.d. hence, not contaminated (see however Section 4.6). It is an interesting exercise to check if the different results extend to contaminated settings and which proportion of outliers is tolerated by different methods.

1.3 What are these notes about

The notes are an attempt to extract important *principles* underlying the construction and theoretical analysis of estimators that are referred to as "robust". The main task is to build estimators that satisfy the same oracle inequalities as the ERM does when data have sub-Gaussian behavior in a relaxed setting where the Gaussian assumption is replaced by moment hypotheses. These principles are divided in four main categories.

- The median-of-means principle allows to build estimators of univariate mean estimation achieving sub-Gaussian deviations, see Chapter 2. This is arguably the simplest construction allowing to achieve such results.

- The minmax principle allows to build from estimators of increments $P[\ell_f - \ell_g]$ (which are univariate means), estimators of "oracles", see Chapters 4 and 5. The idea of using pairwise comparisons or tests to build estimators goes back to the works of Le Cam and Birgé, the minmax principle is an elegant formulation of this construction which makes a bridge between Birgé/Le Cam's construction and the ERM of Vapnik.

- The homogeneity lemma reduces the analysis of minmax estimators to deviation bounds of MOM processes on localized classes, see Chapter 5. The homogeneity lemma is an alternative to peeling arguments that can be used when deviation properties cannot be obtained at any confidence level.

- The *small ball method* allows to prove (uniform and sub-Gaussian) deviation inequalities of the median-of-mean processes, under weak assumptions but only up to a confidence level that decreases geometrically with the sample size N, see Chapter 3.

The combination of these principles allows to prove oracle inequalities simultaneously for the ERM in the sub-Gaussian framework (hence, providing the relevant benchmarks) and for robust alternatives such as minmax MOM estimators. These procedures, thanks to the median step, naturally resist to a small proportion of outliers in the dataset. Chapter 8 presents ρ-estimators of [3, 4, 6]. This presentation is not exhaustive, it stresses some links between this construction and both the minmax principle and the homogeneity lemma. It should be seen as an invitation to learn this powerful theory.

These notes do not cover many important development, they focus on very particular learning tasks and very particular robustness issues, which correspond to problems I have been mostly interested in regarding this subject. I hope that they will convince some readers to contribute to this rapidly growing literature. In particular, Chapter 9 presents (way too) briefly the literature on robust procedures that are computable in polynomial time. In particular, numerically efficient methods are not discussed here. In this direction, important results appeared recently. In particular, [37] presents a first algorithm which produces an estimator of multivariate mean with optimal sub-Gaussian deviation rates using a spectral algorithm rather than SDP relaxations as the one presented in Chapter 9. This new algorithm should behave much better numerically than its concurrent. Moreover, only the problem of multivariate mean expectation is considered in the note from this computational perspective. Going from this problem to more generic learning problems is well understood, see in particular [50]. This material should also be added in future version of these notes.

Chapter 2

Sub-Gaussian estimation of univariate means

This chapter focuses on one of the most simple problem in statistics where we want to estimate the expectation $\mu_P = P[X]$ of a distribution P on $\mathbb{R}$ from the observation of an i.i.d. sample $\mathcal{D}_N = (X_1, \ldots, X_N)$ of real valued random variables with common distribution P. These estimators are natural building blocks for more general learning tasks in the following chapters. We first establish the behaviour of the empirical mean from a *deviation* point of view. We prove that it achieves good subexponential deviation bounds when X is Gaussian and that Chebyshev's inequality is essentially sharp when X is only assumed to have a bounded second moment. Then, we study alternative estimators that achieve a certain type of sub-Gaussian deviation inequalities when X has only 2 finite moments.

Notation All along the chapter, $\mathcal{P}_2$ denotes the class of all probability distributions on $\mathbb{R}$ with finite second moment and $\mathcal{P} \subset \mathcal{P}_2$. For any $P \in \mathcal{P}_2$, μ_P denotes the expectation of P and σ_P^2 its variance. $X_1, \ldots, X_N$ denotes an i.i.d. sample and for any $P \in \mathcal{P}_2$, $\mathbb{P} = P^{\otimes N}$. An estimator $\widehat{\mu}$ of μ_P is a real valued random variable $\widehat{\mu} = F(\mathcal{D}_N)$, where $F : \mathbb{R}^N \to \mathbb{R}$ is a measurable function.

2.1 Empirical mean

The arguably most simple estimator of μ_P is the empirical mean

$$P_N X = \overline{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i \ .$$

2.1.1 Lower bounds

The empirical mean plays an important role in these notes in the case where the random variables are Gaussian. The reason is that the deviation of the empirical mean in this example are somehow extremal as can be seen from the following result.

Proposition 1. *[14, Proposition 6.1] Assume that $\mathcal{P}$ contains all Gaussian distributions $\mathrm{N}(\mu, \sigma^2)$. For any estimator $\widehat{\mu}$ of $\mu_P \in \mathbb{R}$, any $t > 0$, there exists $P \in \mathcal{P}$ such that the empirical mean $\overline{X}_N = N^{-1} \sum_{i=1}^N X_i$ satisfies either*

$$\mathbb{P}\left(\widehat{\mu} - \mu_P > t\right) \geqslant \mathbb{P}\left(\overline{X}_N > \mu_P + t\right) \quad \text{or} \quad \mathbb{P}\left(\widehat{\mu} - \mu_P < -t\right) \geqslant \mathbb{P}\left(\overline{X}_N < \mu_P - t\right) \ .$$

Proof. For any $i \in \{-t, t\}$, let P_i denote the Gaussian distribution with variance 1 and respective mean $\mu_{P_i} = i$. By construction

$$\begin{aligned}\mathbb{P}_t\left(\widehat{\mu} \leqslant \mu_{P_t} - t\right) + \mathbb{P}_{-t}\left(\widehat{\mu} \geqslant \mu_{P_{-t}} + t\right) &= \mathbb{P}_t\left(\widehat{\mu} \leqslant 0\right) + \mathbb{P}_{-t}\left(\widehat{\mu} \geqslant 0\right) \\ &\geqslant \left(\mathbb{P}_{-t} \wedge \mathbb{P}_t\right)\left(\widehat{\mu} \leqslant 0\right) + \left(\mathbb{P}_{-t} \wedge \mathbb{P}_t\right)\left(\widehat{\mu} \geqslant 0\right) \geqslant \left|\mathbb{P}_{-t} \wedge \mathbb{P}_t\right| \ .\end{aligned}$$

Here $\mathbb{P}_{-t} \wedge \mathbb{P}_t$ denotes the measure whose density is the minimum between those of $\mathbb{P}_{-t}$ and $\mathbb{P}_t$ and $|\mathbb{P}_{-t} \wedge \mathbb{P}_t|$ is its total variation. Now, $\mathbb{P}_i$ has density

$$\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|\mathbf{x} - \mu_{P_i}\mathbf{1}\|^2} \ ,$$

therefore, $\mathbb{P}_{-t} \wedge \mathbb{P}_t$ has density $\mathbb{P}_t$ for any $\mathbf{x} \in \mathbb{R}^N$ such that

$$\|\mathbf{x} - \mu_{P_t}\mathbf{1}\|^2 \geqslant \|\mathbf{x} - \mu_{P_{-t}}\mathbf{1}\|^2 \quad \text{that is, such that} \quad \overline{x}_N = N^{-1} \sum_{i=1}^N x_i \leqslant 0 \ .$$

Therefore,

$$\begin{aligned}\left|\mathbb{P}_{-t} \wedge \mathbb{P}_t\right| &= \mathbb{P}_t\left(\overline{X}_N \leqslant 0\right) + \mathbb{P}_{-t}\left(\overline{X}_N \geqslant 0\right) \\ &= \mathbb{P}_t\left(\overline{X}_N \leqslant \mu_{P_t} - t\right) + \mathbb{P}_{-t}\left(\overline{X}_N \geqslant \mu_{P_{-t}} + t\right) \ .\end{aligned}$$

Overall,

$$\begin{aligned}&\mathbb{P}_t\left(\widehat{\mu} \leqslant \mu_{P_t} - t\right) + \mathbb{P}_{-t}\left(\widehat{\mu} \geqslant \mu_{P_{-t}} + t\right) \\ &\qquad \geqslant \mathbb{P}_t\left(\overline{X}_N \leqslant \mu_{P_t} - t\right) + \mathbb{P}_{-t}\left(\overline{X}_N \geqslant \mu_{P_{-t}} + t\right) \ .\end{aligned}$$

This implies the result. □

2.1.2 Upper bounds in the sub-Gaussian case

In order to establish the benchmark for future estimators, recall the following upper bound on the deviations of the empirical mean in the Gaussian case.

Proposition 2. *If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the empirical mean $P_N X$ satisfies*

$$\forall t > 0, \qquad \mathbb{P}\left(|P_N X - \mu| > \sigma\sqrt{\frac{2t}{N}}\right) \leqslant e^{-t} \ .$$

Proof. Since $X \sim \mathrm{N}(\mu, \sigma^2)$, $P_N X \sim \mathrm{N}(\mu, \sigma^2/N)$ and $\sqrt{N}(P_N X - \mu)/\sigma \sim \mathrm{N}(0, 1)$. The Gaussian distribution satisfies

$$\begin{aligned}1 - \Phi(x) = \int_x^{+\infty} e^{-u^2/2} \frac{\mathrm{d}u}{\sqrt{2\pi}} &= \int_0^{+\infty} e^{-(u+x)^2/2} \frac{\mathrm{d}u}{\sqrt{2\pi}} \\ &\leqslant e^{-x^2} \int_0^{+\infty} e^{-u^2/2} \frac{\mathrm{d}u}{\sqrt{2\pi}} = \frac{e^{-x^2/2}}{2} \ .\end{aligned}$$

Therefore,

$$\forall x > 0, \qquad \mathbb{P}\left(\frac{\sqrt{N}(P_N X - \mu)}{\sigma} > x \right) \leqslant \frac{e^{-x^2/2}}{2} \ .$$

This is equivalent to

$$\forall t > 0, \qquad \mathbb{P}\left(P_N X - \mu > \sigma \sqrt{\frac{2t}{N}} \right) \leqslant \frac{e^{-t}}{2} \ .$$

Applying this inequality to $-X_i$ yields

$$\forall t > 0, \qquad \mathbb{P}\left(P_N X - \mu < -\sigma \sqrt{\frac{2t}{N}} \right) \leqslant \frac{e^{-t}}{2} \ .$$

The result follows therefore from a union bound. □

The result on Gaussian distributions naturally extends to any sub-Gaussian distribution, thanks to Hoeffding's inequality. Let $\sigma > 0$. Recall that a random variable X is called σ-sub-Gaussian if, for any $s > 0$,

$$\mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leqslant e^{\sigma^2 s^2/2} \ .$$

A Gaussian random variable with variance σ^2 is σ-sub-Gaussian. Another important are bounded variables as shown by the following result.

Lemma 3 (Hoeffding's Lemma)**.** *If $X \in [a, b]$, then X is $(b-a)/2$-sub-Gaussian.*

Hoeffding's Lemma is proved in Lemma 24 in the following Chapter.

Deviation properties of Sub-Gaussian random variables are easy to get from the Chernoff bound. Let X denote a random variable and, for any s for which it make sense, let $\psi(s) = \log \mathbb{E}[e^{s(X - \mathbb{E}[X])}]$. Chernoff bound is an upper bound on the deviation of the variable X. Let $t > 0$, then, by Markov's inequality, for any $s \geqslant 0$ such that $\psi(s)$ is well defined,

$$\mathbb{P}(X - \mathbb{E}[X] > t) = \mathbb{P}(e^{s(X - \mathbb{E}[X])} > e^{st}) \leqslant e^{-st + \psi(s)} \ . \tag{2.1}$$

When X is σ-sub-Gaussian, $\psi(s) \leqslant s^2\sigma^2/2$, for all $s \geqslant 0$, hence,

$$\mathbb{P}(X - \mathbb{E}[X] > t) \leqslant e^{-st + s^2\sigma^2/2} \ .$$

As this holds for any $s > 0$, one can apply it to $s = t/\sigma^2$ and we obtain

$$\forall t > 0, \qquad \mathbb{P}(X - \mathbb{E}[X] > t) \leqslant e^{-\frac{t^2}{2\sigma^2}} \ . \tag{2.2}$$

The empirical mean of independent sub-Gaussian random variables is sub-Gaussian, as shown by the following inequality.

Lemma 4. *If $X_1, \ldots, X_n$ are independent random variables and if, for any $i \in \{1, \ldots, n\}$, X_i is σ_i-sub-Gaussian, then $n^{-1} \sum_{i=1}^n X_i$ is $n^{-1} \sqrt{\sum_{i=1}^n \sigma_i^2}$-sub-Gaussian.*

Proof. Assume without loss of generality that each $\mathbb{E}[X_i] = 0$. Let $s > 0$, then

$$\mathbb{E}[e^{\frac{s}{n} \sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbb{E}[e^{sX_i}] = e^{\frac{s^2}{2n^2} \sum_{i=1}^n \sigma_i^2} \ .$$

□

Together with (2.2), we obtain the following corollary.

Corollary 5. *If $X_1, \dots, X_n$ are independent random variables and if, for any $i \in \{1, \dots, n\}$, X_i is σ_i-sub-Gaussian, then*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leqslant e^{-\frac{n^2t^2}{2\sum_{i=1}^n \sigma_i^2}} \ .$$

In the particular case of finite valued random variables discussed in Hoeffding's lemma, this corollary yields the standard version of Hoeffding's inequality.

Corollary 6 (Hoeffding's inequality)**.** *If $X_1, \dots, X_n$ are independent random variables and if, for any $i \in \{1, \dots, n\}$, X_i takes values in $[a_i, b_i]$, then*

$$\forall t > 0, \qquad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leqslant e^{-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}} \ .$$

2.1.3 Sub-Gaussian estimators

All along these notes, we build estimators achieving the same deviation rates as the empirical mean in the (sub-)Gaussian case, since these rates are somehow extremal from Proposition 1. Proposition 2 suggests a first definition of "good" estimators of μ_P.

Definition 7 (sub-Gaussian estimator [20])**.** *Let $A \in [0, +\infty]$, $B, C \geqslant 0$. An estimator $\widehat{\mu}$ of μ_P is called (A, B, C)-sub-Gaussian over $\mathcal{P}$ if, for any $P \in \mathcal{P}$,*

$$\forall t \in (0, A), \qquad \mathbb{P}\left(|\widehat{\mu} - \mu_P| > B\sigma_P\sqrt{\frac{1+t}{N}}\right) \leqslant Ce^{-t} \ .$$

Of course, (A, B, C)-sub-Gaussian estimators with $A = +\infty$ are the most desirable. Proposition 2 shows that the empirical mean is $(+\infty, \sqrt{2}, 1)$-sub-Gaussian over the class $\mathcal{P}_{\text{gauss}} = \{\mathrm{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0\}$ and Corollary 5 shows that this result extends to the class of sub-Gaussian distributions. As the empirical mean is satisfying on $\mathcal{P}_{\text{gauss}}$, we may wonder if it is also the case on $\mathcal{P}_2$. The following proposition proves that this is unfortunately not true and that Chebyshev's inequality is sharp in general.

Proposition 8. *[14, Proposition 6.2] For any σ^2 and $t > 0$, there exists a distribution $P \in \mathcal{P}_2$ with variance $\sigma_P^2 = \sigma^2$ (and mean $\mu_P = 0$) such that the empirical mean $\overline{X}_N = N^{-1}\sum_{i=1}^N X_i$ satisfies*

$$\mathbb{P}\left(\overline{X}_N \geqslant t\right) = \mathbb{P}\left(\overline{X}_N \leqslant -t\right) \geqslant \frac{\sigma^2}{2t^2N}\left(1 - \frac{\sigma^2}{t^2N^2}\right)^{N-1} \ .$$

Remark 9. *Proposition 8 implies in particular that, for any value of N and $t > 1/4$, there exists a distribution $P = P_{N,t}$ such that*

$$\mathbb{P}\left(\overline{X}_N - \mu_P > \sigma\sqrt{\frac{2t}{N}}\right) \geqslant \frac{e^{-1/(2t)}}{4t} \geqslant \frac{1}{4e^2t} \ .$$

In words, this means that, for any constants B and C, there exists $A = f(B, C)$ such that the empirical mean is not an (A, B, C) sub-Gaussian estimator over $\mathcal{P}_2$, nor over the subset $\mathcal{P}_2^{\sigma^2}$ of distributions with variance σ^2.

Proof. Consider the distribution taking values in $\{-Nt, 0, Nt\}$ such that

$$\mathbb{P}(Nt) = \mathbb{P}(-Nt) = \frac{1-\mathbb{P}(0)}{2} = \frac{\sigma^2}{2N^2t^2} \ .$$

This distribution is centered with variance σ^2. As this distribution is symmetric,

$$\mathbb{P}\left(\overline{X}_N \geqslant t\right) = \mathbb{P}\left(\overline{X}_N \leqslant -t\right) \ .$$

Moreover,

$$\begin{aligned}\mathbb{P}\left(\overline{X}_N \geqslant t\right) &\geqslant \mathbb{P}\left(\overline{X}_N = t\right) \\ &\geqslant \mathbb{P}\left(\exists! i \in \{1,\dots,N\} : X_i = Nt,\ \forall j \neq i,\ X_j = 0\right) \ .\end{aligned}$$

It is clear that this last event has probability $\frac{\sigma^2}{2t^2N}\left(1-\frac{\sigma^2}{t^2N^2}\right)^{N-1}$ as desired. □

2.2 Level-dependent sub-Gaussian estimators

Proposition 8 implies that, for any choice of B and C, there exists a constant $f(B,C)$ such that, for any $N \geqslant 1$ and any $A \geqslant f(B,C)$, the empirical mean is not an (A,B,C)-sub-Gaussian estimator over $\mathcal{P}_2$. The question is therefore if there exist estimators $\widehat{\mu}$, constants B and C and a sequence $A_N \to \infty$ such that, for any $N \geqslant 1$, $\widehat{\mu}$ is a (A_N,B,C) sub-Gaussian estimator over $\mathcal{P}_2$. The following result shows that, actually, this problem cannot be solved over $\mathcal{P}_2$.

Let $\theta > 0$ and let Po_θ denote the Poisson distribution such that

$$\mathrm{Po}_\theta(k) = \frac{\theta^k}{k!}e^{-\theta}, \qquad \forall k \in \mathbb{Z}_+ \ .$$

For any $\theta > 0$, the expectation and the variance of Po_θ are equal to θ.

Theorem 10. *[20] Assume that $\mathcal{P}$ contains all Poisson's distributions. Then, for any (B,C), there exists $A = f(B,C)$ such that for any $N \geqslant 1$, there does not exist (A,B,C)-sub-Gaussian estimators over $\mathcal{P}$.*

Proof. Let $\widehat{\mu}$, A,B,C such that $\widehat{\mu}$ is (A,B,C)-sub-Gaussian over $\mathcal{P}$. Let $\theta_1 = \sqrt{1/N}$ and $\theta_2 = \sqrt{R/N}$. Let $P_i = \mathrm{Po}_{\theta_i^2}$, $\mathbb{P}_i = P_i^{\otimes N}$ for any $i \in \{1,2\}$. The sub-Gaussian hypothesis on $\widehat{\mu}$ implies that

$$\begin{aligned}\mathbb{P}_1\left(|\widehat{\mu}-\theta_1^2| > B\theta_1\sqrt{\frac{1+t}{N}}\right) &\leqslant Ce^{-t}, \qquad \forall t \leqslant A \ , \\ \mathbb{P}_2\left(|\widehat{\mu}-\theta_2^2| > B\theta_2\sqrt{\frac{1+t}{N}}\right) &\leqslant Ce^{-t}, \qquad \forall t \leqslant A \ .\end{aligned}$$

Denote by $\Omega = \{\sum_{i=1}^N X_i = R\}$. Define $h(x) = (x/2)\log(x/2)$. Applying Stirling's formula as $R \to \infty$ shows that there exists R_0 such that, for any $R \geqslant R_0$,

$$\begin{aligned}\mathbb{P}_1(\Omega) = e^{-1}\frac{1}{R!} &\sim \frac{e^{-R\log R - 2R}}{\sqrt{2\pi R}} \geqslant e^{-h(R)} \ , \\ \mathbb{P}_2(\Omega) = e^{-R}\frac{R^R}{R!} &\sim \frac{1}{\sqrt{2\pi R}} \geqslant \frac{1}{4\sqrt{R}} \ .\end{aligned}$$

We deduce from these estimates that

$$\mathbb{P}_2\left(\widehat{\mu} < \frac{R - B\sqrt{R(1+t)}}{N} | \Omega\right) \leqslant 4C\sqrt{R}e^{-t} \ .$$

If $A \geqslant \log(8C\sqrt{R})$, one can apply this inequality with $t = \log(8C\sqrt{R})$ to get

$$\mathbb{P}_2\left(\widehat{\mu} < \frac{R - B\sqrt{R(1+\log(8C\sqrt{R}))}}{N} | \Omega\right) \leqslant \frac{1}{2} \ .$$

This is equivalent to

$$\mathbb{P}_2\left(\widehat{\mu} \geqslant \frac{R - B\sqrt{R(1+\log(8C\sqrt{R}))}}{N} | \Omega\right) \geqslant \frac{1}{2} \ .$$

Now, we apply the following Poisson's trick

$$\mathcal{D}_1\left(X_1, \ldots, X_N \middle| \sum_{i=1}^N X_i = R\right) = \mathcal{D}_2\left(X_1, \ldots, X_N \middle| \sum_{i=1}^N X_i = R\right) \ .$$

In particular,

$$\mathbb{P}_1\left(\widehat{\mu} \geqslant \frac{R - B\sqrt{R(1+\log(8C\sqrt{R}))}}{N} | \Omega\right) \geqslant \frac{1}{2} \ .$$

Therefore, by the estimate on $\mathbb{P}_2(\Omega)$,

$$\mathbb{P}_1\left(\widehat{\mu} \geqslant \frac{R - B\sqrt{R(1+\log(8C\sqrt{R}))}}{N}\right) \geqslant \frac{e^{-h(R)}}{2} \ .$$

Now, for any R larger than a fixed $R_0(B, C)$,

$$R - B\sqrt{R(1+\log(8C\sqrt{R}))} \geqslant 1 + B\sqrt{1 + R^2/(2B^2)} \ .$$

Pick $R \geqslant R_0(B, C)$ and $A \geqslant R^2/(2B^2)$, the sub-Gaussian property of $\widehat{\mu}$ applied with $t = R^2/(2B^2)$ yields

$$\mathbb{P}_1\left(\widehat{\mu} \geqslant \frac{R - B\sqrt{R(1+\log(8C\sqrt{R}))}}{N}\right) \leqslant Ce^{-R^2/(2B^2)} \ .$$

In other words, $Ce^{-R^2/(2B^2)} \geqslant e^{-h(R)}/2$ which is possible only if $R \leqslant R_1(B, C)$. In conclusion, the result is possible only if $R \leqslant R_0(B, C) \vee R_1(B, C)$ which means that an (A, B, C) sub-Gaussian $\widehat{\mu}$ exists only if $A \leqslant R^2/(2B^2) \vee \log(8C\sqrt{R})$, for some $R \leqslant R_0(B, C) \vee R_1(B, C)$. □

Theorem 10 shows that the notion of sub-Gaussian estimators is a bit too constraining to work on $\mathcal{P}_2$. Hereafter, the following relaxation of the sub-Gaussian property will be used extensively.

Definition 11 (Idea from [14], formally defined in [20])**.** *Let $t \in (0,1)$. A level-dependent estimator $\widehat{\mu}_t$ of μ is a function of the data and t: $\widehat{\mu}_t = F(\mathcal{D}_N, t)$, where F is a measurable map $\mathbb{R}^{N+1} \to \mathbb{R}$.*

The (level-dependent) estimator $\widehat{\mu}_t$ of μ_P is called (B,C)-sub-Gaussian at level t over $\mathcal{P}$ if, for any $P \in \mathcal{P}$,

$$\mathbb{P}\left(|\widehat{\mu}_t - \mu_P| > B\sigma_P\sqrt{\frac{1+t}{N}}\right) \leqslant Ce^{-t} \ .$$

The notion of level dependent estimator may seem surprising at first sight. It is however a key concept in these notes. The first reason is that one can build level-dependent estimators over the class $\mathcal{P}_2$ up to levels $t \asymp N$ as we will see in the following section.

2.3 Median-Of-Means estimators

This section introduces a basic example of level-dependent sub-Gaussian estimators, called median-of-means estimators (MOM). These estimators date back at least from the textbook [48] although they have been around before. For example, a similar construction also appeared independently in [7]. These estimators are used systematically in these notes to build robust extensions of ERM.

Let K and b such that $N = Kb$ and let $B_1, \dots, B_K$ denote a partition of $\{1, \dots, N\}$ into subsets of cardinality b. For any $k \in \{1, \dots, K\}$, let $P_{B_k}X = b^{-1}\sum_{i \in B_k} X_i$. The MOM estimators of μ_P are defined by

$$\mathrm{MOM}_K[X] \in \mathrm{median}\left\{P_{B_k}X,\ k \in \{1, \dots, K\}\right\} \ .$$

The following result shows that $\mathrm{MOM}_K[X]$ is a level dependent estimator over $\mathcal{P}_2$ for a proper choice of $t \asymp K$.

Proposition 12. *For any K, $\epsilon > 0$ and $P \in \mathcal{P}_2$,*

$$\mathbb{P}\left(|MOM_K[X] - \mu_P| > \epsilon\right) \leqslant e^{-2K(1/2 - \sigma_P^2 K/(N\epsilon^2))^2} \ .$$

It follows that, for any $\delta > 0$, choosing $\epsilon = \sigma_P\sqrt{(2+\delta)K/N}$ yields

$$\mathbb{P}\left(|MOM_K[X] - \mu| > \sigma_P\sqrt{(2+\delta)\frac{K}{N}}\right) \leqslant e^{-\frac{\delta^2 K}{2(2+\delta)^2}} \ .$$

Choosing $\delta = 2$ yields

$$\mathbb{P}\left(|MOM_K[X] - \mu| > 2\sigma_P\sqrt{\frac{K}{N}}\right) \leqslant e^{-K/8} \ .$$

$MOM_K[X]$ is a $(4\sqrt{2}, 1)$-sub-Gaussian estimator at level $K/8$.

Proof. Fix $\epsilon > 0$. The first analysis of MOM estimators is based on the remark that, if there are more than $K/2$ blocks B_k such that $|P_{B_k}X - \mu_P| \leqslant \epsilon$, then $|\mathrm{MOM}_K[X] - \mu_P| \leqslant \epsilon$. Formally,

$$\begin{aligned}\left\{|\mathrm{MOM}_K[X] - \mu_P| \leqslant \epsilon\right\} &\supset \left\{|\{k \in \{1, \dots, K\} : |P_{B_k}X - \mu_P| \leqslant \epsilon\}| \geqslant \frac{K}{2}\right\} \\ &= \left\{\sum_{k \in \{1,\dots,K\}} \mathbf{1}_{\{|P_{B_k}X - \mu_P| > \epsilon\}} < \frac{K}{2}\right\} \ .\end{aligned}$$

Denote by $p_\epsilon = \mathbb{P}\left(|P_{B_k} X - \mu_P| > \epsilon\right)$, $Y_k = \mathbf{1}_{\{|P_{B_k} X - \mu_P| > \epsilon\}} - p_\epsilon$. This implies

$$\mathbb{P}\left(|\mathrm{MOM}_K[X] - \mu_P| \leqslant \epsilon\right) \geqslant 1 - \mathbb{P}\left(\sum_{i=1}^{K} Y_k \geqslant K(1/2 - p_\epsilon)\right) \ .$$

As $(Y_k)_{k=1,\dots,K}$ are independent random variables bounded by 1, by Hoeffding's inequality (recalled in (3.13)),

$$\mathbb{P}\left(|\mathrm{MOM}_K[X] - \mu_P| \leqslant \epsilon\right) \geqslant 1 - e^{2(1/2-p_\epsilon)^2 K} \ .$$

The proof is concluded since, by Chebishev's inequality,

$$p_\epsilon \leqslant \frac{\sigma_P^2 K}{N\epsilon^2} \ .$$

□

The previous elementary result shows that median-of-means estimators are level-dependent sub-Gaussian estimators. It is based on a very basic first argument that easily generalises to other frameworks. However, the result can be refined using Gaussian approximation and slightly stronger hypotheses. The following result is due to Minsker and Strawn [46]. It shows that, under slightly stronger assumptions on P, MOM estimators are also sub-Gaussian estimators (not level-dependent). Let

$$\mathcal{P}_3^\gamma = \{P \in \mathcal{P}_2 : P[|X|^3] < \infty \text{ and } \sigma^{-3}\mathbb{E}[|X-\mu|^3] \leqslant \gamma\} \ .$$

Theorem 13. *For any $P \in \mathcal{P}_3^\gamma$ and any $t > 0$ such that $0.5\gamma\sqrt{K/N} + \sqrt{t/2K} \leqslant 1/3$,*

$$\mathbb{P}\left(|MOM_K[X] - \mu_P| > \sigma_P \left(1.5\gamma\frac{K}{N} + 3\sqrt{\frac{t}{2N}}\right)\right) \leqslant 4e^{-t} \ .$$

Remark 14. *As long as $K \leqslant \sqrt{N}$, this result implies that*

$$\forall t \lesssim \sqrt{N}, \qquad \mathbb{P}\left(|MOM_{\sqrt{N}}[X] - \mu_P| > \sigma_P \left((1.5\gamma + 3)\sqrt{\frac{1+t}{2N}}\right)\right) \leqslant 4e^{-t} \ .$$

In other words, the estimator $MOM_{\sqrt{N}}[X]$ is $(O(\sqrt{N}), C(\gamma), 4)$-sub-Gaussian over $\mathcal{P}_3^\gamma$, where $C(\gamma) = (1.5\gamma + 3)/\sqrt{2}$. Theorem 13 is not in contradiction with Theorem 10 since $\mathcal{P}_3^\gamma$ does not contain all Poisson's distributions.

Proof. Denote by

$$Q_K^{(b)}(t) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\left\{\sqrt{b}\frac{P_{B_k} X - \mu}{\sigma_P} > t\right\}} \ .$$

The goal is to find deterministic quantities t_- and t_+ such that, the event Ω_{t_-,t_+} has large probability, where

$$\Omega_{t_-,t_+} = \left\{1 - Q_K^{(b)}(t_+) \geqslant \frac{1}{2}, \quad Q_K^{(b)}(t_-) \geqslant \frac{1}{2}\right\} \ .$$

Indeed, on Ω_{t_-,t_+}, it holds

$$\text{median}\left(\sqrt{b}\frac{P_{B_k}X-\mu}{\sigma_P}, k\in\{1,\dots,K\}\right)\in[t_-,t_+]\ .$$

By homogeneity and translation invariance of the median, this implies that, on Ω_{t_-,t_+},

$$\sigma_P\frac{t_-}{\sqrt{b}}\leqslant \text{MOM}_K[X]-\mu_P\leqslant\sigma_P\frac{t_+}{\sqrt{b}}\ . \tag{2.3}$$

Fix $t\in\mathbb{R}$, to bound $Q_K^{(b)}(t)$, let us first introduce

$$Q^{(b)}(t)=\mathbb{P}\left(\sqrt{b}\frac{P_{B_1}X-\mu}{\sigma}>t\right)\ .$$

By Hoeffding's inequality,

$$\forall x>0,\qquad \mathbb{P}\left(|Q_K^{(b)}(t)-Q^{(b)}(t)|>\sqrt{\frac{x}{2K}}\right)\leqslant 2e^{-x}\ . \tag{2.4}$$

Hence, for any t_-,t_+ in $\mathbb{R}$ such that

$$Q^{(b)}(t_+)\leqslant 1/2-\sqrt{\frac{x}{2K}},\qquad Q^{(b)}(t_-)\geqslant 1/2+\sqrt{\frac{x}{2K}}\ . \tag{2.5}$$

A union bound in (2.4) shows that

$$\mathbb{P}\left(\Omega_{t_-,t_+}\right)\geqslant 1-4e^{-x}\ .$$

Therefore, (2.3) holds for these values of t_-,t_+ with probability at least $1-4e^{-x}$. To evaluate t_-, t_+ in (2.5), introduce now

$$Q(t)=1-\Phi(t)=\int_t^{+\infty}e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}}\ .$$

By Berry-Essen theorem,

$$\left\|Q^{(b)}-Q\right\|_\infty\leqslant 0.5\gamma\sqrt{\frac{K}{N}}\ . \tag{2.6}$$

Therefore, (2.5) is fulfilled if t_- and t_+ satisfy

$$Q(t_+)\leqslant 1/2-\sqrt{\frac{x}{2K}}-0.5\gamma\sqrt{\frac{K}{N}},\qquad Q(t_-)\geqslant 1/2+\sqrt{\frac{x}{2K}}+0.5\gamma\sqrt{\frac{K}{N}}\ .$$

Using the mean value theorem, for any $t\in(0,\sqrt{\log(9/2\pi)})$, $|Q'(t)|=e^{-t^2/2}/\sqrt{2\pi}\geqslant 1/3$, therefore

$$Q(t)\leqslant Q(0)-\frac{t}{3}=\frac{1}{2}-\frac{t}{3}\ .$$

Therefore, (2.5) is fulfilled if

$$t_+=3\sqrt{\frac{x}{2K}}+1.5\gamma\sqrt{\frac{K}{N}},\qquad t_-=-t_+\ .$$

□

Proposition 12 shows that, when $K \geqslant 8t$, $\mathrm{MOM}_K[X]$ is a level-dependent sub-Gaussian estimators at level t. In particular, as K can be equal to N, there exist level-dependent sub-Gaussian estimators at levels t that might be of order N. The following result shows that this rate cannot be improved in general. Let $\lambda \in \mathbb{R}$ and let La_λ denote the Laplace distribution with density

$$f_\lambda(x) = \frac{1}{2} e^{-|x-\lambda|}, \qquad \forall x \in \mathbb{R} \ .$$

This distribution has expectation λ and variance 2.

Proposition 15. *Assume that $\mathcal{P}$ contains all Laplace distributions. Then, for any B and C there exists a constant $f(B,C)$ such that, for any $N \geqslant 1$, there does not exists a (B,C) level-dependent sub-Gaussian estimator at level $t \geqslant f(B,C)N$.*

Proof. Proceed by contradiction and let $\widehat{\mu}_t$ denote such an estimator. Let $P_1 = \mathrm{La}_0$ and $P_2 = \mathrm{La}_\lambda$ and $t = N\lambda^2/(4B^2\sigma_{P_i}^2) = N\lambda^2/(8B^2)$. By the triangular inequality,

$$f_2(x_1, \ldots, x_N) \leqslant e^{\lambda N} f_1(x_1, \ldots, x_N) \ .$$

Therefore

$$\mathbb{P}_2\left(\widehat{\mu}_t > \frac{\lambda}{2}\right) \leqslant e^{\lambda N} \mathbb{P}_1\left(\widehat{\mu}_t > \frac{\lambda}{2}\right) \ .$$

Now $\lambda = P_2 X$, so, by the sub-Gaussian property of $\widehat{\mu}_t$,

$$\begin{aligned}\mathbb{P}_2\left(\widehat{\mu}_t > \frac{\lambda}{2}\right) = \mathbb{P}_2\left(\widehat{\mu}_t - \mu_{P_2} > B\sigma_{P_2}\sqrt{\frac{t}{N}}\right) \\ \geqslant 1 - Ce^{-t} = 1 - C\exp\left(-\frac{N\lambda^2}{8B^2}\right) \ .\end{aligned}$$

Likewise, the sub-Gaussian property of $\widehat{\mu}_t$ yields

$$\mathbb{P}_1\left(\widehat{\mu}_t > \frac{\lambda}{2}\right) \leqslant \mathbb{P}_1\left(|\widehat{\mu}_t - \mu_{P_1}| < B\sigma_{P_1}\sqrt{\frac{t}{N}}\right) \leqslant Ce^{-t} = C\exp\left(-\frac{N\lambda^2}{8B^2}\right) \ .$$

Overall, this yields

$$1 - C\exp\left(-\frac{N\lambda^2}{8B^2}\right) \leqslant C\exp\left(-\frac{N\lambda(1-\lambda)}{8B^2}\right) \ .$$

Whatever the value of $N \geqslant 1$, this relationship is absurd for any $\lambda \geqslant \lambda_0(B,C)$ which implies that the existence of $\widehat{\mu}_t$ is absurd for any $t \geqslant N\lambda_0(B,C)^2/(8B^2)$. □

2.4 M-estimators

This section introduces an alternative to MOM estimators which is extremely popular in robust statistics. These estimators are known as M-estimators. The asymptotic of these estimators is well known and an overview of these results can be found in [30]. Recall that

$$\mu \in \mathrm{argmin}_{\nu\in\mathbb{R}}\, \mathbb{E}[(X-\nu)^2], \quad P_N X \in \mathrm{argmin}_{\nu\in\mathbb{R}} \sum_{i=1}^N (X_i - \nu)^2 \ .$$

The principle of M-estimation is to replace the function $x \mapsto x^2$ in this formulation by another function Ψ and build

$$\widehat{\mu} \in \operatorname{argmin}_{\nu \in \mathbb{R}} \sum_{i=1}^{N} \Psi(X_i - \nu) \ .$$

The most famous example of M-estimator used to estimate μ_P is given by Huber's function

$$\Psi_c(x) = \begin{cases} \frac{x^2}{2} & \text{if} |x| \leqslant c \\ c|x| - \frac{c^2}{2} & \text{if} |x| > c \end{cases} \ .$$

This function is continuously differentiable, with derivative $\psi_c(x) = x\mathbf{1}_{|x| \leqslant c} + c\text{sign}(x)\mathbf{1}_{|x|>c}$, Ψ_c is convex and c-Lipshitz. Huber's estimators interpolate between the empirical mean that would be obtained for $\Psi = x^2$ and the empirical median that would be obtained for $\Psi = |x|$. In this section, we study the Huber estimators defined either by

$$\widehat{\mu}_c \in \operatorname{argmin}_{\nu \in \mathbb{R}} \sum_{i=1}^{N} \Psi_c(X_i - \nu) \tag{2.7}$$

or as a solution of the equation

$$P_N \psi_c(\cdot - \nu) = \frac{1}{N} \sum_{i=1}^{N} \psi_c(X_i - \nu) = 0 \ . \tag{2.8}$$

The formulation (2.8) suggests to analyse these estimators as particular instances of the following larger family of Z-estimators introduced by [14] (although Huber's function does not exactly satisfy Catoni's condition). Let $\psi : \mathbb{R} \to \mathbb{R}$ denote any continuous and non-decreasing function such that

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leqslant \psi(x) \leqslant \log\left(1 + x + \frac{x^2}{2}\right) \ .$$

Let $\alpha > 0$ and define $\widehat{\mu}_\alpha$ as any solution of the equation

$$\sum_{i=1}^{N} \psi[\alpha(X_i - \mu)] = 0 \ . \tag{2.9}$$

The following result establishes the sub-Gaussian behavior of these estimators.

Theorem 16. *Pick $\alpha = \sigma_P^{-1}\sqrt{t/N}$, the estimator $\widehat{\mu}_\alpha$ defined in (2.9) satisfies*

$$\mathbb{P}\left(|\widehat{\mu}_\alpha - \mu| > \sigma_P \sqrt{\frac{2}{1 - 2\epsilon}\frac{t}{N}}\right) \leqslant 2e^{-t} \ ,$$

for any $\epsilon \in (0, 1/2)$ such that

$$\frac{\alpha^2 \sigma_P^2}{2} + \frac{t}{N} = \frac{3t}{2N} \leqslant \epsilon \ . \tag{2.10}$$

Remark 17. *As $N \to \infty$, the constant ϵ can be chosen as small as desired in* (2.10)*, so Catoni's construction shows that almost optimal constant $\sqrt{2}$ can be achieved by t-dependent sub-Gaussian estimators up to levels of order N.*

Besides t, Catoni's estimators are sub-Gaussian if σ_P is known*, that is it can be used on the classes $\mathcal{P}_2^{\sigma^2, L\sigma^2}$ of distributions $P \in \mathcal{P}_2$ with variance $\sigma_P \in [\sigma^2, L\sigma^2]$. It yields optimal constants if $L = 1$. Otherwise, choosing for example $\alpha = \sqrt{t/N}$, Catoni's estimators are* weakly *sub-Gaussian in the sense that the variance σ_P in Definition 7 is replaced by a larger quantity, here $1+\sigma_P^2$.*

Proof. All along the proof, denote, for any $\mu \in \mathbb{R}$, by

$$Z_\alpha(\mu) = \frac{1}{N\alpha} \sum_{i=1}^{N} \psi[\alpha(X_i - \mu)] \ .$$

First, by independence of X_i, for any $s \in \{-1, 1\}$,

$$\mathbb{E}\big[e^{s\alpha N Z_\alpha(\mu)}\big] \leqslant \prod_{i=1}^{N} \mathbb{E}\big[e^{s\psi[\alpha(X_i - \mu)]}\big] \ .$$

Second, the definition of ψ implies that, for any $s \in \{-1, 1\}$,

$$\mathbb{E}\big[e^{s\alpha N Z_\alpha(\mu)}\big] \leqslant \prod_{i=1}^{N} \left(1 + s\alpha(\mu_P - \mu) + \frac{\alpha^2}{2}[\sigma_P^2 + (\mu_P - \mu)^2]\right) \ .$$

By the inequality $1 + x \leqslant e^x$, it follows that, for any $s \in \{-1, 1\}$,

$$\mathbb{E}\big[e^{s\alpha N Z_\alpha(\mu)}\big] \leqslant \exp\left(N\left[s\alpha(\mu_P - \mu) + \frac{\alpha^2}{2}[\sigma_P^2 + (\mu_P - \mu)^2]\right]\right) \ . \tag{2.11}$$

Fix $t > 0$ and, for any $\mu \in \mathbb{R}$, let

$$U_\alpha(\mu, t) = \mu_P - \mu + \frac{\alpha}{2}\big[\sigma_P^2 + (\mu_P - \mu)^2\big] + \frac{t}{N\alpha} \ ,$$
$$L_\alpha(\mu, t) = \mu_P - \mu - \frac{\alpha}{2}\big[\sigma_P^2 + (\mu_P - \mu)^2\big] - \frac{t}{N\alpha} \ .$$

Fix $t > 0$. Then, using the inequality $\mathbb{P}(sZ_\alpha(\mu) > u) \leqslant e^{-u}\mathbb{E}[e^{N\alpha s Z_\alpha(\mu)}]$ respectively with $u = U_\alpha(\mu, t)$, $s = 1$ and $u = L_\alpha(\mu, t)$, $s = -1$ yields

$$\mathbb{P}\left(Z_\alpha(\mu) < U_\alpha(\mu, t)\right) \geqslant 1 - e^{-t}, \qquad \mathbb{P}\left(Z_\alpha(\mu) > L_\alpha(\mu, t)\right) \geqslant 1 - e^{-t} \ . \tag{2.12}$$

By (2.10), the smallest solution μ_+ of the equation $U_\alpha(\mu, t) = 0$ and the largest solution μ_- of $L_\alpha(\mu, t) = 0$ satisfy

$$\mu_+ \leqslant \mu + \frac{1}{\sqrt{1 - 2\epsilon}}\left(\frac{\alpha\sigma_P^2}{2} + \frac{t}{\alpha N}\right) \ ,$$
$$\mu_- \geqslant \mu - \frac{1}{\sqrt{1 - 2\epsilon}}\left(\frac{\alpha\sigma_P^2}{2} + \frac{t}{\alpha N}\right) \ .$$

Consider the event

$$\Omega = \big\{L_\alpha(\mu_-, t) < Z_\alpha(\mu_-),\ Z_\alpha(\mu_+) < U_\alpha(\mu_+, t)\big\} \ .$$

By (2.12), $\mathbb{P}(\Omega) \geqslant 1 - 2e^{-t}$. As the map $\mu \mapsto Z_\alpha(\mu)$ is non-increasing, on Ω, $Z_\alpha(\mu_+) < U_\alpha(\mu_+, t) = 0 = Z_\alpha(\widehat{\mu})$, so $\widehat{\mu} \leqslant \mu_+$. Likewise $\widehat{\mu} \geqslant \mu_-$. It follows that

$$\mathbb{P}\left(\mu_- < \widehat{\mu} < \mu_+\right) \geqslant \mathbb{P}(\Omega) \geqslant 1 - 2e^{-t} \ .$$

This concludes the proof. □

2.5 Level free sub-Gaussian estimators

Theorem 13 showed that $\mathrm{MOM}_{\sqrt{N}}[X]$ is a level free $(A_N, B(\gamma), C)$-sub-Gaussian estimator over $\mathcal{P}_3^\gamma$ with A_N of order $\sqrt{N}$. The purpose here is to present a method to derive level free estimators from level dependent ones, provided, for example that informations on the variance are available. The central tool is due to Lepski.

Theorem 18. *Assume that, for any K in a finite set $\mathcal{K}$, there exists a confidence interval $\widehat{I}_K$ such that*

(i) for any K and K' in $\mathcal{K}$ such that $K \leqslant K'$, $|\widehat{I}_K| \leqslant |\widehat{I}_{K'}|$,

(ii) $\mathbb{P}\left(\mu \in \widehat{I}_K\right) \geqslant 1 - \alpha_K$.

Then, if one defines

$$\widehat{K} = \min\left\{K \in \mathcal{K} : \cap_{J\in\mathcal{K}, J\geqslant K} \widehat{I}_J \neq \emptyset\right\}, \qquad \widehat{\mu} \in \widehat{I}_{\widehat{K}} \ ,$$

we have

$$\forall K \in \mathcal{K}, \qquad \mathbb{P}\left(|\widehat{\mu} - \mu| > 2|\widehat{I}_K|\right) \leqslant \sum_{J\in\mathcal{K}, J\geqslant K} \alpha_J \ .$$

Proof. For any $K \in \mathcal{K}$, denote by $\mathcal{K}_K = \{J \in \mathcal{K} : J \geqslant K\}$. Fix $K \in \mathcal{K}$ and consider the event $\Omega = \{\mu \in \cap_{J\in\mathcal{K}_K} \widehat{I}_J\}$. A union bound grants that

$$\mathbb{P}(\Omega) \geqslant 1 - \sum_{J\in\mathcal{K}_K} \alpha_J \ .$$

On Ω, $\cap_{J\in\mathcal{K}_K} \widehat{I}_J \neq \emptyset$, therefore, $\widehat{K} \leqslant K$ and there exists $\mu_0 \in \cap_{J\in\mathcal{K}_{\widehat{K}}} \widehat{I}_J$. As $\mu_0, \widehat{\mu} \in \widehat{I}_{\widehat{K}}$, $|\mu_0 - \mu| \leqslant |\widehat{I}_{\widehat{K}}|$ and as $\widehat{K} \leqslant K$, $|\widehat{I}_{\widehat{K}}| \leqslant |\widehat{I}_K|$, so $|\mu_0 - \widehat{\mu}| \leqslant |\widehat{I}_K|$. Moreover, as $\widehat{K} \leqslant K$ and $\mu_0 \in \cap_{J\in\mathcal{K}_{\widehat{K}}} \widehat{I}_J$, $\mu_0 \in \cap_{J\in\mathcal{K}_K} \widehat{I}_J$ and as $\mu \in \cap_{J\in\mathcal{K}_K} \widehat{I}_J$, $|\mu_0 - \mu| \leqslant |\widehat{I}_K|$. Hence,

$$|\widehat{\mu} - \mu| \leqslant |\widehat{\mu} - \mu_0| + |\mu_0 - \mu| \leqslant 2|\widehat{I}_K| \ .$$

□

We are now in position to prove the result.

Theorem 19. *For any $\sigma^2 > 0$ and $L \geqslant 1$, there exists an $((N/2-1)/8, 8\sqrt{2L}, 9)$-sub-Gaussian estimator on $\mathcal{P}_2^{[\sigma^2, L\sigma^2]}$.*

Proof. For any $K = 1, \ldots, N/2$, let $b = \lfloor N/K \rfloor$ and let $\mathrm{MOM}_K[X]$ denote the MOM estimators based on $X_1, \ldots, X_{bK}$. Define, for any $K \in \{1, \ldots, N/2\}$, the intervals

$$\widehat{I}_K = \left[\mathrm{MOM}_K[X] \pm 2\sigma\sqrt{L\frac{K}{N}}\right] \supset \left[\mathrm{MOM}_K[X] \pm 2\sigma_P\sqrt{\frac{K}{N}}\right] \ .$$

Proposition 12 shows that the intervals $\widehat{I}_K$ satisfy Condition (i) of Theorem 18 with $|\widehat{I}_K| = 2\sigma\sqrt{LK/N}$ and Condition (ii) with $\alpha_K = e^{-K/8}$. It follows that, if

$$\widehat{K} = \min\left\{K \in \{1, \ldots, N/2\} : \cap_{J=K}^{N/2} \widehat{I}_J \neq \emptyset\right\}, \qquad \widehat{\mu} = \mathrm{MOM}_{\widehat{K}}[X] \ ,$$

the estimator $\widehat{\mu}$ satisfies, for any $K \in \{1, \ldots . N/2\}$,

$$\mathbb{P}\left(|\widehat{\mu} - \mu_P| > 4\sigma \sqrt{\frac{LK}{N}} \right) \leqslant \sum_{J=K}^{+\infty} e^{-J/8} \leqslant \frac{e^{-K/8}}{1 - e^{-1/8}} \ .$$

Fix $x \in (0, (N/2-1)/8)$ and choose $K = \lfloor 8x \rfloor + 1$. It follows from this result that

$$\mathbb{P}\left(|\widehat{\mu} - \mu_P| > 8\sigma_P \sqrt{\frac{2L(1+x)}{N}} \right) \leqslant 9e^{-x} \ .$$

□

Chapter 3

Concentration/deviation inequalities

Given a random variable Z of interest, concentration inequalities are upper bounds on the probabilities $\mathbb{P}(Z - \mathbb{E}[Z] > t)$ for any $t > 0$, deviation inequalities are upper bounds on the probabilities $\mathbb{P}(Z - D > t)$ for some deterministic quantity D and at least some $t > 0$. Deviation inequalities is therefore a more generic term that refers to less precise results than concentration inequalities. Both are natural tools to show deviation properties of estimators. Deviation inequalities have been widely used in statistics since the 1990's and their introduction for model selection by Birgé and Massart [9]. This chapter presents useful deviation inequalities for the following chapters. We briefly present the powerful entropy method of Ledoux for concentration inequalities and recall sufficient results to establish Bousquet's version of Talagrand's concentration inequality for suprema of empirical processes. All the material of Sections 3.1 and 3.2 is borrowed from [10] that the interested reader is invited to read to learn much more on concentration inequalities. Section 4.4 presents a PAC-Bayesian inequality that will be used to analyse M-estimators for multivariate mean estimation. This result is borrowed from [12] where PAC-Bayesian approaches are developed in various other learning problems. Finally, Section 3.5 presents the most useful result in these notes, which is a deviation result for suprema of MOM processes. This result is obtained using the small ball approach, following arguments originally introduced in [39].

All along the chapter, $X = (X_1, \dots, X_N)$ denotes a vector of independent random variables taking values in a measurable space $\mathcal{X}$. For any $i \in \{1, \dots, N\}$, $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$ and $\mathbb{E}^{(i)}$ denote expectation conditionally on $X^{(i)}$. The function $\Phi : x \mapsto x \log(x)$ for any $x > 0$ is extended by continuity $\Phi(0) = 0$. For any positive random variable Y such that $\mathbb{E}[\Phi(Y)] < \infty$, the entropy of Y is defined by $\mathrm{Ent}(Y) = \mathbb{E}[\Phi(Y)] - \Phi(\mathbb{E}[Y])$. The conditional entropies are defined, for any $i \in \{1, \dots, N\}$ by $\mathrm{Ent}^{(i)}(Y) = \mathbb{E}^{(i)}[\Phi(Y)] - \Phi(\mathbb{E}^{(i)}[Y])$. f denotes a measurable map $\mathcal{X}^n \to [0, +\infty)$ and $Z = f(X) = f(X_1, \dots, X_N)$.

3.1 The entropy method

The entropy method is a series of steps introduced by Ledoux [36] that allows to establish concentration inequalities for $Z = f(X)$ around its expectation $\mathbb{E}[Z]$. The starting point is the Chernoff bound. Assume that $Z \leqslant 1$ so, for any $s > 0$, the log Laplace-transform $\psi(s) = \log(\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}])$ is well defined. For any $t > 0$, by Markov's inequality, it holds that

$$\forall s > 0, \qquad \mathbb{P}(Z - \mathbb{E}[Z] > t) = \mathbb{P}(e^{s(Z-\mathbb{E}[Z])} > e^{st}) \leqslant e^{-st+\psi(s)} \ .$$

Introduce the Fenchel-Legendre transform of Z, $\psi^*(t) = \sup_{s>0}\{st - \psi(s)\}$. Optimizing over $s > 0$ in the previous bound shows the Chernoff bound

$$\forall t > 0, \qquad \mathbb{P}(Z - \mathbb{E}[Z] > t) \leqslant e^{-\psi^*(t)} \ .$$

Chernoff's bound shows that one can bound the deviation probabilities of $Z - \mathbb{E}[Z]$ by bounding from bellow the Fenchel-Legendre transform $\psi^*(t)$ of Z, which can be done by bounding from above the log-Laplace transform $\psi(s)$ of Z. As important examples, basic analysis allows to check the following result.

Lemma 20. *Let $\sigma > 0$. The random variable Z is called σ-sub-Gaussian if $\psi(s) \leqslant s^2\sigma^2/2$. If Z is σ-sub-Gaussian, $\psi^*(t) \geqslant t^2/(2\sigma^2)$. In particular,*

$$\forall t > 0, \qquad \mathbb{P}(Z - \mathbb{E}[Z] > t) \leqslant e^{-t^2/(2\sigma^2)} \ .$$

Let $\nu > 0$, $\phi(s) = e^s - 1 - s$ and $h(t) = (1+t)\log(1+t) - t$. The random variable Z is called ν-sub-Poissonian if $\psi(s) \leqslant \nu\phi(s)$. If Z is ν-sub-Poissonian, $\phi^*(t) \geqslant \nu h(t/\nu)$. In particular,

$$\forall t > 0, \qquad \mathbb{P}(Z - \mathbb{E}[Z] > t) \leqslant e^{-\nu h(t/\nu)} \ .$$

Thanks to Chernoff's bound, concentration inequalities follow from upper bounds on $\psi(s)$. The idea of the entropy method is to obtain these bound by bounding from above the entropy of $e^{s(Z-\mathbb{E}[Z])}$. The method can be summarized in the following lemma.

Lemma 21. *The entropy satisfies*

$$\mathrm{Ent}(e^{sZ}) = \mathbb{E}[e^{sZ}](s\psi'(s) - \psi(s)) \ . \tag{3.1}$$

Therefore, if there exists a function g such that

$$\mathrm{Ent}(e^{sZ}) \leqslant g(s)\mathbb{E}[e^{sZ}] \ , \tag{3.2}$$

then, the log-Laplace transform of Z satisfies

$$\psi(s) \leqslant s\int_0^s \frac{g(t)}{t^2}\mathrm{d}t \ . \tag{3.3}$$

For example, if (3.2) holds with $g(s) = \sigma^2 s^2/2$, then $\psi(s) \leqslant s^2\sigma^2/2$, so Z is σ-sub-Gaussian.

Proof. Notice that $\mathrm{Ent}(e^{s(Z-\mathbb{E}[Z])}) = \mathbb{E}[e^{-s\mathbb{E}[Z]}]\,\mathrm{Ent}(e^{sZ})$. Thus, Equation (3.1) is equivalent to

$$\mathrm{Ent}(e^{s(Z-\mathbb{E}[Z])}) = \mathbb{E}[e^{s(Z-\mathbb{E}[Z])}](s\psi'(s) - \psi(s)) \ .$$

As

$$\psi'(s) = \frac{\mathbb{E}\big[(Z-\mathbb{E}[Z])e^{s(Z-\mathbb{E}[Z])}\big]}{\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]} \ ,$$

we have

$$\begin{aligned}
\mathrm{Ent}(e^{s(Z-\mathbb{E}[Z])}) &= \mathbb{E}[e^{s(Z-\mathbb{E}[Z])}\log(e^{s(Z-\mathbb{E}[Z])})] - \mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]\log(\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]) \\
&= \mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]\left(s\frac{\mathbb{E}\big[(Z-\mathbb{E}[Z])e^{s(Z-\mathbb{E}[Z])}\big]}{\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}]} - \psi(s)\right) \\
&= \mathbb{E}[e^{s(Z-\mathbb{E}[Z])}](s\psi'(s) - \psi(s)) \ .
\end{aligned}$$

This shows Equation (3.1).

The entropy condition (3.2) is equivalent to

$$\mathrm{Ent}(e^{s(Z-\mathbb{E}[Z])}) \leqslant g(s)\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}] \ . \tag{3.4}$$

Under this condition, the function ψ satisfies the following differential inequality

$$s\psi'(s) - \psi(s) \leqslant g(s) \ .$$

Dividing by s^2 on both sides shows that

$$\left(\frac{\psi(s)}{s}\right)' \leqslant \frac{g(s)}{s^2} \ .$$

As $\psi(0) = \psi'(0) = 0$, the function $u : s \mapsto \psi(s)/s$ can be extended continuously in 0 by defining $u(0) = 0$ and the previous inequality implies that

$$\psi(s) \leqslant s\int_0^s \frac{g(t)}{t^2}\mathrm{d}t \ .$$

□

The Entropy lemma is well known in the sub-Gaussian case where it is referred to as Herbst's argument, which is both simple and elegant while surprisingly powerful. The entropy method (Lemma 21) shows that bounding the entropy from above *can be useful*. The success of the method comes from the fact that *it is actually possible* to obtain such upper bounds. An important reason is the sub-additivity property of the entropy which allows to bound the entropy of functions depending on only one variable X_i. This property is shown in the following section.

3.1.1 Sub-additivity of the entropy

To prove the sub-additivity property, we need a first variational formula for the entropy.

Theorem 22 (Duality formula of entropy)**.** *Let Y denote a positive random variable such that $\mathbb{E}[\Phi(Y)] < \infty$ and let $\mathcal{U}$ denote the set of real valued random variables U such that $\mathbb{E}[e^U] = 1$. Then*

$$\mathrm{Ent}(Y) = \sup_{U \in \mathcal{U}} \mathbb{E}[UY] \ . \tag{3.5}$$

Equivalently, let $\mathcal{T}$ denote the set of non negative and integrable random variables, then

$$\mathrm{Ent}(Y) = \sup_{T \in \mathcal{T}} \mathbb{E}[Y(\log(T) - \log(\mathbb{E}[T]))] \ .$$

Proof. The second part being a direct consequence of the first one, it is sufficient to show the first part. Let $U \in \mathcal{U}$, then

$$\begin{aligned}\mathrm{Ent}(Y) - \mathbb{E}[UY] &= \mathbb{E}[Ye^{-U}\log(Ye^{-U})e^U] - \mathbb{E}[Ye^{-U}e^U]\log(\mathbb{E}[Ye^{-U}e^U]) \\ &= \mathbb{E}[\Phi(Ye^{-U})e^U] - \Phi(\mathbb{E}[Ye^{-U}e^U]) \ .\end{aligned}$$

If P' denotes the measure such that $P'(\mathrm{d}u) = e^u P(\mathrm{d}u)$ (note that this is a probability measure), then $\mathrm{Ent}(Y) - \mathbb{E}[UY]$ is the entropy of Ye^{-U} with respect to the measure P'. Hence, $\mathrm{Ent}(Y) - \mathbb{E}[UY] \geqslant 0$, so the right-hand side of (3.5) is smaller than the left-hand side.

Conversely, if $U = \log(Y) - \log(\mathbb{E}[Y])$, then $\mathbb{E}[e^U] = 1$ so $U \in \mathcal{U}$ and $\mathbb{E}[UY] = \mathrm{Ent}(Y)$. This proves the second inequality in (3.5) and therefore the theorem. □

The Duality formula is used to prove the sub-additivity property. The idea is to bound the entropy $\mathrm{Ent}(Z)$ of any function $Z = f(X)$ by the entropies of "simpler" functions depending on a single variable X_i only. Recall that $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$, $\mathbb{E}^{(i)} = \mathbb{E}[\cdot|X^{(i)}]$ and $\mathrm{Ent}^{(i)}(Z) = \mathbb{E}^{(i)}[\Phi(Z)] - \Phi(\mathbb{E}^{(i)}[Z])$. By conditioning on $X^{(i)}$, $\mathrm{Ent}^{(i)}(Z)$ is therefore the entropy of Z with respect to X_i only, while $X^{(i)}$ is left fixed. The sub-additivity property bounds the entropy $\mathrm{Ent}(Z)$ from above using the simpler entropies $\mathrm{Ent}^{(i)}(Z)$.

Theorem 23. *[Sub-additivity of entropy] If $Z > 0$, then*

$$\mathrm{Ent}(Z) \leqslant \mathbb{E}\left[\sum_{i=1}^N \mathrm{Ent}^{(i)}(Z)\right] \ .$$

Proof. The proof relies on the martingale method. Introduce $\mathbb{E}_i = \mathbb{E}[\cdot|X_1, \ldots, X_i]$, $\mathbb{E}_0 = \mathbb{E}$. As $\mathbb{E}_N[Z] = Z$, one can decompose $\log(Z) - \log(\mathbb{E}[Z])$ into the following sum of martingale increments

$$\log(Z) - \log(\mathbb{E}[Z]) = \sum_{i=1}^N (\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}_{i-1}[Z])) \ .$$

It yields

$$Z(\log(Z) - \log(\mathbb{E}[Z])) = \sum_{i=1}^N Z(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}_{i-1}[Z])) \ . \tag{3.6}$$

Now by independence of X_i and $X_1, \ldots, X_{i-1}$,

$$\mathbb{E}^{(i)}[\mathbb{E}_i[Z]] = \mathbb{E}_{i-1}[Z] \ . \tag{3.7}$$

Plugging (3.7) into (3.6) yields

$$Z(\log(Z) - \log(\mathbb{E}[Z])) = \sum_{i=1}^{N} Z(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}^{(i)}[\mathbb{E}_i[Z]])) \ . \tag{3.8}$$

The second duality formula in Theorem 22 applied conditionally on $X^{(i)}$ with $Y = Z$ and $T = \mathbb{E}_i[Z]$ implies that

$$\mathbb{E}^{(i)}[Z(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}^{(i)}[\mathbb{E}_i[Z]]))] \leqslant \mathrm{Ent}^{(i)}(Z) \ . \tag{3.9}$$

Therefore, taking expectation in (3.8) and using (3.9) shows that

$$\begin{aligned}\mathrm{Ent}(Y) &= \mathbb{E}\left[\sum_{i=1}^{N} \mathbb{E}^{(i)}[Z(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}^{(i)}[\mathbb{E}_i[Z]]))]\right] \\ &\leqslant \mathbb{E}\left[\sum_{i=1}^{N} \mathrm{Ent}^{(i)}(Z)\right] \ .\end{aligned}$$

This proves the theorem. □

3.1.2 Bounded difference inequality

The entropy method shows that concentration derives from upper bounds on the entropy. Sub-additivity of the entropy shows that it is sufficient to bound the entropy of functions depending on one of the X_i only. One can bound the entropy of a function of one X_i only if the function takes value in a compact space. This is the purpose of Hoeffding's lemma.

Lemma 24 (Hoeffding's lemma)**.** *Let X_0 denote a random variable taking values in $[a, b]$ and let $\psi(s) = \log \mathbb{E}[e^{s(X_0 - \mathbb{E}[X_0])}]$. Then*

$$\psi(s) \leqslant \frac{s^2(b-a)^2}{8}, \qquad \mathrm{Ent}(e^{sX_0}) \leqslant \frac{s^2(b-a)^2}{8}\mathbb{E}[e^{sX_0}] \ .$$

Proof. Assume, without loss of generality, that $\mathbb{E}[X_0] = 0$. Check that $\psi(0) = \psi'(0) = 0$ and that

$$\psi''(s) = \mathbb{E}\left[X_0^2 \frac{e^{sX_0}}{\mathbb{E}[e^{sX_0}]}\right] - \left(\mathbb{E}\left[X_0 \frac{e^{sX_0}}{\mathbb{E}[e^{sX_0}]}\right]\right)^2 \ . \tag{3.10}$$

As $e^{sX_0}/\mathbb{E}[e^{sX_0}]$ is non-negative with expectation with respect to the measure $\mathbb{E}$ equal 1, one can consider the measure $\mathbb{F}$ such that

$$\frac{\mathrm{d}\mathbb{F}}{\mathrm{d}\mathbb{E}}(x) = \frac{e^{sx}}{\mathbb{E}[e^{sX_0}]} \ .$$

Equation (3.10) shows that

$$\psi''(s) = \mathrm{Var}_{\mathbb{F}}(X_0) = \mathrm{Var}_{\mathbb{F}}\left(X_0 - \frac{a+b}{2}\right) \ .$$

As X_0 takes value in $[a, b]$ $\mathbb{F}$-a.s., $|X_0 - (a+b)/2| \leqslant (b-a)/2$ $\mathbb{F}$-a.s. so

$$\mathrm{Var}_{\mathbb{F}}\left(X_0 - \frac{a+b}{2}\right) \leqslant \frac{(b-a)^2}{4} \ .$$

Integrating twice shows that

$$\psi(s) = \psi(s) - \psi(0) = \int_0^s \psi'(t)\mathrm{d}t = \int_0^s (\psi'(t) - \psi'(0))\mathrm{d}t = \int_0^s \int_0^t \psi''(u)\mathrm{d}u\mathrm{d}t$$
$$\leqslant \int_0^s \int_0^t \frac{(b-a)^2}{4}\mathrm{d}u\mathrm{d}t = \int_0^s \frac{(b-a)^2}{4}t\mathrm{d}t = \frac{(b-a)^2 s^2}{8} \ .$$

For the second inequality, note that

$$s\psi'(s) - \psi(s) = \int_0^s u\psi''(u)\mathrm{d}u \leqslant \frac{s^2(b-a)^2}{8} \ .$$

Plugging this bound into (3.1) gives the second inequality. □

The association of the sub-additivity of entropy with Hoeffding's lemma is useful when the functions $x_i \mapsto f(x)$ have bounded range. This property of the function is known as the bounded difference property of f.

Definition 25 (Bounded difference property)**.** *Let $\boldsymbol{c} = (c_1, \ldots, c_N)$ denote a vector of positive real numbers. The set $\mathcal{B}(\boldsymbol{c})$ is the set of functions $f : \mathcal{X}^N \to \mathbb{R}$ such that, for any $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ in $\mathcal{X}^N$,*

$$|f(x) - f(y)| \leqslant \sum_{i=1}^N c_i \mathbf{1}_{\{x_i \neq y_i\}} \ .$$

The bounded difference property is a Lipschitz property of f with respect to the Hamming distance. It implies that the functions $x_i \mapsto f(x)$ have range with length at most c_i. The bounded difference inequality provides the concentration inequality satisfied by $f(X)$ when f has bounded differences.

Theorem 26 (Bounded Difference Inequality, BDI)**.** *Assume that $\boldsymbol{c} \in \mathbb{R}_+^N$, $f \in \mathcal{B}(\boldsymbol{c})$ and let $\sigma^2 = \|\boldsymbol{c}\|^2/4$. Z is σ-sub-Gaussian, in particular,*

$$\forall t > 0, \qquad \mathbb{P}(Z - \mathbb{E}[Z] > t) \leqslant e^{-t^2/(2\sigma^2)} \ .$$

Proof. By sub-additivity of the entropy,

$$\mathrm{Ent}(e^{sZ}) \leqslant \mathbb{E}\left[\sum_{i=1}^N \mathrm{Ent}^{(i)}(e^{sZ})\right] \ .$$

As $f \in \mathcal{B}(\mathbf{c})$, conditionally on $X^{(i)}$, Z belongs to a set with range at most c_i. By the second part of Hoeffding's lemma,

$$\mathrm{Ent}^{(i)}(e^{sZ}) \leqslant \frac{s^2 c_i^2}{8}\mathbb{E}^{(i)}[e^{sZ}] \ .$$

Summing up over $i \in \{1, \ldots, N\}$ and taking expectation yields

$$\mathrm{Ent}(e^{sZ}) \leqslant \mathbb{E}\left[\sum_{i=1}^N \mathrm{Ent}^{(i)}(e^{sZ})\right] \leqslant \mathbb{E}\left[\sum_{i=1}^N \frac{s^2 c_i^2}{8}\mathbb{E}^{(i)}[e^{sZ}]\right]$$
$$= \sum_{i=1}^N \frac{s^2 c_i^2}{8}\mathbb{E}[e^{sZ}] = \frac{s^2\sigma^2}{2}\mathbb{E}[e^{sZ}] \ .$$

Herbst's argument, see Lemma 21 in the sub-Gaussian case, concludes the proof. □

The bounded difference inequality is of particular interest when f is the supremum of bounded empirical processes. Let $X = (X_1, \dots, X_N)$ denote independent $\mathcal{X}$-valued random variables, each $x_i \in \mathcal{X}$ being a vector $x_i = (x_{i,t})_{t\in T}$. Assume that

$$\forall t \in T, \qquad \mathbb{E}[X_{i,t}] = 0, \quad \text{and} \quad X_{i,t} \in [a_i, b_i] \ .$$

For any $x = (x_1, \dots, x_N) \in \mathcal{X}^N$, let

$$f(x) = \sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} x_{i,t} \ .$$

It is clear that $f \in \mathcal{B}(\mathbf{c})$, with $c_i = (b_i - a_i)/N$, therefore, the BDI applies to f and yields the following concentration inequality for suprema of empirical processes.

$$\forall u > 0, \qquad \mathbb{P}\bigg(\sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} X_{i,t} > \mathbb{E}\big[\sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} X_{i,t} \big] + u \bigg) \leqslant e^{-\frac{2N^2u^2}{\sum_{i=1}^N (b_i-a_i)^2}} \ . \tag{3.11}$$

In particular, if each $X_{i,t} \in [a_i, a_i + 1]$,

$$\forall u > 0, \qquad \mathbb{P}\bigg(\sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} X_{i,t} > \mathbb{E}\big[\sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} X_{i,t} \big] + u \bigg) \leqslant e^{-2Nu^2} \ , \tag{3.12}$$

or equivalently

$$\forall u > 0, \qquad \mathbb{P}\bigg(\sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} X_{i,t} > \mathbb{E}\big[\sup_{t\in T} \frac{1}{N} \sum_{i=1}^{N} X_{i,t} \big] + \sqrt{\frac{u}{2N}} \bigg) \leqslant e^{-u} \ .$$

Another classical application of (3.11) is when T is reduced to a singleton. In that case, the result, known as Hoeffding's inequality, see Corollary 6, states that, if $X_1, \dots, X_N$ are independent random variables taking values respectively in $[a_i, b_i]$, then

$$\forall u > 0, \qquad \mathbb{P}\bigg(\frac{1}{N} \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) > u \bigg) \leqslant e^{-\frac{2N^2u^2}{\sum_{i=1}^N (b_i-a_i)^2}} \ . \tag{3.13}$$

3.1.3 Gaussian concentration inequality

The second application of the entropy method is the Gaussian concentration inequality, whose proof also uses Herbst's argument but coupled with the Gaussian logarithmic Sobolev inequality. These inequalities bound the entropy of $f^2(X)$ for regular functions f by some variance-like term. To establish this result, start with the basic log-Sobolev inequality for Rademacher random variables.

Theorem 27 (Rademacher logarithmic Sobolev inequality)**.** *Let X denote a vector of independent Rademacher random variables. For any $i \in \{1, \dots, N\}$, let $\bar{X}^{(i)} = (X_1, \dots, X_{i-1}, -X_i, X_{i+1}, \dots, X_N)$ and*

$$\mathcal{E}(f) = \frac{1}{2}\mathbb{E}\bigg[\sum_{i=1}^{N} (f(X) - f(\bar{X}^{(i)}))^2 \bigg] \ .$$

Then

$$\mathrm{Ent}(f^2(X)) \leqslant \mathcal{E}(f) \ .$$

Proof. By sub-additivity of the entropy,

$$\mathrm{Ent}(f^2(X)) \leqslant \mathbb{E}\Big[\sum_{i=1}^{N} \mathrm{Ent}^{(i)}(f^2(X))\Big] \ .$$

Hence, it is sufficient to show that

$$\mathrm{Ent}^{(i)}(f^2(X)) \leqslant \frac{1}{2}(f(X) - f(\bar{X}^{(i)}))^2 \ .$$

Given $X^{(i)}$, $f(X)$ can take two values, say a and b, each with probability $1/2$, so it is sufficient to show that, for any a, b,

$$a^2 \log a^2 + b^2 \log b^2 - (a^2 + b^2) \log\left(\frac{a^2 + b^2}{2}\right) \leqslant (a - b)^2 \ .$$

We may assume without loss of generality that a and b are non-negative and that $a > b$. Therefore, if

$$h(a) = a^2 \log a^2 + b^2 \log b^2 - (a^2 + b^2) \log\left(\frac{a^2 + b^2}{2}\right) - (a - b)^2 \ ,$$

it is sufficient to show that $h(b) = h'(b) = 0$, which is obvious and that h is concave, which follows from basic calculus. □

The Rademacher log-Sobolev inequality is sufficient to derive the Gaussian log-Sobolev inequality. This is the main tool to prove the Gaussian concentration inequality.

Theorem 28 (Gaussian log-Sobolev inequality)**.** *Let* $X \sim \mathrm{N}(0, I_N)$ *and* $f : \mathbb{R}^N \to \mathbb{R}$ *be continuously differentiable, then*

$$\mathrm{Ent}(f^2) \leqslant 2\mathbb{E}\big[\|\nabla f(X)\|^2\big] \ .$$

Proof. Assume first that $N = 1$. If $\mathbb{E}[f'(X)^2] = \infty$, the result is trivial so we can assume that $\mathbb{E}[f'(X)^2] < \infty$. By standard density arguments, one can assume furthermore that f is twice continuously differentiable with bounded support. Under this assumption, let K denote the sup-norm of f''. Let $\varepsilon_1, \ldots, \varepsilon_n$ denote i.i.d. Rademacher random variables. Define $S_n = \sum_{i=1}^n \varepsilon_i/\sqrt{n}$. By the Rademacher logarithmic Sobolev inequality,

$$\mathrm{Ent}(f^2(S_n)) \leqslant \frac{1}{2}\sum_{j=1}^{n}\left(f(S_n) - f\big(S_n - \frac{2\varepsilon_j}{\sqrt{n}}\big)\right)^2 \ . \tag{3.14}$$

As f is uniformly bounded and continuous, by the central limit theorem, the left-hand side in (3.14) satisfies

$$\lim_{n\to\infty} \mathrm{Ent}(f^2(S_n)) = \mathrm{Ent}(f^2(X)) \ .$$

On the other hand, for any $j \in \{1, \dots, n\}$, by a Taylor expansion,

$$|f(S_n - 2\varepsilon_j/\sqrt{n}) - f(S_n)| \leqslant \frac{2}{\sqrt{n}}|f'(S_n)| + \frac{2K}{n} \ .$$

Thus,

$$\frac{1}{4}\sum_{j=1}^{n}\left(f\big(S_n - \frac{2\varepsilon_j}{\sqrt{n}}\big) - f(S_n)\right)^2 \leqslant f'(S_n)^2 + \frac{2K}{\sqrt{n}}|f'(S_n)| + \frac{K^2}{n} \ .$$

By the central limit theorem, it follows that

$$\limsup_{n\to\infty} \frac{1}{4}\sum_{j=1}^{n}\left(f\big(S_n - \frac{2\varepsilon_j}{\sqrt{n}}\big) - f(S_n)\right)^2 \leqslant \mathbb{E}[f'(X)^2] \ .$$

Hence, the result for $N = 1$ follows by taking limits in (3.14). To extend the results in dimension $N \geqslant 1$, apply sub-additivity of entropy to get

$$\mathrm{Ent}(f^2(X)) \leqslant \mathbb{E}\Big[\sum_{i=1}^{N} \mathrm{Ent}^{(i)}(f^2(X))\Big] \ .$$

The result for $N = 1$ shows that

$$\mathrm{Ent}^{(i)}(f^2(X)) \leqslant 2\mathbb{E}^{(i)}[(\partial_i f(X))^2] \ .$$

Hence, $\mathrm{Ent}(f^2(X)) \leqslant 2\mathbb{E}[\sum_{i=1}^{N}(\partial_i f(X))^2]$ and the proof is concluded since $\|\nabla f(X)\|^2 = \sum_{i=1}^{N}(\partial_i f(X))^2$. □

Together with Herbst's argument, the Gaussian log-Sobolev inequality shows the Gaussian concentration inequality which is the main result of this section.

Theorem 29 (Borel's Gaussian concentration inequality)**.** *Assume that f is L-Lipschitz, that is $|f(x) - f(y)| \leqslant L\|x - y\|$ for any x and y in $\mathbb{R}^N$. Then $Z = f(X)$ is L-sub-Gaussian, that is, for any $s \in \mathbb{R}$,*

$$\log(\mathbb{E}[e^{s(f(X)-\mathbb{E}[f(X)])}]) \leqslant \frac{s^2L^2}{2} \ .$$

In particular,

$$\forall u > 0, \qquad \mathbb{P}\big(f(X) - \mathbb{E}[f(X)] > u\big) \leqslant e^{-u^2/(2L^2)} \ .$$

Proof. Using standard density argument, one may assume that f is differentiable with gradient bounded by L and that $\mathbb{E}[f(X)] = 0$. The Gaussian log-Sobolev inequality applied with $f = e^{sf/2}$ shows that

$$\mathrm{Ent}(e^{sf}) \leqslant 2\mathbb{E}[\|\nabla e^{sf(X)/2}\|^2] = \frac{s^2}{2}\mathbb{E}[e^{sf(X)}\|\nabla f(X)\|^2] \leqslant \frac{s^2L^2}{2}\mathbb{E}[e^{sf(X)}] \ .$$

The proof is concluded by Herbst's argument. □

Borel's inequality can be applied to show concentration for suprema of Gaussian processes.

Theorem 30 (Concentration for suprema of Gaussian processes)**.** *Let $(X_t)_{t\in T}$ denote a collection of Gaussian random variables $\mathrm{N}(\mu_t, \sigma_t^2)$ indexed by a separable set T. Let $\sigma^2 = \sup_{t\in T} \sigma_t^2$.*

$$\forall u > 0, \qquad \mathbb{P}\big(\sup_{t\in T}(X_t - \mu_t) > \mathbb{E}[\sup_{t\in T}(X_t - \mu_t)] + u \big) \leqslant e^{-u^2/2\sigma^2} \ .$$

Proof. Assume that T is finite, the extension to separable sets follows by density arguments. Denote $T = \{1, \dots, d\}$, $Y = (X_t - \mu_t)_{t\in T}$ is a centered Gaussian vector. Denote by Σ its covariance matrix and $A = \Sigma^{1/2}$ a symmetric positive semi-definite square-root of Σ. Y has the distribution of AX, where $X \sim \mathrm{N}(0, I_d)$. Define the function $f : \mathbb{R}^d \to \mathbb{R},\ x \mapsto \sup_{i\in\{1,\dots,d\}} (Ax)_i$. For any x and y, it follows that

$$|f(x) - f(y)| \leqslant \sup_{i\in\{1,\dots,d\}} (A(x-y))_i \leqslant \|x - y\| \sup_{\|v\|=1} |(Av)_i| \ .$$

Now, by Cauchy-Schwarz inequality,

$$\sup_{\|v\|=1} |(Av)_i| = \sup_{\|v\|=1} \left| \sum_{j=1}^d A_{i,j} v_j \right| \leqslant \sqrt{\sum_{j=1}^d A_{i,j}^2} = \sqrt{\Sigma_{i,i}} \leqslant \sigma \ .$$

The last equality uses the symmetry of A. Thus f is σ-Lipschitz and the result follows from Borel's Gaussian concentration inequality. □

Theorem 31. *Let $X = (X_1, \dots, X_N)$ denote i.i.d. Gaussian random vectors in $\mathbb{R}^d$ and let T denote a set of functions $t : \mathbb{R}^d \to \mathbb{R}$ such that, for all $t \in T$, t is 1-Lipshitz. Let $Z = \sup_{t\in T} \frac{1}{N} \sum_{i=1}^N [t(X_i) - Pt]$ or $\sup_{t\in T} \frac{1}{N} \big| \sum_{i=1}^N [t(X_i) - Pt] \big|$. Let Σ denote the covariance matrix of X_1. Then*

$$\forall u > 0, \qquad \mathbb{P}\big(Z > \mathbb{E}[Z] + u\big) \leqslant \exp\left(- \frac{Nu^2}{2\|\Sigma\|_{op}} \right) \ .$$

Proof. Write $X_i = \mu + AY_i$ with $A = \Sigma^{1/2}$ and Y_i standard Gaussian. Let

$$f : (\mathbb{R}^d)^n \to \mathbb{R},\ \mathbf{x} \mapsto \sup_{t\in T} \frac{1}{N} \sum_{i=1}^N (t(\mu + Ax_i) - Pt) \ .$$

Then, for any $\mathbf{x}, \mathbf{y} \in (\mathbb{R}^d)^n$,

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{y}) &\leqslant \sup_{t\in T} \frac{1}{N} \sum_{i=1}^N (t(\mu + Ax_i) - t(\mu + Ay_i)) \leqslant \frac{1}{N} \sum_{i=1}^N A(x_i - y_i) \\ &\leqslant \frac{\|A\|_{\mathrm{op}}}{N} \sum_{i=1}^N \|x_i - y_i\| \leqslant \frac{\|A\|_{\mathrm{op}}}{\sqrt{N}} \|\mathbf{x} - \mathbf{y}\| \ . \end{aligned}$$

The result follows from Borel's Gaussian concentration inequality as $\|A\|_{\mathrm{op}}^2 = \|\Sigma\|_{\mathrm{op}}$. □

3.2 Talagrand's concentration inequality

In this section, $f_i : \mathcal{X}^{N-1} \to \mathbb{R}$ denotes any measurable function and let $Z_i = f_i(X^{(i)})$. Let $\phi(x) = e^x - 1 - x$. The goal of this section is to establish a concentration result for suprema of empirical processes. Let $X = (X_1, \ldots, X_N)$ denote independent $\mathcal{X}$-valued random variables, each $x_0 \in \mathcal{X}$ being a vector $x_0 = (x_{0,t})_{t\in T}$. Assume that

$$\forall t \in T, \forall i \in \{1, \ldots, N\}, \qquad \mathbb{E}[X_{i,t}] = 0, \qquad X_{i,t} \leqslant 1, \text{ a.s. } .$$

Talagrand's concentration inequality shows sub-Poissonian deviations of $Z = \sup_{t\in T} X_{i,t}$ above its expectation $\mathbb{E}[Z]$. It proceeds by bounding from above the log-Laplace transform $\psi(s)$ of Z, using the entropy $\mathrm{Ent}(e^{sZ})$, but in a more involved way than the Herbst's argument.

3.2.1 Modified logarithmic Sobolev inequality

The starting point of this analysis is a modified version of log-Sobolev inequality. To establish this inequality, the following variational formulation of entropy is useful.

Theorem 32. *Let Y denote a nonnegative random variable such that $\mathbb{E}[\Phi(Y)] < \infty$. Then*

$$Ent(Y) = \inf_{u>0} \mathbb{E}[Y(\log(Y) - \log(u)) - (Y - u)] \ .$$

Proof. Recall that Φ is convex, so $\mathrm{Ent}(Y) \geqslant 0$ by Jensen's inequality, and

$$\mathrm{Ent}(Y) = \mathbb{E}[\Phi(Y) - \Phi(\mathbb{E}[Y])] \ .$$

Then, for any $u > 0$,

$$\begin{aligned}\mathbb{E}[\Phi(Y) - \Phi(u) - \Phi'(u)(Y - u)] &= \mathbb{E}[Y\log(Y) - u\log(u) - (1 + \log(u))(Y - u)] \\ &= \mathbb{E}[Y(\log(Y) - \log(u)) - (Y - u)] \ .\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}[Y(\log(Y) - \log(u)) - (Y - u)] - \mathrm{Ent}(Y) \\ &\qquad = \mathbb{E}[\Phi(Y) - \Phi(u) - \Phi'(u)(Y - u) - (\Phi(Y) - \Phi(\mathbb{E}[Y]))] \\ &\qquad = \mathbb{E}[\Phi(\mathbb{E}[Y]) - \Phi(u) - \Phi'(u)(Y - u)] \\ &\qquad = \Phi(\mathbb{E}[Y]) - \Phi(u) - \Phi'(u)(\mathbb{E}[Y] - u) \ .\end{aligned}$$

By convexity of Φ, this last term is always nonnegative and it is clearly null when $u = \mathbb{E}[Y]$. □

The modified log-Sobolev inequality bounds from above the entropy of Z using the increments $Z - Z_i$ via the function ϕ rather than the square function.

Theorem 33 (Modified log-Sobolev inequality)**.** *For any $s \in \mathbb{R}$,*

$$\mathrm{Ent}(e^{sZ}) \leqslant \sum_{i=1}^{n} \mathbb{E}[e^{sZ}\phi(s(Z_i - Z))] \ .$$

Proof. Basic algebra shows that

$$\begin{aligned} e^{sZ}\phi(s(Z_i - Z)) &= e^{sZ}(e^{s(Z_i - Z)} - s(Z_i - Z) - 1) \\ &= e^{sZ_i} - e^{sZ} + se^{sZ}(Z - Z_i) \ . \end{aligned}$$

Applying Theorem 32 conditionally on $X^{(i)}$, to $Y = e^{sZ}$ and $u = e^{sZ_i}$, it follows that

$$\begin{aligned} \mathrm{Ent}^{(i)}(e^{sZ}) &\leqslant \mathbb{E}^{(i)}[e^{sZ}(sZ - sZ_i) - (e^{sZ} - e^{sZ_i})] \\ &= \mathbb{E}^{(i)}[e^{sZ}\phi(s(Z_i - Z))] \ . \end{aligned}$$

Therefore, by sub-additivity of the entropy, see Theorem 23,

$$\begin{aligned} \mathrm{Ent}(e^{sZ}) &\leqslant \mathbb{E}\Big[\sum_{i=1}^{N} \mathrm{Ent}^{(i)}(e^{sZ})\Big] \leqslant \mathbb{E}\Big[\sum_{i=1}^{N} \mathbb{E}^{(i)}[e^{sZ}\phi(s(Z_i - Z))]\Big] \\ &= \sum_{i=1}^{n} \mathbb{E}[e^{sZ}\phi(s(Z_i - Z))] \ . \end{aligned}$$

□

3.2.2 Bousquet's version of Talagrand's inequality

Define $\sigma^2 = \sum_{i=1}^{n} \sup_{t\in T} \mathbb{E}[X_{i,t}^2]$ and $\nu = 2\mathbb{E}[Z] + \sigma^2$. Recall that $h(u) = (1+u)\log(1+u) - u$. Talagrand's inequality shows that $Z = \sup_{t\in T} X_{i,t}$ is a ν-sub-Poissonian random variables. The following version of this inequality, with sharp constants, was first established by Bousquet [11].

Theorem 34 (Bousquet's version of Talagrand's concentration inequality)**.** *The random variable $Z - \mathbb{E}[Z]$ is ν-sub-Poissonian, that is, for any $s > 0$, $\mathbb{E}[e^{s(Z-\mathbb{E}[Z])}] \leqslant \nu\phi(s)$. Moreover,*

$$\forall x > 0, \qquad \mathbb{P}(Z > \mathbb{E}[Z] + x) \leqslant e^{-\nu h(t/\nu)} \ .$$

Proof. The proof relies on the following result from calculus.

Lemma 35. *For any $s \geqslant 0$ and any $x \leqslant 1$,*

$$\frac{\phi(-sx)}{\phi(-s)} \leqslant \frac{x + (x^2/2 - x)e^{-sx}}{1 - e^{-s}/2} \ .$$

The proof of the lemma is omitted. Going back to the proof of the theorem, define

$$Z_i = \sup_{t\in T} \sum_{1\leqslant j\leqslant n, j\neq i} X_{j,t} \ .$$

Let also t_0 such that $Z = \sum_{1\leqslant i\leqslant n} X_{i,t_0}$ and t_i such that $Z_i = \sum_{1\leqslant j\leqslant n, j\neq i} X_{j,t_i}$. Remark that $X_{i,t_i} \leqslant Z - Z_i \leqslant X_{i,t_0} \leqslant 1$, so $\mathbb{E}^{(i)}[Z - Z_i] \geqslant \mathbb{E}^{(i)}[X_{i,t_i}] = 0$ and

$$\sum_{i=1}^{n} Z - Z_i \leqslant Z \ .$$

By the modified logarithmic Sobolev inequality,

$$\mathrm{Ent}(e^{sZ}) \leqslant \sum_{i=1}^{n} \mathbb{E}[e^{sZ}\phi(-s(Z-Z_i))] \ .$$

Let

$$g(s) = \frac{\phi(-s)}{1-e^{-s}/2} = \frac{e^{-s}-1+s}{1-e^{-s}/2} = \frac{1-e^s+se^s}{e^s-1/2} = \frac{-\phi(s)-s+se^s}{e^s-1/2} \ .$$

Lemma 35 applied with $x = (Z - Z_i)$ implies

$$\begin{aligned} e^{sZ}\phi(-s(Z-Z_i)) &\leqslant \frac{(Z-Z_i)e^{sZ} + ((Z-Z_i)^2/2 - (Z-Z_i))e^{sZ-s(Z-Z_i)}}{1-e^{-s}/2}\phi(-s) \\ &\leqslant g(s)\left(e^{sZ_i}\left[\frac{(Z-Z_i)^2}{2} - (Z-Z_i)\right] + (Z-Z_i)e^{sZ}\right) \ . \end{aligned}$$

Now, as $Z - Z_i - X_{i,t_i} \geqslant 0$ and $Z - Z_i + X_{i,t_i} - 2 \leqslant 0$,

$$(Z-Z_i)^2 - 2(Z-Z_i) - [X_{i,t_i}^2 - 2X_{i,t_i}] = (Z-Z_i-X_{i,t_i})(Z-Z_i+X_{i,t_i}-2) \leqslant 0 \ .$$

It follows that

$$\mathbb{E}^{(i)}\left[\frac{(Z-Z_i)^2}{2} - (Z-Z_i)\right] \leqslant \mathbb{E}^{(i)}\left[\frac{X_{i,t_i}^2}{2} - X_{i,t_i}\right] = \frac{\mathbb{E}^{(i)}[X_{i,t_i}^2]}{2} \ .$$

Therefore,

$$\begin{aligned} \mathbb{E}^{(i)}[e^{sZ}\phi(-s(Z-Z_i))] &\leqslant g(s)\left(\mathbb{E}^{(i)}[(Z-Z_i)e^{sZ}] + \frac{1}{2}\mathbb{E}^{(i)}[X_{i,t_i}^2]e^{sZ_i}\right) \\ &\leqslant g(s)\left(\mathbb{E}^{(i)}[(Z-Z_i)e^{sZ}] + \frac{1}{2}\sup_{t\in T}\mathbb{E}[X_{i,t}^2]e^{sZ_i}\right) \ . \end{aligned}$$

As $\mathbb{E}^{(i)}[Z - Z_i] \geqslant 0$, $Z_i \leqslant \mathbb{E}^{(i)}[Z]$ and, by Jensen's inequality,

$$e^{sZ_i} \leqslant e^{s\mathbb{E}^{(i)}[Z]} \leqslant \mathbb{E}^{(i)}[e^{sZ}] \ ,$$

thus

$$\mathbb{E}^{(i)}[e^{sZ}\phi(-s(Z-Z_i))] \leqslant g(s)\mathbb{E}^{(i)}\left[\left(Z - Z_i + \frac{1}{2}\sup_{t\in T}\mathbb{E}[X_{i,t}^2]\right)e^{sZ}\right] \ .$$

Summing up over i and taking the expectation, it follows from $\sum_{i=1}^{n}(Z-Z_i) \leqslant Z$ that

$$\mathrm{Ent}(e^{sZ}) \leqslant g(s)\mathbb{E}\left[\left(Z + \frac{\sigma^2}{2}\right)e^{sZ}\right] = g(s)\mathbb{E}\left[\left(Z - \mathbb{E}[Z] + \frac{\nu}{2}\right)e^{sZ}\right]$$

By (3.1), this can be rewritten

$$(s - g(s))\psi'(s) - \psi(s) \leqslant g(s)\frac{\nu}{2} \ . \tag{3.15}$$

Let $\zeta(s) = \phi(s) + s/2$, so $\zeta'(s) = e^s - 1 + 1/2 = e^s - 1/2$ and

$$s - g(s) = s + \frac{\phi(s) + s - se^s}{e^s - 1/2} = \frac{se^s - s/2 + \phi(s) + s - se^s}{e^s - 1/2} = \frac{\phi(s) + s/2}{e^s - 1/2} = \frac{\zeta(s)}{\zeta'(s)} \ .$$

In particular thus $g(s) = s - \zeta(s)/\zeta'(s)$ so, multiplying inequality (3.15) by $\zeta'(s)$ shows

$$\zeta(s)\psi'(s) - \zeta'(s)\psi(s) \leqslant (s\zeta'(s) - \zeta(s))\frac{\nu}{2} \ .$$

Dividing by $\zeta^2(s)$ yields

$$\frac{\zeta(s)\psi'(s) - \zeta'(s)\psi(s)}{\zeta^2(s)} \leqslant \frac{s\zeta'(s) - \zeta(s)}{\zeta^2(s)}\frac{\nu}{2} \ .$$

that is

$$\left(\frac{\psi(s)}{\zeta(s)}\right)' \leqslant -\frac{\nu}{2}\left(\frac{s}{\zeta(s)}\right)' \ .$$

As $\psi(0) = \psi'(0) = 0$, the function $u : s \mapsto \psi(s)/\zeta(s)$ can be continuously extended in 0 by defining $u(0) = 0$. Therefore, integrating over s shows that

$$\frac{\psi(s)}{\zeta(s)} \leqslant -\frac{\nu}{2}\left(\frac{s}{\zeta(s)} - \lim_{s\to 0}\frac{s}{\zeta(s)}\right) = -\frac{\nu}{2}\left(\frac{s}{\zeta(s)} - 2\right) \ .$$

Finally, multiplying by $\zeta(s)$,

$$\psi(s) \leqslant -\nu\left(\frac{s}{2} - \zeta(s)\right) = \nu\phi(s) \ .$$

This shows the first part of the theorem. The second part comes then from the first part and Lemma 20. □

3.3 PAC-Bayesian inequalities

Let $X \in \mathcal{X}$ denote a random variable and let F denote a measurable space. Let $\Gamma : F \times \mathcal{X} \to \mathbb{R}$ denote a bounded measurable function. For any measures μ and ρ on F, let

$$K(\rho, \mu) = \begin{cases} \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right)\mathrm{d}\rho & \text{if } \rho \ll \mu \\ +\infty & \text{otherwise} \end{cases} \ .$$

Let $X_1, \ldots, X_N$ denote i.i.d. copies of X. The entropy method is not the only method to show uniform deviation inequalities for empirical processes. A famous alternative, that has been fruitfully exploited by Catoni for example, see [13], is known as PAC-Bayesian inequality. The idea is to exploit a variational formula for the Kullback divergence to obtain this uniformity.

Theorem 36 (PAC-Bayesian inequality)**.** *For any probability measure μ on F, for any $t > 0$, with probability $1 - e^{-t}$, for any probability measure ρ on F,*

$$P_N\left[\int \Gamma_f \mathrm{d}\rho(f)\right] \leqslant \int \log P\left[e^{\Gamma_f}\right]\mathrm{d}\rho(f) + \frac{K(\rho, \mu) + t}{N} \ .$$

Proof. The proof relies on the following variational formula.

$$\log \int e^h \mathrm{d}\mu = \sup_\rho \int h \mathrm{d}\rho - K(\rho, \mu) \ , \tag{3.16}$$

where the supremum is taken over all probability measures ρ on F.

Proof of (3.16). Choose ρ such that $\mathrm{d}\rho = e^h \mathrm{d}\mu / \int e^h \mathrm{d}\mu$. Then,

$$\sup_\rho \int h \mathrm{d}\rho - K(\rho, \mu) \geqslant \int (h - h + \log \int e^h \mathrm{d}\mu) \mathrm{d}\rho = \log \int e^h \mathrm{d}\mu \ .$$

In words, the left-hand side of (3.16) is smaller than the right-hand side. Conversely, by Jensen's inequality

$$\int \left(h + \log \frac{\mathrm{d}\mu}{\mathrm{d}\rho} \right) \mathrm{d}\rho = \int \log \left(e^h \frac{\mathrm{d}\mu}{\mathrm{d}\rho} \right) \mathrm{d}\rho \leqslant \log \int e^h \mathrm{d}\mu \ .$$

This shows that the right-hand side in (3.16) is also smaller than the left-hand side, which concludes the proof of this inequality. □

Applying (3.16) with $h = N(P_N \Gamma_f - \log P e^{\Gamma_f})$ yields

$$\begin{aligned}
\mathbb{E}\left[e^{\sup_\rho N \int (P_N \Gamma_f - \log P e^{\Gamma_f}) \mathrm{d}\rho - K(\rho,\mu)} \right] &= \mathbb{E}\left[\int e^{N(P_N \Gamma_f - \log P e^{\Gamma_f})} \mathrm{d}\mu \right] \\
&= \int \mathbb{E}\left[e^{N(P_N \Gamma_f - \log P e^{\Gamma_f})} \mathrm{d}\mu \right] \\
&= \int \prod_{i=1}^N P\left[\frac{e^{\Gamma_f}}{P e^{\Gamma_f}} \right] \mathrm{d}\mu = 1 \ .
\end{aligned}$$

By the Chernoff bound, any random variable W such that $\mathbb{E}[e^W] \leqslant 1$ satisfies

$$\forall t > 0, \qquad \mathbb{P}(W > t) \leqslant e^{-t + \log \mathbb{E}[e^W]} = e^{-t} \ .$$

The result follows by applying this basic inequality to

$$W = N \int (P_N \Gamma_f - \log P e^{\Gamma_f}) \mathrm{d}\rho - K(\rho, \mu) \ .$$

□

3.4 Hanson-Wright inequality

In this section, $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{X}$ is a random vector in $\mathbb{R}^n$ such that $\mathbb{E}[X] = 0$ and, for any $\mathbf{u} \in \mathbb{R}^n$,

$$\mathbb{E}[\mathrm{e}^{\mathbf{u}^T \mathbf{X}}] \leqslant \exp\left(\frac{\|\mathbf{u}\|^2 v^2}{2} \right) \ .$$

The following result is usually refer to as Hanson-Wright's inequality.

Theorem 37 (Hanson-Wright)**.** *Let $\Sigma = \boldsymbol{A}^T\boldsymbol{A}$. For any $t > 0$,*

$$\mathbb{P}\big(\|\boldsymbol{AX}\|^2 > v^2\big(\, Tr(\Sigma) + 2\sqrt{Tr(\Sigma^2)t} + 8\|\Sigma\|_{op}t\big)\big) \leqslant \mathrm{e}^{-t} \ .$$

Proof. Let $\mathbf{Z}$ denote a random vector independent from $\mathbf{X}$ such that $\mathbf{Z} \sim \mathrm{N}(0, \mathbf{I}_n)$. We have, for any $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{u}^T\mathbf{Z} \sim \mathrm{N}(0, \|u\|^2)$, so

$$\mathbb{E}[\mathrm{e}^{\mathbf{u}^T\mathbf{Z}}] \leqslant \exp\left(\frac{\|\mathbf{u}\|^2}{2}\right) \ .$$

Therefore, for any $s > 0$, $\epsilon > 0$,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}] &\geqslant \mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}|\|\mathbf{AX}\| > \epsilon]\mathbb{P}(\|\mathbf{AX}\| > \epsilon) \\ &\geqslant \exp\left(\frac{s^2\epsilon^2}{2}\right)\mathbb{P}(\|\mathbf{AX}\| > \epsilon) \ . \end{aligned} \tag{3.17}$$

Moreover,

$$\mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}] = \mathbb{E}\big[\mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}|\mathbf{X}]\big] \leqslant \mathbb{E}\left[\exp\left(\frac{s^2v^2\|\mathbf{A}^T\mathbf{Z}\|^2}{2}\right)\right] \ .$$

Now, let $\mathbf{A} = \mathbf{UDV}^T$ denote the singular value decomposition of $\mathbf{A}$, since $\mathbf{V}$ is orthogonal, we have

$$\|\mathbf{A}^T\mathbf{Z}\| = \|\mathbf{VDU}^T\mathbf{Z}\| = \|\mathbf{DU}^T\mathbf{Z}\| \ .$$

Moreover, by rotational invariance of the Gaussian distribution, $\mathbf{U}^T\mathbf{Z} \sim \mathbf{Z}$, hence

$$\mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}] \leqslant \mathbb{E}\left[\exp\left(\frac{s^2v^2\|\mathbf{DZ}\|^2}{2}\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{s^2v^2d_{i,i}^2Z_i^2}{2}\right)\right] \ .$$

As for any $\alpha \in (0, 1/2)$, $\mathbb{E}[\mathrm{e}^{\alpha Z_i^2}] = \frac{1}{\sqrt{1-2\alpha}}$, we get, for any $s < 1/v^2\|A\|_{\mathrm{op}}^2$,

$$\mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}] \leqslant \prod_{i=1}^n \frac{1}{\sqrt{1 - s^2v^2d_{i,i}^2}} = \exp\left(-\frac{1}{2}\sum_{i=1}^n \log(1 - s^2v^2d_{i,i}^2)\right) \ .$$

As, for any $u \in (0, 1)$,

$$-\log(1-u) = \sum_{j=1}^{\infty}\frac{u^j}{j} \leqslant u + \frac{u^2}{2(1-u)} \ ,$$

we get, on one side

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{s\mathbf{Z}^T\mathbf{AX}}] &\leqslant \exp\left(\frac{1}{2}\sum_{i=1}^n s^2v^2d_{i,i}^2 + \frac{s^4v^4d_{i,i}^4}{1 - s^2v^2d_{i,i}^2}\right) \\ &\leqslant \exp\left(\frac{s^2v^2}{2}\operatorname{Tr}(\Sigma) + \frac{s^4v^4\operatorname{Tr}(\Sigma^2)}{4(1 - s^2v^2\|\Sigma\|_{\mathrm{op}})}\right) \\ &= \exp\left(\frac{s^2v^2}{2}\operatorname{Tr}(\Sigma) + \frac{s^4v^4\operatorname{Tr}(\Sigma^2)}{4(1 - s^2v^2\|\Sigma\|_{\mathrm{op}})}\right) \ . \end{aligned}$$

Together with (3.17), this yields, for any $\epsilon > 0$ and $s > 0$,

$$\mathbb{P}(\|\mathbf{AX}\| > \epsilon) \leqslant \exp\left(-\frac{s^2\epsilon^2}{2} + \frac{s^2v^2}{2}\operatorname{Tr}(\Sigma) + \frac{s^4v^4\operatorname{Tr}(\Sigma^2)}{4(1-s^2v^2\|\Sigma\|_{\mathrm{op}})}\right) .$$

Optimizing in s, we get, for $h_1(x) = 1 + x - \sqrt{1+2x}$, for any $\epsilon > 0$,

$$\mathbb{P}(\|\mathbf{AX}\| > \epsilon) \leqslant \exp\left(-\frac{\operatorname{Tr}(\Sigma^2)}{2\|\Sigma\|_{\mathrm{op}}^2} h_1\left(\frac{\|\Sigma\|_{\mathrm{op}}(\epsilon^2 - v^2\operatorname{Tr}(\Sigma))}{2v^2\operatorname{Tr}(\Sigma^2)}\right)\right) .$$

This is equivalent to

$$\forall u > 0, \qquad \mathbb{P}\big(\|\mathbf{AX}\|^2 > v^2\big(\operatorname{Tr}(\Sigma) + 2\sqrt{\operatorname{Tr}(\Sigma^2)u} + 8\|\Sigma\|_{\mathrm{op}}u\big)\big) \leqslant \mathrm{e}^{-u} .$$

□

Corollary 38 (Concentration for Euclidean norm of sub-Gaussian vectors)**.** *Let $\boldsymbol{X}, \boldsymbol{X}_1, \dots, \boldsymbol{X}_N$ denote i.i.d. vectors of $\mathbb{R}^d$ with invertible covariance matrix $\Sigma = \mathbb{E}[(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])^T]$. Assume that $\boldsymbol{X}$ is v^2 sub-Gaussian in the sense that $\boldsymbol{Y} = \Sigma^{-1/2}(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])$ satisfies*

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \qquad \mathbb{E}[\mathrm{e}^{\boldsymbol{u}^T\boldsymbol{Y}}] \leqslant \mathrm{e}^{\|\boldsymbol{u}\|^2v^2/2} .$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_{i=1}^N(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}])\right\|^2 > \frac{v^2}{N}\big(\, Tr(\Sigma) + 2\sqrt{Tr(\Sigma^2)t} + 8\|\Sigma\|_{op}t\big)\right) \leqslant \mathrm{e}^{-t} .$$

As $Tr(\Sigma^2) \leqslant \|\Sigma\|_{op}\, Tr(\Sigma)$, the inequality $2ab \leqslant a^2 + b^2$ also implies that, for any $t > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_{i=1}^N(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}])\right\|^2 > \frac{v^2}{N}\big(2\, Tr(\Sigma) + 9\|\Sigma\|_{op}t\big)\right) \leqslant \mathrm{e}^{-t} .$$

Proof. Assume that T is finite, the extension to separable sets can be done by standard density arguments.

□

3.5 Deviation of suprema of median-of-means processes

To conclude this chapter, we present two deviation results for suprema of MOM processes. Both show deviations of this process above a term involving the *Rademacher complexity* of F. Recall that the Rademacher complexity of a class F of functions $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$D(F) = \left(\mathbb{E}\left[\sup_{f\in F}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i f(X_i)\right\}\right]\right)^2 .$$

The quantity $D(F)$ can easily be evaluated when F is a set linear functionals. Let $r > 0$, $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$ and $r\mathbf{B} = \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\| \leqslant r\}$

$$F = \{f : \mathbb{R}^d \to \mathbb{R} : \exists \mathbf{a} \in r\mathbf{B},\ f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}\} \ .$$

Let $X_0 \in \mathbb{R}^d$ be such that $PX_0 = 0$, $P\|X_0\|^2 < \infty$ and let $\Sigma_P = P[X_0X_0^T]$. In this case,

$$\begin{aligned} D(F) &= \left(\mathbb{E}\left[\sup_{\mathbf{a}\in r\mathbf{B}}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i \mathbf{a}^T X_i\right\}\right]\right)^2 \\ &= r^2\left(\mathbb{E}\left[\sup_{\mathbf{a}\in \mathbf{B}}\left\{\mathbf{a}^T\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i X_i\right)\right\}\right]\right)^2 \\ &= r^2\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i X_i\right\|\right]\right)^2 \ . \end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned} D(F) &\leqslant r^2\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i X_i\right\|^2\right] \\ &= \frac{2r^2}{N}\sum_{1\leqslant i,j\leqslant N}\mathbb{E}\big[\epsilon_i\epsilon_j X_i^T X_j\big] \\ &= r^2\mathbb{E}\big[X_0^T X_0\big] \\ &= r^2\mathbb{E}\big[\mathrm{Tr}\big(X_0X_0^T\big)\big] \\ &= r^2\mathrm{Tr}(\Sigma) \ . \end{aligned} \tag{3.18}$$

In particular, when $r = 1$ and Σ is the identity matrix $\Sigma = \mathbf{I}_d$, $D(F)$ is the dimension of the state space $D(F) = d$. The first result is a deviation for suprema of MOM processes above $\sqrt{D(F)/N}$. It is established using the tools introduced by Lugosi and Mendelson [39].

Theorem 39 (Concentration for suprema of MOM processes)**.** *Let F denote a separable set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\sup_{f\in F}\sigma^2(f) = \sigma^2 < \infty$, where $\sigma^2(f) = Var(f(X))$. Then, for any $K \in \{1, \ldots, N/2\}$,*

$$\mathbb{P}\left(\sup_{f\in F}|MOM_K[f] - Pf| \geqslant 128\sqrt{\frac{D(F)}{N}} \vee 4\sigma\sqrt{\frac{2K}{N}}\right) \leqslant e^{-K/32} \ .$$

Proof. Assume that F is finite, the general case follows by a standard density argument. The basic idea is that, for any $\epsilon > 0$,

$$\sup_{f\in F}|\mathrm{MOM}_K[f] - Pf| \leqslant \epsilon \quad \text{if} \quad \sup_{f\in F}\sum_{k=1}^K \mathbf{1}_{\{|(P_{B_k}-P)f|>\epsilon\}} \leqslant \frac{K}{2} \ .$$

Introduce ϕ, a 1-Lipschitz function such that $\mathbf{1}_{x\geqslant 2} \leqslant \phi(x) \leqslant \mathbf{1}_{x\geqslant 1}$. We have

$$\begin{aligned} &\sup_{f\in F}\sum_{k=1}^K \mathbf{1}_{\{|(P_{B_k}-P)f|>\epsilon\}} \leqslant \sup_{f\in F}\sum_{k=1}^K \phi\left(\frac{2|(P_{B_k}-P)f|}{\epsilon}\right) \\ &\leqslant K\sup_{f\in F}\mathbb{P}\left(|(P_{B_1}-P)f| > \frac{\epsilon}{2}\right) + \sup_{f\in F}\left\{\sum_{k=1}^K \phi\left(\frac{2|(P_{B_k}-P)f|}{\epsilon}\right) - \mathbb{E}\left[\phi\left(\frac{2|(P_{B_k}-P)f|}{\epsilon}\right)\right]\right\} \ . \end{aligned}$$

The first term in this upper-bound can be bounded from above using Chebyshev's inequality as follows.

$$\sup_{f\in F}\mathbb{P}\left(|(P_{B_k}-P)f|>\frac{\epsilon}{2}\right)\leqslant\frac{4\sigma^2K}{\epsilon^2N}\ .$$

Using the bounded difference inequality, the second term is bounded from above, with probability at least $1-e^{-2x^2/K}$, by

$$\mathbb{E}\left[\sup_{f\in F}\left\{\sum_{k=1}^K\phi\left(\frac{2|(P_{B_k}-P)f|}{\epsilon}\right)-\mathbb{E}\left[\phi\left(\frac{2|(P_{B_k}-P)f|}{\epsilon}\right)\right]\right\}\right]+x\ .$$

Using the symmetrization trick, the expectation is now bounded from above by

$$2\mathbb{E}\left[\sup_{f\in F}\left\{\sum_{k=1}^K\epsilon_k\phi\left(\frac{2|(P_{B_k}-P)f|}{\epsilon}\right)\right\}\right]\ .$$

By Ledoux and Talagrand's contraction lemma, this term is bounded from above by

$$\frac{16}{\epsilon}\mathbb{E}\left[\sup_{f\in F}\left\{\sum_{k=1}^K\epsilon_k(P_{B_k}-P)f\right\}\right]\ .$$

By the symmetrization trick, this term is bounded from above by

$$\frac{32K}{\epsilon}\sqrt{\frac{D(F)}{N}}\ .$$

Overall, with probability at least $1-e^{-2x^2/K}$,

$$\sup_{f\in F}\sum_{k=1}^K\mathbf{1}_{\{|(P_{B_k}-P)f|>\epsilon\}}\leqslant\frac{32K}{\epsilon}\sqrt{\frac{D(F)}{N}}+\frac{4\sigma^2K}{\epsilon^2N}+x$$

Choose $\delta\in 1/2$, $\epsilon=128\sqrt{\frac{D(F)}{N}}\vee\sqrt{32\frac{\sigma^2K}{N}}$ and $x=K/8$, this shows that, with probability $1-e^{-K/32}$,

$$\sup_{f\in F}|\mathrm{MOM}_K[f]-Pf|\leqslant 128\sqrt{\frac{D(F)}{N}}\vee 4\sigma\sqrt{\frac{2K}{N}}\ .$$

□

Some results require the following extension of the previous result whose proof follows exactly the same arguments and is left to the reader.

Theorem 40 (General concentration bound for suprema of MOM processes)**.** *Let F denote a separable set of functions $f:\mathcal{X}\to\mathbb{R}$ such that $\sup_{f\in F}\sigma^2(f)=\sigma^2<\infty$, where $\sigma^2(f)=Var(f(X))$. Let $\alpha\in(0,1)$. There exists a constant c_α such that, for any $K\geqslant 1/\alpha$, with probability at least $1-e^{-K/c_\alpha}$, there exists at least $(1-\alpha)K$ blocks B_k where*

$$\forall f\in F,\qquad |(P_{B_k}-P)f|\leqslant c_\alpha\left(\sqrt{\frac{D(F)}{N}}\vee\sigma\sqrt{\frac{K}{N}}\right)\ .$$

This general result admits the following corollary that was first proved in [42] and that will be used repeatedly in the following.

Corollary 41. *Assume that $X_1, \dots, X_N$ are i.i.d. random vectors of $\mathbb{R}^d$, with common distribution P such that $P[\|X\|^2] < \infty$. Let $\Sigma = P[(X - PX)(X - PX)^T]$, $\alpha \in (0,1)$ and $r > 0$. There exists a constant c_α such that, for any $K \geqslant 1/\alpha$, with probability at least $1 - e^{-K/c_\alpha}$, there exists at least $(1-\alpha)K$ blocks B_k where*

$$\forall \boldsymbol{a} \in \mathbb{R}^d : \|\boldsymbol{a}\| \leqslant r, \qquad |(P_{B_k} - P)[\boldsymbol{a}^T \cdot]| \leqslant c_\alpha r \sqrt{\frac{Tr(\Sigma) \vee \|\Sigma\|_{op} K}{N}} \ .$$

Proof. Apply Theorem 40 to the class $F = \{\mathbf{a}^T \cdot : \|\mathbf{a}\| \leqslant r\}$. By (3.18) , $D(F) \leqslant r^2 \mathrm{Tr}(\Sigma)$ and, for any $\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\| \leqslant r$,

$$\mathrm{Var}(\mathbf{a}^T X) = \mathbf{a}^T \Sigma \mathbf{a} \leqslant r^2 \|\Sigma\|_{\mathrm{op}} \ .$$

The result follows. □

As for univariate mean estimate, this first analysis can be refined under stronger moments assumptions using Minsker-Strawn's approach. Denote by Q the tail function of a standard Gaussian, $n = N/K$ and

$$g(n, f) = \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sqrt{n} \frac{(P_{B_1} - P)f}{\sigma(f)} > t \right) - Q(t) \right| \ .$$

Recall that if F is a class of functions such that $P|f|^3 < \infty$ for any $f \in F$ and $\sup_{f \in F} P[|f - Pf|^3]/\sigma(f)^3 = \gamma < \infty$, then, Berry-Esseen theorem implies that

$$\sup_{f \in F} g(n, f) := g(n) \leqslant \frac{\gamma}{\sqrt{n}} \ .$$

The key-point is that one has to use a "smoothed" version of median of means estimators. Define the function

$$\rho(t) = \begin{cases} -1 & \text{if } t \leqslant -1 \\ t & \text{if } -1 \leqslant t \leqslant 1 \\ 1 & \text{if } t \geqslant 1 \end{cases} \tag{3.19}$$

Then, let $\Delta \geqslant \sup_{f \in F} \sigma(f)$ and let $\hat{P}_K f$ be solution of the equation

$$\sum_{k=1}^{K} \rho\left(\sqrt{n} \frac{P_{B_k} f - z}{\Delta} \right) = 0 \ .$$

Theorem 42. *[Minsker's deviation bound for suprema of smoothed MOM processes] Assume that s and K satisfy*

$$300\left(\frac{16}{\Delta} \sqrt{\frac{D(F)}{N}} + \sqrt{\frac{2s}{N}} + 4\frac{g(n)}{\sqrt{n}} \right) \leqslant \sqrt{\frac{K}{N}} \ ,$$

Then

$$\mathbb{P}\left(\sup_{f \in F} |\hat{P}_K f - Pf| \geqslant 300\sqrt{\frac{D(F)}{N}} + 20\Delta\left(\sqrt{\frac{2s}{N}} + 4\frac{g(n)}{\sqrt{n}} \right) \right) \leqslant e^{-s} \ .$$

Remark 43. *Assume that F is a class of functions such that $P|f|^3 < \infty$ for any $f \in F$ and $\sup_{f\in F} P[|f - Pf|^3]/\sigma(f)^3 = \gamma < \infty$, then, Berry-Esseen theorem implies that*

$$\sup_{f\in F} g(n,f) := g(n) \leqslant \frac{\gamma}{\sqrt{n}} \ .$$

Assume moreover that $D(F) \leq \Delta^2\sqrt{N}$. Let $K = C\sqrt{N}$, where C is a sufficiently large absolute constant. Then Theorem 42 implies that, simultaneously for all $s \leq C'\sqrt{N}$, with probability larger than $1 - e^{-s}$,

$$\sup_{f\in F} |\hat{P}_K f - Pf| \leqslant C\left(\sqrt{\frac{D(F)}{N}} + \Delta\sqrt{\frac{\gamma^2 + s}{N}}\right) \ .$$

Proof. Let us first remark that $(\hat{P}_K - P)f$ is solution of $Q_K^{(n)}(f,\cdot) = 0$, where

$$Q_K^{(n)}(f,z) := \frac{1}{K}\sum_{k=1}^{K} \rho\left(\sqrt{n}\frac{(P_{B_k} - P)f - z}{\Delta}\right) \ .$$

The strategy is then to find a deterministic function $U(\cdot)$ such that, for any $z \in \mathbb{R}$, w.h.p., for any $f \in F$,

$$Q_K^{(n)}(f,z) \leqslant U(z) \ .$$

Then, if z_+ denotes the smallest solution of $U(z) = 0$, on the event $\sup_{f\in F} Q_K^{(n)}(f,z_+) \leqslant U(z_+)$, for any $f \in F$,

$$Q_K^{(n)}(f,z_+) \leqslant U(z_+) = 0 = Q_K^{(n)}(f,(\hat{P}_K - P)f) \ .$$

As $Q_K^{(n)}(f,\cdot)$ is non-increasing, this implies

$$\mathbb{P}\big(\forall f \in F,\ (\hat{P}_K - P)f \leqslant z_+\big) \geqslant \mathbb{P}\big(\sup_{f\in F} Q_K^{(n)}(f,z_+) \leqslant U(z_+)\big) \ .$$

Fix z and bound uniformly from above $Q_K^{(n)}(f,z)$. Let $G_k(f) = \sqrt{n}(P_{B_k} - P)f/\sigma(f)$, then

$$Q_K^{(n)}(f,z) = \frac{1}{K}\sum_{k=1}^{K} \rho\left(\frac{\sigma(f)}{\Delta}G_k(f) - \frac{\sqrt{n}z}{\Delta}\right) \ .$$

Therefore

$$Q_K^{(n)}(f,z) \leqslant (Q_K^{(n)}(f,z) - Q^{(n)}(f,z)) + (Q^{(n)}(f,z) - Q(f,z)) + Q(f,z) \ ,$$

where G is a standard Gaussian random variable and

$$Q^{(n)}(f,z) := \mathbb{E}\left[\rho\left(\frac{\sigma(f)}{\Delta}G_k(f) - \frac{\sqrt{n}z}{\Delta}\right)\right] \ ,$$
$$Q(f,z) := \mathbb{E}\left[\rho\left(\frac{\sigma(f)}{\Delta}G - \frac{\sqrt{n}z}{\Delta}\right)\right] \ .$$

Let $\psi_1(f,z) = Q_K^{(n)}(f,z) - Q^{(n)}(f,z)$. As ρ takes values in $[-1,1]$, the bounded difference inequality grants that, with probability larger than $1-e^{-s}$

$$\sup_{f\in F} \psi_1(f,z) \leqslant \mathbb{E}\big[\sup_{f\in F} \psi_1(f,z)\big] + \sqrt{\frac{2s}{K}} \ .$$

By symmetrization and contraction $(\rho(\cdot - x) - \rho(-x))$ being 1-Lipshitz,

$$\mathbb{E}\big[\sup_{f\in F} \psi_1(f,z)\big] \leqslant 16\frac{\sqrt{n}}{\Delta}\sqrt{\frac{D(F)}{N}} \ .$$

For any real numbers α and β and any real valued random variable X with c.d.f. F_X,

$$\begin{aligned}\mathbb{E}[\rho(\alpha X - \beta)] &= -F_X\left(\frac{\beta-1}{\alpha}\right) + 1 - F_X\left(\frac{\beta+1}{\alpha}\right) + \int_{\frac{\beta-1}{\alpha}}^{\frac{\beta+1}{\alpha}} (\alpha t - \beta) dF_X(t) \\ &= -2F_X\left(\frac{\beta-1}{\alpha}\right) + 1 - \alpha\int_{\frac{\beta-1}{\alpha}}^{\frac{\beta+1}{\alpha}} F_X(t)dt \ . \end{aligned} \tag{3.20}$$

By definition of $g(n)$, it follows that

$$\sup_{f\in F} |Q^{(n)}(f,z) - Q(f,z)| \leqslant 4g(n) \ .$$

Now, let $f \in F$, $\alpha = \sigma(f)/\Delta$ and $\beta = \sqrt{n}z/\Delta$,

$$\begin{aligned} Q(f,z) &= \mathbb{P}\big(\alpha G - \beta \geqslant 1\big) - \mathbb{P}\big(\alpha G - \beta \leqslant -1\big) + \mathbb{E}\big[\big(\alpha G - \beta\big)\mathbf{1}_{|\alpha G - \beta| \leqslant 1}\big] \\ &\leqslant \alpha\mathbb{E}\big[G\mathbf{1}_{|\alpha G-\beta|\leqslant 1}\big] - \beta\mathbb{P}\big(|\alpha G - \beta| \leqslant 1\big) \\ &= \frac{\alpha}{\sqrt{2\pi}}[e^{-(\beta-1)^2/2\alpha^2} - e^{-(\beta+1)^2/2\alpha^2}] - \beta\mathbb{P}\big(|\alpha G - \beta| \leqslant 1\big) \\ &\leqslant \sqrt{\frac{2}{\pi}}\frac{\beta}{\alpha}e^{-(\beta-1)^2/2\alpha^2} - \beta\mathbb{P}\big(|\alpha G - \beta| \leqslant 1\big) \\ &= -\frac{\beta}{\alpha\sqrt{2\pi}}\int_{\beta-1}^{\beta+1} e^{-x^2/2\alpha^2} - e^{-(\beta-1)^2/2\alpha^2}\mathrm{d}x \ . \end{aligned}$$

Assume now that $\beta \leqslant 1/16$, and, as $\alpha \leqslant 1$, write

$$\begin{aligned}\int_{\beta-1}^{\beta+1} e^{-x^2/2\alpha^2} - e^{-(\beta-1)^2/2\alpha^2}\frac{\mathrm{d}x}{\alpha\sqrt{2\pi}} &\geqslant \int_{-1/2}^{1/2} e^{-x^2/2\alpha^2}(1 - e^{x^2-(\beta-1)^2/2\alpha^2})\frac{\mathrm{d}x}{\alpha\sqrt{2\pi}} - \frac{2\beta e^{-(\beta-1)^2/2\alpha^2}}{\alpha\sqrt{2\pi}} \\ &\geqslant (1 - e^{-161/512\alpha^2})\mathbb{P}(-1/2 \leqslant G \leqslant 1/2) - \frac{e^{-225/512\alpha^2}}{8\alpha\sqrt{2\pi}} \\ &\geqslant (1 - e^{-161/512})\mathbb{P}(-1/2 \leqslant G \leqslant 1/2) - \frac{e^{-1/2}}{8\sqrt{2\pi}} \geqslant 0.06 \ . \end{aligned}$$

It follows that, if $z \leqslant \Delta/16\sqrt{n}$,

$$Q(f,z) \leqslant -0.06\sqrt{n}z/\Delta \ .$$

Overall, for any $z \leqslant \Delta/16\sqrt{n}$, with probability larger than $1-e^{-s}$,

$$Q_K^{(n)}(f,z) \leqslant 16\frac{\sqrt{n}}{\Delta}\sqrt{\frac{D(F)}{N}} + \sqrt{\frac{2s}{K}} + 4g(n) - 0.06\frac{\sqrt{n}z}{\Delta} \ .$$

As a conclusion, one can pick

$$z_+ = 300\sqrt{\frac{D(F)}{N}} + 20\Delta\left(\sqrt{\frac{2s}{N}} + 4\frac{g(n)}{\sqrt{n}}\right) \ ,$$

since, by assumption, this quantity is smaller than $\Delta/16\sqrt{n}$. □

Chapter 4

Multivariate mean estimation

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$. Let $\mathcal{P}_2$ denote the set of probability distributions on $\mathbb{R}^d$ such that

$$P[\|X\|^2] < \infty \ .$$

For any $P \in \mathcal{P}_2$, denote by

$$\mu_P = PX \in \mathbb{R}^d, \qquad \Sigma_P = P[(X-\mu_P)(X-\mu_P)^T] \in \mathbb{R}^{d\times d} \ .$$

The goal of the chapter is to build estimators of μ_P based on an i.i.d. sample of P, $\mathcal{D}_N = (X_1, \ldots, X_N)$, with deviations bounded for all $P \in \mathcal{P}_2$ by those of the empirical mean $\widehat{\mu}_e = N^{-1}\sum_{i=1}^N X_i$ when the vectors X_i are Gaussian. To compute the deviation bounds in the Gaussian case, we need to define the trace of Σ_P, $\mathrm{Tr}(\Sigma_P)$ and its largest eigenvalue $\|\Sigma_P\|_{\mathrm{op}}$.

Example: Least-squares density estimation The multivariate mean estimation problem is highly connected to a particular instance of unsupervized learning where one wants to recover, from an i.i.d. sample $X_1, \ldots, X_N$ taking values in a measurable space $\mathcal{X}$, the distribution P of X. Assume that P has density $\bar{f}$ with respect to a known reference measure μ, so recovering P is equivalent to recover $\bar{f}$. Assume that $\bar{f} \in L^2(\mu)$. To estimate $\bar{f}$, choose an orthonormal basis $(\varphi_i)_{i\in\mathbb{N}}$ of $L^2(\mu)$. The function $\bar{f}$ can be decomposed onto this basis $\bar{f} = \sum_{i\in\mathbb{N}} \beta_i\varphi_i$ (the convergence of the series being in $L^2(\mu)$-sense). Moreover, the coefficient β_i in this decomposition is the inner product in $L^2(\mu)$ between φ_i and $\bar{f}$, $\beta_i = \int \varphi_i \bar{f} \mathrm{d}\mu$, that is, $\beta_i = P[\varphi_i]$. Overall

$$\bar{f} = \sum_{i\in\mathbb{N}} P[\varphi_i]\varphi_i \ .$$

The projection method proceeds by cutting the sum and estimate the projections of $\bar{f}$ onto finite dimensional subspaces:

$$\bar{f}_d = \sum_{i=1}^{d} P[\varphi_i]\varphi_i \ .$$

Estimating $\bar{f}_d$ is then equivalent to estimate the vector

$$\mu_P = P\mathbf{X}, \qquad \text{where} \qquad \mathbf{X} = \begin{bmatrix} \varphi_1(X) \\ \vdots \\ \varphi_d(X) \end{bmatrix} \in \mathbb{R}^d \ .$$

The 2-moment assumption is equivalent to the assumption that $P[\varphi_i^2] < \infty$ for all $i \in \{1, \dots, d\}$. It is a weaker requirement than the connection between L^∞ and L^2-norms that is made to analyse the empirical mean estimator of μ_P, $\forall \mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}^T\mathbf{X}\|_\infty \leqslant L\sqrt{d}\|\mathbf{a}^T\mathbf{X}\|_{L^2(\mu)}$, which, as $\varphi_1, \dots, \varphi_d$ is an orthonormal system in $L^2(\mu)$ reduces to $\|\mathbf{a}^T\mathbf{X}\|_\infty \leqslant L\sqrt{d}\|\mathbf{a}\|$.

4.1 Deviations of the empirical mean in the Gaussian case

In order to establish a relevant benchmark, start by computing the deviations of the empirical mean in the Gaussian case.

Theorem 44 (Hanson-Wright)**.** *If the dataset $\mathcal{D}_N = (X_1, \dots, X_N)$ is a collection of i.i.d. Gaussian vectors with common distribution $\mathrm{N}(\mu, \Sigma)$, the empirical mean $\widehat{\mu}_e = N^{-1}\sum_{i=1}^N X_i$ satisfies*

$$\forall t > 0, \qquad \mathbb{P}\left(\|\widehat{\mu}_e - \mu\| > \sqrt{\frac{Tr(\Sigma)}{N}} + \sqrt{\frac{2\|\Sigma\|_{op} t}{N}} \right) \leqslant e^{-t} \ .$$

Proof. Let $\mathbf{S} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$ and, for any $\mathbf{u} \in \mathbf{S}$, let $X_\mathbf{u} = \mathbf{u}^T\widehat{\mu}_e$. The random variables $X_\mathbf{u}$ are Gaussian with expectation $\mu_\mathbf{u} = \mathbf{u}^T\mu$ and variance $\sigma_\mathbf{u}^2 = \mathbf{u}^T\Sigma\mathbf{u}/N$. It follows that

$$\sigma^2 = \sup_{\mathbf{u}\in\mathbf{S}} \sigma_\mathbf{u}^2 = \frac{\|\Sigma\|_{\mathrm{op}}}{N} \ . \tag{4.1}$$

Moreover,

$$\|\widehat{\mu}_e - \mu\| = \sup_{\mathbf{u}\in\mathbf{S}} \mathbf{u}^T(\widehat{\mu}_e - \mu) = \sup_{\mathbf{u}\in\mathbf{S}} (X_\mathbf{u} - \mu_\mathbf{u}) \ .$$

It comes from the concentration theorem for suprema of Gaussian processes that

$$\forall t > 0, \qquad \mathbb{P}\left(\|\widehat{\mu}_e - \mu\| > \mathbb{E}[\|\widehat{\mu}_e - \mu\|] + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}} t}{N}} \right) \leqslant e^{-t} \ .$$

Now, by Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E}[\|\widehat{\mu}_e - \mu\|] &\leqslant \sqrt{\mathbb{E}\left[\|\widehat{\mu}_e - \mu\|^2\right]} = \sqrt{\frac{1}{N^2} \sum_{1\leqslant i,j\leqslant N} \mathbb{E}[(X_i - \mu)^T(X_j - \mu)]} \\ &= \sqrt{\frac{1}{N}\mathbb{E}[(X - \mu)^T(X - \mu)]} \\ &= \sqrt{\frac{1}{N}\mathrm{Tr}(\mathbb{E}[(X - \mu)(X - \mu)^T])} = \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} \ . \end{aligned}$$

□

4.2 A first glimpse at minmax strategies

Recall that $\mathbf{S}$ denotes the unit sphere in $\mathbb{R}^d$: $\mathbf{S} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$. The proof of Hanson-Wright theorem is based on the following representation of the risk:

$$\|\widehat{\mu}_e - \mu\| = \sup_{\mathbf{u}\in\mathbf{S}} \mathbf{u}^T(P_N X - \mu_P) \ .$$

As P_N is linear, this can be rewritten

$$\|\widehat{\mu}_e - \mu\| = \sup_{\mathbf{u}\in\mathbf{S}} \{P_N[\mathbf{u}^T X] - P[\mathbf{u}^T X]\} \ .$$

The risk bound is then based on the fact the empirical estimators $P_N[\mathbf{u}^T X]$ of the univariate expectations $P[\mathbf{u}^T X]$ have uniform deviations over the sphere $\mathbf{S}$. Therefore, there are three ingredients to prove Hanson-Wright's inequality:

(i) build estimators $\widehat{P}[\mathbf{u}^T X]$ of the *univariate* expectations $P[\mathbf{u}^T X]$,

(ii) bound the deviations of $|\widehat{P}[\mathbf{u}^T X] - P[\mathbf{u}^T X]|$ uniformly over the unit sphere $\mathbf{S}$,

(iii) deduce from the collection $\{\widehat{P}[\mathbf{u}^T X],\ \mathbf{u} \in \mathbf{S}\}$ an estimator of μ_P.

In Chapter 2, we presented various constructions that can be used to estimate the univariate expectations with sub-Gaussian guarantee when $P \in \mathcal{P}_2$, therefore, extending step (i) will not be difficult. In Chapter 3, we showed uniform deviation bounds for these estimators that will be sufficient to extend step (ii). Step (iii) is obvious for the empirical mean since, by linearity, $P_N[\mathbf{u}^T X] = \mathbf{u}^T P_N[X]$. However, none of the "robust" estimators presented in Chapter 2 is linear. Therefore, extending step (iii) requires a new idea. A first idea, that appeared independently in various works such as [12, 42] for example, is to consider the minmax estimator

$$\widehat{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu - \hat{P}[\mathbf{u}^T X]| \ .$$

If the minimum is not achieved, then $\widehat{\mu}$ in this definition can be replaced by any $\widehat{\mu}_N$ satisfying

$$\sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\widehat{\mu}_N - \hat{P}[\mathbf{u}^T X]| \leqslant \inf_{\mu\in\mathbb{R}^d} \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu - \hat{P}[\mathbf{u}^T X]| + \frac{1}{N} \ .$$

This would not affect the results of this section.

A very nice feature of this construction is that this estimator has a risk bounded from above by the uniform deviations of $\widehat{P}[\mathbf{u}^T X]$ around $P[\mathbf{u}^T X]$. Actually, using successively the representation of the Euclidean norm as a supremum, the triangle inequality and the definition of $\widehat{\mu}$, it holds

$$\begin{aligned}\|\widehat{\mu} - \mu_P\| &= \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T(\widehat{\mu} - \mu_P)| \leqslant \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\widehat{\mu} - \hat{P}[\mathbf{u}^T X]| + \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu_P - \hat{P}[\mathbf{u}^T X]| \\ &\leqslant 2\sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu_P - \hat{P}[\mathbf{u}^T X]| = 2\sup_{\mathbf{u}\in\mathbf{S}} \{|(P - \hat{P})[\mathbf{u}^T X]|\} \ .\end{aligned}$$

We deduce from these remarks the following result.

Lemma 45. *For any $\boldsymbol{u} \in \boldsymbol{S}$, let $\widehat{P}[\boldsymbol{u}^T X]$ denote an estimator of the univariate expectation $P[\boldsymbol{u}^T X]$. On the event Ω_r where these estimators have uniform deviations bounded from above by r,*

$$\Omega_r = \left\{ \sup_{\boldsymbol{u} \in \boldsymbol{S}} \{|(P - \hat{P})[\boldsymbol{u}^T X]|\} \leqslant r \right\} \ , \tag{4.2}$$

the minmax estimator

$$\widehat{\mu} \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} \sup_{\boldsymbol{u} \in \boldsymbol{S}} |\boldsymbol{u}^T \mu - \hat{P}[\boldsymbol{u}^T X]| \ ,$$

satisfies $\|\widehat{\mu} - \mu_P\| \leqslant 2r$.

Lemma 45 shows that the risk of minmax estimators is bounded from above by $2r$ on the event Ω_r. To show that the risk of the minmax estimator is bounded by $2r$ with high probability, it is therefore sufficient to compute r such that Ω_r has high probability. As a first example, consider the case where $\hat{P}[\mathbf{u}^T X] = \mathrm{MOM}_K\big[\mathbf{u}^T X\big]$. The following result is a corollary of the concentration theorem for suprema of MOM processes given in Theorem 39.

Theorem 46. *Let $P \in \mathcal{P}_2$, $K \in \{1, \ldots, N\}$ and*

$$r_K = 128\sqrt{\frac{Tr(\Sigma_P)}{N}} \vee 4\sqrt{\frac{2\|\Sigma_P\|_{op} K}{N}} \ .$$

Then,

$$\mathbb{P}\bigg(\sup_{\boldsymbol{u} \in \boldsymbol{S}} |MOM_K\big[\boldsymbol{u}^T X\big] - P[\boldsymbol{u}^T X]| > r_K \bigg) \leqslant e^{-K/32} \ . \tag{4.3}$$

In particular, the minmax MOM estimator

$$\widehat{\mu} \in \operatorname{argmin}_{\mu \in \mathbb{R}^d} \sup_{\boldsymbol{u} \in \boldsymbol{S}} |\boldsymbol{u}^T \mu - MOM_K\big[\boldsymbol{u}^T X\big]|$$

satisfies

$$\mathbb{P}(\|\widehat{\mu} - \mu_P\| \leqslant 2r_K) \geqslant 1 - e^{-K/32} \ .$$

Proof. The second result comes from (4.3) and Lemma 45. To prove (4.3), consider the class of functions $F = \{\mathbf{u}^T \cdot, \ \mathbf{u} \in \mathbf{S}\}$. By (3.18), this class satisfies $D(F) \leqslant \mathrm{Tr}(\Sigma)$. Moreover, for any $\mathbf{u} \in \mathbf{S}$,

$$\mathrm{Var}(\mathbf{u}^T X) = \mathbf{u}^T \Sigma_P \mathbf{u} \leqslant \|\Sigma_P\|_{\mathrm{op}} \ . \tag{4.4}$$

Therefore, Theorem 39 shows (4.3). □

As a second application, consider smoothed MOM estimators.

Theorem 47. *Assume that there exists a known constant v such that $v \geqslant \sqrt{\|\Sigma_P\|_{op}}$, let ρ denote the function defined in* (3.19) *and consider the estimator $\hat{P}_K[\boldsymbol{u}^T X]$ to be the solution of*

$$\sum_{k=1}^{K} \rho\bigg(\sqrt{\frac{N}{K}} \frac{P_{B_k}[\boldsymbol{u}^T X] - z}{v} \bigg) = 0 \ .$$

Let

$$r_{K,s} = \sqrt{\frac{Tr(\Sigma_P)}{N}} + v\left(\sqrt{\frac{s}{N}} + g(N/K)\sqrt{\frac{K}{N}}\right) \ .$$

Then, there exists an absolute constant $C > 0$ such that, if

$$K \geqslant C\left(\frac{Tr(\Sigma_P)}{v^2} \vee g(N/K)^2\right) \ ,$$

then, for any $s \leqslant K/C$,

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\boldsymbol{S}} |\hat{P}_K[\boldsymbol{u}^T X] - P[\boldsymbol{u}^T X]| > C r_{K,s}\right) \leqslant e^{-s} \ . \tag{4.5}$$

In particular, the minmax smooth-MOM estimator

$$\widehat{\mu} \in \mathrm{argmin}_{\mu\in\mathbb{R}^d} \sup_{\boldsymbol{u}\in\boldsymbol{S}} |\boldsymbol{u}^T\mu - \hat{P}_K[\boldsymbol{u}^T X]|$$

satisfies, for any $s \leqslant K/C$,

$$\mathbb{P}(\|\widehat{\mu} - \mu_P\| \leqslant 2C r_{K,s}) \geqslant 1 - e^{-s} \ .$$

Remark 48. *Theorem 47 improves upon Theorem 46 as it shows, for example, that, when $g(N/K) \leqslant \gamma\sqrt{K/N}$, the minmax estimator $\widehat{\mu}$ based on smoothed MOM preliminary estimates with $K = \sqrt{N}$ is sub-Gaussian at any level $t \lesssim \sqrt{N}$. On the other hand, this improvement holds under L^3/L^2 comparison to bound $g(N/K)$ and requires the knowledge of an upper bound v on $\|\Sigma_P\|_{op}$.*

Proof. The second result comes from (4.5) and Lemma 45. Eq (4.5) comes from Theorem 42 applied to the class F of linear functionals $F = \{\mathbf{u}^T\cdot, \ \mathbf{u} \in \mathbf{S}\}$. For this class of functions, $D(F) = \mathrm{Tr}(\Sigma_P)$, see (3.18) and $\sigma^2 = \|\Sigma_P\|_{\mathrm{op}} \leqslant v$, see Eq (4.4). □

4.3 Working with other norms

Suppose here that one wants to estimate μ_P and that we measure the risk of an estimator $\widehat{\mu}$ by $|\widehat{\mu} - \mu_P|_*$, where $|\cdot|_*$ denote the dual norm of a norm $|\cdot|$ in $\mathbb{R}^d$. In this section, denote the sphere of the norm $|\cdot|$ by

$$\mathbf{S} = \{u \in \mathbb{R}^d : |u| = 1\} \ .$$

Recall that the dual norm $|\cdot|_*$ is defined, for any $v \in \mathbb{R}^d$, by

$$|v|_* = \sup_{u\in\mathbf{S}} u^T v \ .$$

Let $\mathbf{S}_*$ denote the unit sphere for the dual norm $\mathbf{S}_* = \{v \in \mathbb{R}^d : |v|_* = 1\}$. The construction of the previous section naturally extends to this framework. Define the estimator

$$\widehat{\mu} \in \mathrm{argmin}_{\mu\in\mathbb{R}^d} \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu - \hat{P}[\mathbf{u}^T X]| \ .$$

The triangle inequality and the definition of $\widehat{\mu}$ show that

$$|\widehat{\mu} - \mu_P|_* = \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T(\widehat{\mu} - \mu_P)| \leqslant \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\widehat{\mu} - \hat{P}[\mathbf{u}^T X]| + \sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu_P - \hat{P}[\mathbf{u}^T X]|$$
$$\leqslant 2\sup_{\mathbf{u}\in\mathbf{S}} |\mathbf{u}^T\mu_P - \hat{P}[\mathbf{u}^T X]| = 2\sup_{\mathbf{u}\in\mathbf{S}} \{|(P - \hat{P})[\mathbf{u}^T X]| \ .$$

Assume that $\hat{P}[u^T X] = \mathrm{MOM}_K\left[u^T X\right]$ for any $u \in \mathbf{S}$. We have

$$\sup_{u\in\mathbf{S}} \mathrm{Var}(u^T X) = \sup_{u\in\mathbf{S}} u^T\Sigma u = \|\Sigma\|_{**} \ .$$

Moreover, the Rademacher complexity of the class F of linear functions $x \mapsto u^T x$, for all $u \in \mathbf{S}$, can be computed as follows

$$\sqrt{D(F)} = \mathbb{E}\left[\sup_{u\in\mathbf{S}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i^T u\right] = \mathbb{E}\left[\left|\frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i\right|_*\right] \ .$$

Proceeding as in Lemma 45 yields the following result.

Lemma 49. *Let $|\cdot|$ denote a norm on $\mathbb{R}^d$, let $\boldsymbol{S}$ denote the unit sphere for $|\cdot|$ and let $|\cdot|_*$ denote the dual norm of $|\cdot|$. The minmax estimator*

$$\widehat{\mu} \in \mathrm{argmin}_{\mu\in\mathbb{R}^d} \sup_{\boldsymbol{u}\in\boldsymbol{S}} |\boldsymbol{u}^T\mu - MOM_K\left[\boldsymbol{u}^T X\right]| \ ,$$

satisfies, with probability larger than $1 - e^{-K/32}$,

$$|\widehat{\mu} - \mu_P|_* \leqslant 128\mathbb{E}\left[\left|\frac{1}{N} \sum_{i=1}^N \epsilon_i X_i\right|_*\right] \vee 4\sqrt{\frac{2\|\Sigma\|_{**}K}{N}}$$

Example: error rates in ℓ_1-norm Assume that

$$|u| = \|u\|_\infty = \max_{i\in\{1,\ldots,d\}} |u_i|$$

so $|u|_* = \|u\|_1 = \sum_{i=1}^d |u_i|$. As the term $\|\Sigma\|_{**} = \sup_{\mathbf{u}:\|\mathbf{u}\|_\infty\leqslant 1} \mathbf{u}^T\Sigma\mathbf{u}$ would also appear in the concentration of the empirical mean using Theorem 30, it is sufficient to bound the main term in Lemma 49. We have

$$\mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^N \epsilon_i(X_i - \mu_P)\right\|_1\right] = \sum_{j=1}^d \mathbb{E}\left[\left|\frac{1}{N} \sum_{i=1}^N \epsilon_i(X_{i,j} - \mu_{P,j})\right|\right] \ .$$

Applying Cauchy-Schwarz inequality, we get

$$\mathbb{E}\left[\left|\frac{1}{N} \sum_{i=1}^N \epsilon_i(X_{i,j} - \mu_{P,j})\right|\right] \leqslant \sqrt{\frac{1}{N^2} \sum_{i=1}^N \mathbb{E}[(X_{i,j} - \mu_{P,j})^2]} = \sqrt{\frac{\mathrm{Var}(X_{1,j})}{N}} \ .$$

Hence,

$$\mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^N \epsilon_i X_i\right\|_1\right] \leqslant \frac{\sum_{j=1}^d \sqrt{\mathrm{Var}(X_{i,j})}}{\sqrt{N}} \ .$$

This is, up to multiplicative numerical constant the order of $\mathbb{E}[\|X - \mu_P\|_1]$ when $X \sim \mathrm{N}(\mu_P, \Sigma_P/N)$

Example: rates in sup-norm Assume that $|u| = \|u\|_1 = \sum_{i=1}^d |u_i|$ so $|u|_* = \|u\|_\infty = \max_{i\in\{1,\dots,d\}} |u_i|$. For any $u \in \mathbf{S}$,

$$u^T\Sigma u \leqslant \max_{1\leqslant i\leqslant d} \Big|\sum_{j=1}^d \Sigma_{i,j}u_j\Big| \leqslant \max_{1\leqslant i,j\leqslant d} |\Sigma_{i,j}| \ .$$

By Cauchy-Schwarz inequality,

$$\max_{1\leqslant i,j\leqslant d} |\Sigma_{i,j}| = \max_{1\leqslant i\leqslant d} \Sigma_{i,i} \ .$$

As this upper bound is achieved for u a vector in the canonical basis, it follows that

$$\|\Sigma\|_{**} = \|\Sigma\|_\infty \ .$$

The main term

$$\mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_{i,j}\Big|\Big] \ ,$$

can be evaluated using higher moment assumptions on $X_{i,j}$. Assume that $\mathbb{E}[|X_{i,j}|^p] < \infty$, for some $p \geqslant 2$. Then, for any $q \in \{2,\dots,p\}$, Pisier's trick applies and gives

$$\begin{aligned}\mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_{i,j}\Big|\Big] &\leqslant \Big(\mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_{i,j}\Big|^q\Big]\Big)^{1/q} \\ &\leqslant \frac{1}{N}\Big(\sum_{j=1}^d \mathbb{E}\Big[\Big|\sum_{i=1}^N \epsilon_i X_{i,j}\Big|^q\Big]\Big)^{1/q}\end{aligned}$$

Now, apply Khinchine's inequality on moments of order p for sums of independent random variables, see for examples [10, Chapter 15]. It shows that

$$\Big(\mathbb{E}\Big[\Big|\sum_{i=1}^N \epsilon_i X_{i,j}\Big|^q\Big]\Big)^{1/q} \leqslant 3\sqrt{q}\,\mathbb{E}\Big[\Big(\sum_{i=1}^N X_{i,j}^2\Big)^{q/2}\Big]^{1/q} \ .$$

Then, by convexity of $x \mapsto x^{q/2}$,

$$\Big(\sum_{i=1}^N X_{i,j}^2\Big)^{q/2} = N^{q/2}\Big(\frac{1}{N}\sum_{i=1}^N X_{i,j}^2\Big)^{q/2} \leqslant N^{q/2-1}\sum_{i=1}^N |X_{i,j}|^q \ .$$

Therefore,

$$\mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_{i,j}\Big|\Big] \leqslant \frac{3\sqrt{q}}{N^{1/2+1/q}}\Big(\sum_{j=1}^d\sum_{i=1}^N \mathbb{E}\big[|X_{i,j}|^q\big]\Big)^{1/q} \ .$$

As X_i are i.i.d., this bound reduces to

$$\mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_{i,j}\Big|\Big] \leqslant \frac{3\sqrt{q}}{N^{1/2}}\Big(\sum_{j=1}^d \mathbb{E}\big[|X_{1,j}|^q\big]\Big)^{1/q} \ . \tag{4.6}$$

We have proved the following result:

Theorem 50. *Assume that there exists $p \geqslant 2$ such that $\mathbb{E}[|X_{i,j}|^p] < \infty$ and, for any $q \leqslant p$, let $M_q = \left(\sum_{j=1}^d \mathbb{E}\left[\left|X_{1,j}\right|^q\right]\right)^{1/q}$. Let $\boldsymbol{S}_1$ denote the sphere for the ℓ_1-norm on $\mathbb{R}^d$. The estimator*

$$\widehat{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\boldsymbol{u}\in\boldsymbol{S}_1} \left|\boldsymbol{u}^T\mu - MOM_K\left[\boldsymbol{u}^T X\right]\right| \ ,$$

satisfies, with probability larger than $1 - e^{-K/32}$,

$$\|\widehat{\mu} - \mu_P\|_\infty \leqslant \frac{384}{\sqrt{N}} \inf_{q\leqslant p}(\sqrt{q}M_q) \vee 4\sqrt{\frac{2\|\Sigma\|_\infty K}{N}} \ .$$

Remark 51. *Since $p \geqslant 2$, $\inf_{q\leqslant p}(\sqrt{q}M_q) \leqslant \sqrt{2}M_2 = \sqrt{2\,Tr(\Sigma)}$.*

If the coordinates $X_{1,j}$ have a finite moment of order $p \geqslant 2\log d$. Denote by $C_d = \max_{j\in\{1,\dots,d\}} \mathbb{E}[|X_{1,j}|^{2\log d}]^{1/(2\log d)}$. We have, for any $q \leqslant 2\log d$, $M_q \leqslant C_d d^{1/q}$, hence

$$\inf_{q\leqslant p}(\sqrt{q}M_q) \leqslant C_d\sqrt{2\log d}d^{1/(2\log d)} = C_d\sqrt{2e\log d} \ .$$

In this case, the optimal sub-Gaussian inequality is therefore recovered only under stronger moment assumption on the vectors X_i.

4.4 PAC-Bayesian analysis

Applying the minmax strategy with MOM (or smoothed MOM) estimators $\mathrm{MOM}_K\left[\mathbf{u}^T X\right]$ of univariate expectations yields estimators $\widehat{\mu}$ of PX with sub-Gaussian tails but the constants involved in the deviation property are a bit loose compared to the Gaussian case of Hanson-Wright theorem. As for univariate mean estimation, there exist alternatives with much better performance from this perspective. The material of this section is borrowed from [12]. Let ψ denote a function such that

$$\forall t \in \mathbb{R}, \qquad -\log(1 - t + t^2/2) \leqslant \psi(t) \leqslant \log(1 + t + t^2/2) \ . \tag{4.7}$$

For example, one can verify that the following function satisfies (4.7),

$$\psi(t) = \begin{cases} -2\sqrt{2}/3 & \text{if } \ t < -\sqrt{2} \\ t - t^3/6 & \text{if } \ t \in [-\sqrt{2}, \sqrt{2}] \\ 2\sqrt{2}/3 & \text{if } \ t > \sqrt{2} \end{cases} \ .$$

Let $\lambda > 0$, $\beta > 0$, $\mathbf{I}_d$ denote the identity matrix in $\mathbb{R}^d$ and, for any $\mathbf{u} \in \mathbf{S}$, $\rho_{\mathbf{u}} = \mathrm{N}(u, \beta\mathbf{I}_d)$. Define the estimators of univariate expectations $P[\mathbf{u}^T X]$ as

$$\hat{P}_{\lambda,\beta}[\mathbf{u}^T X] = \frac{1}{N\lambda}\sum_{i=1}^N \int \psi(\lambda\mathbf{v}^T X_i)\mathrm{d}\rho_{\mathbf{u}}(\mathbf{v}) \ .$$

These new estimators are not translation invariant which means that, if b denotes a deterministic quantity, one cannot guarantee that $\hat{P}_{\lambda,\beta}[\mathbf{u}^T X + b] = \hat{P}_{\lambda,\beta}(\mathbf{u}^T X) + b$. In particular, $\hat{P}_{\lambda,\beta}[\mathbf{u}^T X] - P[\mathbf{u}^T X]$ may not be equal to $\hat{P}_{\lambda,\beta}[\mathbf{u}^T X - P[\mathbf{u}^T X]]$. Therefore, the analysis of these estimators is a bit more tricky than for MOM. The following result shows the deviation properties of the minmax estimators based on the preliminary estimates $\hat{P}_{\lambda,\beta}[\mathbf{u}^T X]$.

Theorem 52. *Assume that there exist* known *constants T and v, $v \leqslant T$, such that the matrix $\overline{\Sigma}_P = P[XX^T]$ satisfies*

$$Tr(\overline{\Sigma}_P) \leqslant T, \qquad \|\overline{\Sigma}_P\|_{op} \leqslant v \ .$$

Let λ and β denote the following quantities

$$\lambda = \sqrt{\frac{2\log(\delta^{-1})}{Nv}}, \qquad \beta = \frac{1}{\lambda\sqrt{NT}} = \sqrt{\frac{v}{2T\log(1/\delta)}} \ .$$

Then, with probability at least $1-\delta$,

$$\sup_{u\in\boldsymbol{S}} |\hat{P}_{\lambda,\beta}[\boldsymbol{u}^T X] - P[\boldsymbol{u}^T X]| \leqslant \sqrt{\frac{T}{N}} + \sqrt{\frac{2v\log(1/\delta)}{N}} \ . \tag{4.8}$$

In particular, the minmax estimator

$$\widehat{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\boldsymbol{u}\in\boldsymbol{S}} |\boldsymbol{u}^T\mu - \hat{P}_{\lambda,\beta}[\boldsymbol{u}^T X]|$$

satisfies

$$\mathbb{P}\left(\|\widehat{\mu} - \mu_P\| \leqslant 2\left(\sqrt{\frac{T}{N}} + \sqrt{\frac{2v\log(1/\delta)}{N}}\right)\right) \geqslant 1-\delta \ .$$

Proof. The second result comes from (4.8) and Lemma 45. Let us focus on proving (4.8). Fix $\mu = \rho_0$. Let $\Gamma : \mathbf{S}\times\mathbb{R}^d \to \mathbb{R}$, $(\mathbf{v},\mathbf{x}) \mapsto \Gamma_{\mathbf{v}}(\mathbf{x}) = \psi(\lambda\mathbf{v}^T\mathbf{x})$, so

$$P_N\left[\int \Gamma_{\mathbf{v}} \mathrm{d}\rho_{\mathbf{u}}(\mathbf{v})\right] = \lambda\hat{P}_{\lambda,\beta}[\mathbf{u}^T X] \ .$$

Moreover, by definition of ψ, (4.7),

$$\begin{aligned}\int \log P[e^{\Gamma_{\mathbf{v}}}]\mathrm{d}\rho_{\mathbf{u}}(\mathbf{v}) &= \int \log P[e^{\psi(\lambda\mathbf{v}^T X)}]\mathrm{d}\rho_{\mathbf{u}}(\mathbf{v}) \\ &\leqslant \int \log\left(1 + \lambda P[\mathbf{v}^T X] + \frac{\lambda^2}{2}P[(\mathbf{v}^T X)^2]\right)\mathrm{d}\rho_{\mathbf{u}}(\mathbf{v}) \ .\end{aligned}$$

As $\log(1+x) \leqslant x$ for any $x > -1$,

$$\begin{aligned}\int \log\left(1 + \lambda P[\mathbf{v}^T X] + \frac{\lambda^2}{2}P[(\mathbf{v}^T X)^2]\right)\mathrm{d}\rho_{\mathbf{u}}(\mathbf{v}) \\ \leqslant \lambda P\left[\int \mathbf{v}^T X \mathrm{d}\rho_{\mathbf{u}}(\mathbf{v})\right] + \frac{\lambda^2}{2}P\left[\int (\mathbf{v}^T X)^2 \mathrm{d}\rho_{\mathbf{u}}(\mathbf{v})\right] \ .\end{aligned}$$

Conditionally on X, when $\mathbf{v}$ is distributed as $\rho_{\mathbf{u}}$, $\lambda\mathbf{v}^T X$ is distributed according to a Gaussian $\mathrm{N}(\lambda\mathbf{u}^T X, \beta\lambda^2\|X\|^2)$. Hence,

$$\begin{aligned}\int \log\left(1 + \lambda P[\mathbf{v}^T X] + \frac{\lambda^2}{2}P[(\mathbf{v}^T X)^2]\right)\mathrm{d}\rho_{\mathbf{u}}(\mathbf{v}) \\ &\leqslant \lambda\left(P\big[\mathbf{u}^T X\big] + \frac{\lambda}{2}P\big[(\mathbf{u}^T X)^2 + \beta\|X\|^2\big]\right) \\ &\leqslant \lambda\left(P\big[\mathbf{u}^T X\big] + \frac{\lambda}{2}\big(v + \beta T\big)\right) \ .\end{aligned}$$

Moreover, as $K(\rho_{\mathbf{u}}, \rho_0) = 1/(2\beta)$, it follows from the PAC-Bayesian inequality, see Theorem 36, that, with probability $1-\delta$, for any $\mathbf{u} \in \mathbf{S}$,

$$\hat{P}_{\lambda,\beta}[\mathbf{u}^T X] \leqslant P[\mathbf{u}^T X] + \frac{\lambda}{2}\big(v + \beta T\big) + \frac{(1/2\beta) + \log(1/\delta)}{\lambda N} \ . \tag{4.9}$$

The choice of parameters now ensures that

$$\begin{aligned}
\frac{\lambda v}{2} &= \frac{1}{2}\sqrt{\frac{2v\log(1/\delta)}{N}} \ , \\
\frac{\lambda \beta T}{2} &= \frac{1}{2}\sqrt{\frac{T}{N}} \ , \\
\frac{1}{2\beta\lambda N} &= \frac{1}{2}\sqrt{\frac{T}{N}} \ , \\
\frac{\log(1/\delta)}{\lambda N} &= \frac{1}{2}\sqrt{\frac{2v\log(1/\delta)}{N}} \ .
\end{aligned}$$

Plugging these estimates into (4.9) shows that

$$\mathbb{P}\bigg(\sup_{\mathbf{u}\in\mathbf{S}} (\hat{P}_{\lambda,\beta}[\mathbf{u}^T X] - P[\mathbf{u}^T X]) \leqslant \sqrt{\frac{T}{N}} + \sqrt{\frac{2v\log(1/\delta)}{N}} \bigg) \geqslant 1-\delta \ .$$

As $\mathbf{S}$ is symmetric, Eq (4.8) is proved and therefore the theorem is established. □

The problem with Theorem 52 is that it involves upper bounds on the L^2 moments $\overline{\Sigma}_P$ rather than on the covariance matrix Σ_P. Fortunately, there is a simple trick to deduce from the estimator $\widehat{\mu}$ an estimator with sub-Gaussian deviations based on the actual covariance matrix Σ_P.

Theorem 53. *Assume that there exist* known *constants $\bar{T}$, $\bar{v}$ and b such that*

$$Tr(\Sigma_P) \leqslant \bar{T}, \qquad \|\Sigma_P\|_{op} \leqslant \bar{v}, \qquad \|\mu_P\|^2 \leqslant b \ . \tag{4.10}$$

Fix $\delta \in (0,1)$ and let

$$A := 4\bigg(\sqrt{\bar{T} + b} + \sqrt{2(\bar{v}+b)\log(1/\delta)} \bigg)^2 \ .$$

For any $k \in \{1, \dots, N-1\}$, there exists an estimator $\widehat{\mu}$ of μ such that

$$\mathbb{P}\bigg(\|\widehat{\mu} - \mu_P\| \leqslant 2\bigg(\sqrt{\frac{\bar{T} + A/k}{N-k}} + \sqrt{\frac{2(\bar{v}+A/k)\log(1/\delta)}{N-k}} \bigg) \bigg) \leqslant 1 - 2\delta \ .$$

Remark 54. *If d is fixed and $N \to \infty$, one can choose $k \asymp \sqrt{N}$ to deduce that, for any $\epsilon > 0$, there exists N_0 such that, for any $N \geqslant N_0$, there exists an estimator $\widehat{\mu}$ of μ such that*

$$\mathbb{P}\bigg(\|\widehat{\mu} - \mu_P\| \leqslant (2+\epsilon)\bigg(\sqrt{\frac{\bar{T}}{N}} + \sqrt{\frac{2\bar{v}\log(1/\delta)}{N}} \bigg) \bigg) \leqslant 1 - 2\delta \ .$$

Therefore, $\widehat{\mu}$ achieves much better constants than the minmax estimator based on MOM preliminary estimators, see Theorem 46 or on smoothed MOM estimators, see Theorem 47. Actually, these constants are asymptotically not worse than twice the optimal constants of the Hanson-Wright theorem. On the other hand, the knowledge of upper bounds on both $Tr(\Sigma_P)$ and $\|\Sigma_P\|_{op}$ is mandatory for the construction of these estimators.

Proof. Under Assumption (4.10), the constants T and v satisfy the requirements of Theorem 52, where

$$T \geqslant \bar{T} + b, \qquad v \geqslant \bar{v} + b \ .$$

To build an estimator with correct sub-Gaussian deviations, split the sample in two parts $(X_1, \dots, X_k)$ and $(X_{k+1}, \dots, X_N)$. With the first sample $X_1, \dots, X_k$, build the minmax estimator $\bar{\mu}$ of μ_P of Theorem 52 with the constants T and v. According to Theorem 52, $\mathbb{P}(\Omega_1) \geqslant 1 - \delta$, where

$$\Omega_1 = \left\{ \|\bar{\mu} - \mu_P\| \leqslant \sqrt{\frac{A}{k}} \right\} \ .$$

The estimator $\widehat{\mu}$ will be a minmax estimator of μ_P based on Theorem 52, using the sample $(X_{k+1} - \bar{\mu}, \dots, X_N - \bar{\mu})$. To choose appropriate constants T' and v' in this theorem, one has to bound the L^2-moments of the sample $(X_{k+1} - \bar{\mu}, \dots, X_N - \bar{\mu})$. The idea is to work conditionally on $\mathcal{F}_k$, the σ-algebra generated by $X_1, \dots, X_k$. It holds

$$P[(X - \bar{\mu})(X - \bar{\mu})^T | \mathcal{F}_k] = \Sigma_P + (\bar{\mu} - \mu_P)(\bar{\mu} - \mu_P)^T \ .$$

In particular,

$$\begin{aligned} \mathrm{Tr}[P[(X - \bar{\mu})(X - \bar{\mu})^T | \mathcal{F}_k]] &\leqslant \mathrm{Tr}(\Sigma_P) + \|\bar{\mu} - \mu_P\|^2 \ , \\ \|P[(X - \bar{\mu})(X - \bar{\mu})^T | \mathcal{F}_k]\|_{\mathrm{op}} &\leqslant \|\Sigma_P\|_{\mathrm{op}} + \|\bar{\mu} - \mu_P\|^2 \ . \end{aligned}$$

This bound cannot be used to build $\widehat{\mu}$ but it holds

$$\begin{aligned} \mathrm{Tr}[P[(X - \bar{\mu})(X - \bar{\mu})^T | \mathcal{F}_k, \Omega_1]] &\leqslant \mathrm{Tr}(\Sigma_P) + \frac{A}{k} \ , \\ \|P[(X - \bar{\mu})(X - \bar{\mu})^T | \mathcal{F}_k, \Omega_1]\|_{\mathrm{op}} &\leqslant \|\Sigma_P\|_{\mathrm{op}} + \frac{A}{k} \ . \end{aligned}$$

This suggests to build the minmax estimator using Theorem 52, based on the sample $(X_{k+1} - \bar{\mu}, \dots, X_N - \bar{\mu})$, with the constants $\bar{T} + A/k$ and $\bar{v} + A/k$. Let

$$r = 2\left(\sqrt{\frac{\bar{T} + A/k}{N}} + \sqrt{\frac{2(\bar{v} + A/k)\log(1/\delta)}{N}} \right) \ ,$$

According to these preliminary computations, it holds

$$\mathbb{P}\left(\left\{ \|\widehat{\mu} - \mu_P\| > 2\left(\sqrt{\frac{\bar{T} + A/k}{N}} + \sqrt{\frac{2(\bar{v} + A/k)\log(1/\delta)}{N}} \right) \right\} \middle| \mathcal{F}_k, \Omega_1 \right) \leqslant \delta \ .$$

Therefore,

$$\begin{aligned} \mathbb{P}\big(\|\widehat{\mu} - \mu_P\| > r\big) &\leqslant 1 - \mathbb{P}(\Omega_1) + \mathbb{P}\big(\Omega_1 \cap \big\{\|\widehat{\mu} - \mu_P\| > r\big\}\big) \\ &\leqslant \delta + \mathbb{E}\big[\mathbb{P}\big(\Omega_1 \cap \big\{\|\widehat{\mu} - \mu_P\| > r\big\} \big| \mathcal{F}_k\big)\big] \leqslant 2\delta \ . \end{aligned}$$

□

4.5 Toward a generic minmax strategy

This section introduces a minmax strategy that can be extended more easily than the one presented in Section 4.2 to any learning problem where ERM can be used. The starting point of this generic construction is that the multivariate expectation μ_P is solution of a minimization problem where the objective function is a univariate expectation

$$\mu_P \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} P[\|X-\mu\|^2] \ .$$

A first "natural" idea to build an estimator from this formulation would be to consider

$$\widehat{\mu}_{\mathrm{nat}} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \mathrm{MOM}_K\big[\|X-\mu\|^2\big] \ .$$

This construction would be similar to that of the empirical mean $\widehat{f}_{\mathrm{emp}}$ which is the ERM associated to the loss $\|f-x\|^2$, i.e. $\widehat{f}_{\mathrm{emp}} \in \operatorname{argmin}_{f\in F} P_N[\|X-f\|^2]$. It turns out that min MOM estimators have suboptimal deviation bounds even under stronger assumption on F, see for example [35]. The reason is that the lack of linearity of the median prevents from using localization ideas that yields optimal deviation rates of the ERM. Instead of simply minimizing MOM, the idea is to reformulate the problem in order to build an estimator based on estimators of the expectations of the *increments* of loss rather than on the losses themselves. To clarify this idea, remark that, by linearity of P, the target μ_P is also solution of the following minmax problem:

$$\mu_P \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\nu\in\mathbb{R}^d} P[\|X-\mu\|^2 - \|X-\nu\|^2] \ .$$

Now, one can obtain an estimator of μ_P simply by plugging in this formulation estimators $\widehat{P}[\|X-\mu\|^2-\|X-\nu\|^2]$ of the univariate expectations $P[\|X-\mu\|^2-\|X-\nu\|^2]$. Contrary to the previous minmax strategy that used the specific form of the risk function, this new construction is completely generic and can be extended to many learning tasks. In the remaining of this section, we present as a first example of application an analysis of the minmax MOM estimator

$$\widehat{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\nu\in\mathbb{R}^d} \mathrm{MOM}_K\big[\|X-\mu\|^2 - \|X-\nu\|^2\big] \ .$$

The minmax MOM estimator differs from the min MOM since the median is not linear. In the following chapters, we shall extend the analysis of minmax MOM estimators to much more general learning tasks and show sub-Gaussian oracle inequalities in several classical problems.

Theorem 55. *Let $X_1,\dots,X_N$ denote i.i.d. realizations of a distribution $P\in\mathcal{P}_2$. Let $K\leqslant N$ and let*

$$\widehat{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\nu\in\mathbb{R}^d} MOM_K\big[\|X-\mu\|^2 - \|X-\nu\|^2\big] \ .$$

Then

$$\mathbb{P}\bigg(\|\widehat{\mu}-\mu_P\| > (\sqrt{2}+1)\bigg(128\sqrt{\frac{Tr(\Sigma_P)}{N}} \vee 4\sqrt{\frac{\|\Sigma_P\|_{op}K}{N}}\bigg)\bigg) \leqslant e^{-K/32} \ .$$

Remark 56. *Remark that the constants here are slightly worse than for the first minmax strategy presented in Theorem 46.*

Proof. Define for any $\mu \in \mathbb{R}^d$ its score

$$S(\mu) = \sup_{\nu \in \mathbb{R}^d} \mathrm{MOM}_K\big[\|X-\mu\|^2 - \|X-\nu\|^2\big] \ .$$

On one hand, we have

$$S(\widehat{\mu}) \leqslant S(\mu_P) = \sup_{\nu \in \mathbb{R}^d} \mathrm{MOM}_K\big[\|X-\mu_P\|^2 - \|X-\nu\|^2\big] \ .$$

On the other hand, for any $\nu \in \mathbb{R}^d$,

$$\begin{aligned} S(\nu) &\geqslant \mathrm{MOM}_K\big[\|X-\nu\|^2 - \|X-\mu_P\|^2\big] \\ &= -\mathrm{MOM}_K\big[\|X-\mu_P\|^2 - \|X-\nu\|^2\big] \ . \end{aligned} \tag{4.11}$$

This suggests to analyse the process

$$\{\mathrm{MOM}_K\big[\|X-\mu_P\|^2 - \|X-\nu\|^2\big],\ \nu \in \mathbb{R}^d\} \ .$$

As, for any $\nu \in \mathbb{R}^d$,

$$\|X-\mu_P\|^2 - \|X-\nu\|^2 = 2(X-\mu_P)^T(\nu-\mu_P) - \|\nu-\mu_P\|^2 \ ,$$

we have

$$\mathrm{MOM}_K\big[\|X-\mu_P\|^2 - \|X-\nu\|^2\big] = 2\|\nu-\mu_P\|\mathrm{MOM}_K\left[(X-\mu_P)^T\frac{\nu-\mu_P}{\|\nu-\mu_P\|}\right] - \|\nu-\mu_P\|^2 \ . \tag{4.12}$$

Therefore, it is sufficient to analyse the process

$$\{\mathrm{MOM}_K\big[(X-\mu_P)^T\mathbf{u}\big],\ \mathbf{u} \in \mathbf{S}\}, \qquad \mathbf{S} = \{\mathbf{u} \in \mathbb{R}^d : \|u\| = 1\} \ .$$

Recall that in Theorem 46, we showed that, for

$$r_K = 128\sqrt{\frac{\mathrm{Tr}(\Sigma_P)}{N}} \vee 4\sqrt{\frac{\|\Sigma_P\|_{\mathrm{op}}K}{N}} \ ,$$

then,

$$\mathbb{P}\left(\sup_{\mathbf{u}\in\mathbf{S}} |\mathrm{MOM}_K\big[\mathbf{u}^TX\big] - P[\mathbf{u}^TX]| > r_K\right) \leqslant e^{-K/32} \ .$$

Let $\Omega = \{\forall \mathbf{u} \in \mathbf{S},\ \sup_{\mathbf{u}\in\mathbf{S}} \mathrm{MOM}_K\big[(X-\mu_P)^T\mathbf{u}\big] \leqslant r_K\}$, so $\mathbb{P}(\Omega) \geqslant 1 - e^{-K/32}$. On Ω, by (4.12),

$$S(\mu_P) \leqslant \sup_{a\in\mathbb{R}}\{2ar_K - a^2\} \leqslant r_K^2 \ .$$

Therefore, by definition of $\widehat{\mu}$, $S(\widehat{\mu}) \leqslant r_K^2$. On the other hand, on Ω, by (4.11) and (4.12), for any $\nu \in \mathbb{R}^d$,

$$S(\nu) \geqslant -2\|\nu-\mu_P\|r_K + \|\nu-\mu_P\|^2$$

In particular, therefore, on Ω,

$$-2\|\widehat{\mu}-\mu_P\|r_K + \|\widehat{\mu}-\mu_P\|^2 \leqslant r_K^2 \ .$$

Solving this inequality shows that, on Ω,

$$\|\widehat{\mu}-\mu_P\| \leqslant (\sqrt{2}+1)r_K \ .$$

This concludes the proof of the theorem. □

4.6 Resistance to outliers

In this section, consider the $O \cup I$ framework where $(X_i)_{i\in I}$ are independent random variables such that

$$\forall i \in I, \quad \mathbb{E}[X_i] = \mu_P, \quad \mathbb{E}[(X_i - \mu)(X_i - \mu)^T] = \Sigma_P \ .$$

No assumption is granted on the outliers $(X_i)_{i\in O}$. Denote by $\epsilon = |O|/N$.

4.6.1 Resistance of MOM estimators

This section investigate minmax MOM estimator in this framework.

Theorem 57. *Assume that $K \geqslant 20N\epsilon/9$. Denote by $\widehat{f}_K$ minmax MOM estimator*

$$\widehat{\mu}_K \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\nu\in\mathbb{R}^d} MOM_K\left[\|X-\mu\|^2 - \|X-\nu\|^2\right] \ .$$

Then there exists an absolute constant C such that, with probability at least $1 - e^{-K/C}$,

$$\|\widehat{\mu}_K - \mu\|^2 \leqslant C\left(\frac{Tr(\Sigma)}{N} \vee \frac{\|\Sigma\|_{op}K}{N}\right) \ .$$

Proof. The proof uses intensively results obtained in the proof of Theorem 55. Proceeding as in this proof, denote, for any $\xi \in \mathbb{R}^d$, by

$$S(\xi) = \sup_{\nu\in\mathbb{R}^d} \mathrm{MOM}_K\left[\|X-\xi\|^2 - \|X-\nu\|^2\right] \ ,$$

and recall that $S(\mu) \leqslant R_K^2$, where

$$R_K = \sup_{\mathbf{u}\in\mathbf{S}} \mathrm{MOM}_K\left[\mathbf{u}^T(X-\mu)\right] \ .$$

Denote by $\mathcal{K}$ the indexes of blocks $B_k \subset I$. It is clear that $|\mathcal{K}| \geqslant K - N\epsilon$. Applying the general version of Lugosi and Mendelson concentration bound for median-of-means processes, with probability at least $1 - e^{-(K-N\epsilon)/c^*}$, there exists at leat $9(K-N\epsilon)/10 \geqslant K/2$ blocks $B_k \subset I$ where

$$\forall \mathbf{u} \in \mathbf{S}, \qquad P_{B_k}[\mathbf{u}^T(X-\mu)] \leqslant c^*\left(\sqrt{\frac{20\mathrm{Tr}(\Sigma)}{9N}} \vee \sqrt{\frac{\|\Sigma\|_{\mathrm{op}}K}{N}}\right) \ .$$

This implies in particular that, with probability at least $1 - e^{-K/2c^*}$,

$$r_K \leqslant c^*\left(\sqrt{\frac{20\mathrm{Tr}(\Sigma)}{9N}} \vee \sqrt{\frac{\|\Sigma\|_{\mathrm{op}}K}{N}}\right) \ .$$

The proof terminates as the one of Theorem 55. □

Remark 58. *The condition $K \gtrsim N\epsilon$ implies that the convergence rate of the minmax MOM estimator $\widehat{\mu}_K$ is bounded from above by*

$$C\left(\frac{Tr(\Sigma)}{N} \vee \frac{\|\Sigma\|_{op}K}{N} \vee \|\Sigma\|_{op}\epsilon\right) \ .$$

In particular, these rates match those obtained on clean datasets in Theorem 55 as long as $\epsilon \lesssim r(\Sigma)/N$, where $r(\Sigma)$ is the effective rank of Σ, $r(\Sigma) = Tr(\Sigma)/\|\Sigma\|_{op}$.

4.6.2 Depth

The purpose of this section is to investigate optimality of the proportion of outliers $\epsilon \lesssim r(\Sigma)/N$ allowed by MOM estimators by comparing with the Gaussian case. The material of this section is an adaptation of results obtained in [15]. Assume that inliers have Gaussian distribution $P_I = \mathrm{N}(\mu, \sigma^2 \mathbf{I}_d)$. Consider the following Gaussian $O \cup I$ framework where the dataset $\mathcal{D}_N$ contains $|I|$ data $(X_i)_{i\in I}$ i.i.d. with common distribution P_I and $|O|$ outliers $(X_i)_{i\in O}$ that can be anything. Tuckey's depth (hereafter called depth) of any $\nu \in \mathbb{R}^d$ relatively to a distribution $\mathbb{P}$ on $\mathbb{R}^d$ is defined by

$$D(\nu, \mathbb{P}) = \inf_{\mathbf{u}\in\mathbf{S}} \mathbb{P}(\mathbf{u}^T X \leqslant \mathbf{u}^T \nu) \ .$$

Tuckey's depth (hereafter called depth) of any $\nu \in \mathbb{R}^d$ relatively to the dataset $\mathcal{D}_N$ on $\mathbb{R}^d$ is the empirical version of $D(\nu, \mathbb{P})$:

$$D(\nu, \mathcal{D}_N) = \inf_{\mathbf{u}\in\mathbf{S}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{\mathbf{u}^T X_i \leqslant \mathbf{u}^T \nu\}} \ .$$

In other words, $D(\nu, \mathcal{D}_N)$ is Tuckey's depth relative to the empirical measure P_N. Tuckey's median is the deepest point in $\mathbb{R}^d$, that is

$$\widehat{\mu}_{\mathrm{Tuc}} \in \mathrm{argmax}_{\nu\in\mathbb{R}^d} D(\nu, \mathcal{D}_N) \ .$$

The purpose of this section is to establish the following result.

Theorem 59. *Denote by $\widehat{\mu}_{Tuc}$ Tuckey's median. Assume the Gaussian $O \cup I$ framework, denote by $\epsilon = |O|/N$. There exist absolute constants C_1, C_2 such that, for any $\delta \in (0,1)$ satisfying $C_1(d + \epsilon^2 + \log(1/\delta))/N < 1$,*

$$\mathbb{P}\left(\|\widehat{\mu}_{Tuc} - \mu\|^2 \leqslant C_2 \left(\frac{d}{N} + \epsilon^2 + \frac{\log(1/\delta)}{N} \right) \right) \ .$$

Remark 60. *In this example, the covariance matrix Σ of the inliers is the identity matrix $\boldsymbol{I}_d$, so $Tr(\Sigma) = d$, $\|\Sigma\|_{op} = 1$ and the effective rank $r(\Sigma) = d$. It follows that the rates of the convergence of MOM estimators in the clean case are not downgraded if the proportion of outliers $\epsilon \lesssim d/N$. It comes from Theorem 59 that Tuckey's median tolerates much outliers since a proportion $\epsilon \lesssim \sqrt{d/N}$ is allowed here. Of course, the result for Tuckey's median only holds when inliers are Gaussian and it provides the optimal sub-Gaussian dependence on the covariance matrix of X only in the case where this covariance is bounded from bellow by the identity. These conditions are way more restrictive than those required for MOM estimators. Moreover, one can show that the dependence $\epsilon \leqslant d/N$ is optimal if we allow inliers with heavier tails than Gaussian. Nevertheless, this shows that the number of outliers allowed by minmax MOM estimators is not optimal in general and opens an interesting question: is there some estimator achieving optimal sub-Gaussian deviation bounds assuming only that $P \in \mathcal{P}_2$ and whose dependency in the number of outliers is always optimal?*

Proof. Assume, without loss of generality, that $\mu = 0$. Define, for any $\mathbf{u} \in \mathbf{S}$ and any $\nu \in \mathbb{R}^d$, the half space $H_{\mathbf{u},\nu} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{u}^T \mathbf{x} \leqslant \mathbf{u}^T \nu\}$. Define

$$P_N^{(I)}(H_{u,\nu}) = \frac{1}{|I|} \sum_{i\in I} \mathbf{1}_{\{X_i \in H_{u,\nu}\}} \ .$$

The set of Half spaces $H_{u,\nu}$ is the set of all affine half spaces in $\mathbb{R}^d$, it has Vapnik-Chervonenkis VC dimension $d+1$. It comes from standard VC theory, see [58] that there exists an absolute constant C such that, with probability larger than $1-\delta$,

$$\sup_{\mathbf{u}\in\mathbf{S},\nu\in\mathbb{R}^d}(P_N^{(I)}-P_I)(H_{u,\nu}) \leqslant C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

This result implies that

$$\sup_{\mathbf{u}\in\mathbf{S},\nu\in\mathbb{R}^d}(D(\nu,(X_i)_{i\in I})-D(\nu,P_I)) \leqslant C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) . \tag{4.13}$$

In particular,

$$D(\widehat{\mu}_{\mathrm{Tuc}},P_I) \geqslant D(\widehat{\mu}_{\mathrm{Tuc}},(X_i)_{i\in I}) - C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Now, for any $\nu\in\mathbb{R}^d$,

$$\begin{aligned} D(\nu,(X_i)_{i\in I}) &= \inf_{\mathbf{u}\in\mathbf{S}}\frac{1}{|I|}\sum_{i\in I}\mathbf{1}_{\{\mathbf{u}^TX_i\leqslant\mathbf{u}^T\nu\}} \\ &\geqslant \frac{N}{|I|}D(\nu,\mathcal{D}_N)-\frac{|O|}{|I|} \\ &= \frac{N}{|I|}(D(\nu,\mathcal{D}_N)-\epsilon) \geqslant \frac{1}{1-\epsilon}(D(\nu,\mathcal{D}_N)-\epsilon) . \end{aligned}$$

Hence,

$$D(\widehat{\mu}_{\mathrm{Tuc}},P_I) \geqslant \frac{1}{1-\epsilon}(D(\widehat{\mu}_{\mathrm{Tuc}},\mathcal{D}_N)-\epsilon) - C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

By definition of $\widehat{\mu}_{\mathrm{Tuc}}$, this implies that

$$D(\widehat{\mu}_{\mathrm{Tuc}},P_I) \geqslant \frac{1}{1-\epsilon}(D(\mu,\mathcal{D}_N)-\epsilon) - C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

As for any $\nu\in\mathbb{R}^d$, $ND(\nu,\mathcal{D}_N) \geqslant |I|D(\nu,(X_i)_{i\in I})$, this implies $D(\nu,\mathcal{D}_N) \geqslant (1-\epsilon)D(\nu,(X_i)_{i\in I})$, thus

$$D(\widehat{\mu}_{\mathrm{Tuc}},P_I) \geqslant D(\mu,(X_i)_{i\in I}) - \frac{\epsilon}{1-\epsilon} - C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Applying (4.13) one more time shows that, with probability larger than $1-2\delta$,

$$D(\widehat{\mu}_{\mathrm{Tuc}},P_I) \geqslant D(\mu,P_I) - \frac{\epsilon}{1-\epsilon} - 2C\left(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Introduce Φ, the c.d.f of the standard Gaussian distribution on $\mathbb{R}$, $\mathrm{N}(0,1)$. It holds that, for any $\nu\in\mathbb{R}^d$,

$$D(\nu,P_I) = \inf_{u\in\mathbb{R}^d}P_I(u^TX\leqslant u^T\nu) = \inf_{u\in\mathbb{R}^d}\Phi(u^T\nu) = 1-\Phi(\|\nu\|) .$$

In particular, as $\mu = 0$, $D(\mu, P_I) = 1/2$, so

$$D(\widehat{\mu}_{\text{Tuc}}, P_I) = 1 - \Phi(\|\widehat{\mu}_{\text{Tuc}}\|) \geqslant \frac{1}{2} - \frac{\epsilon}{1-\epsilon} - 2C\left(\sqrt{\frac{d}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Equivalently, with probability larger than $1 - 2\delta$,

$$\Phi(\|\widehat{\mu}_{\text{Tuc}}\|) \leqslant \frac{1}{2} + \frac{\epsilon}{1-\epsilon} + 2C\left(\sqrt{\frac{d}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Now, the proof terminates since there exists an absolute constant c such that $\Phi(x) \geqslant 1/2 + x/4$ for any $0 < x < c$. □

Chapter 5

The homogeneity lemma

The homogeneity lemma is one of the most important tools in these notes. Roughly speaking, it allows to reduce the analysis of minmax estimators to deviation bounds of the underlying process on localized classes of functions. It is an alternative to the peeling argument that has been repeatedly used in the analysis of the ERM to benefit from localization ideas [31] and prove fast rates of convergence in statistical learning theory. It is particularly well adapted to problems where deviation inequalities are only available up to a certain confidence parameter, as it is the case of MOM processes, see Section 3.5 in Chapter 3. The version presented here is an extension of the "deterministic argument" presented in [18] that allows to deal with convex losses as will be done in Chapters 6 and 7 and with the tests of ρ-estimation presented in Chapter 8.

5.1 Learning, ERM, minmax aggregation of tests

Consider the statistical learning framework of Vapnik. Let Z denote a random variable taking values in a measurable space $\mathcal{Z}$, with distribution P. Let F denote a set of parameters and let $\ell : F \times \mathcal{Z} \to \mathbb{R}$, $(f, z) \mapsto \ell_f(z)$ denote a function called loss. Assume that there exists $f_0 \in F$ such that, for any $f \in F$, $\ell_f(\cdot) - \ell_{f_0}(\cdot) \in L^1(P)$. Under this assumption, $\ell_f - \ell_g \in L^1(P)$ for any $f, g \in F$. We want to estimate

$$f^* \in \operatorname{argmin}_{f \in F} P[\ell_f - \ell_{f_0}] \ . \tag{5.1}$$

It is clear that, for any $g \in F$, we also have $f^* \in \operatorname{argmin}_{f \in F} P[\ell_f - \ell_g]$. The arguably most simple example of such problem is multivariate mean estimation where one wants to estimate the expectation μ_P of a measure P on $\mathbb{R}^d$. In this example, let $\mathcal{Z} = F = \mathbb{R}^d$, $\|\cdot\|$ denote the Euclidean norm and $\ell : (f, z) \mapsto \|f - z\|^2$, $f_0 = 0$, then $\ell_f(z) - \ell_{f_0}(z) = \|f\|^2 - 2f^T z \in L^1(P)$ when $Z \in L^1(P)$ and $P[\ell_f - \ell_{f_0}] = \|f - \mu_P\|^2 - \|\mu_P\|^2$ is obviously minimized when $f = \mu_P$, that is $f^* = \mu_P$.

To estimate f^*, a dataset $Z_1, \dots, Z_N$ i.i.d. with common distribution P is available. Let P_N denote the empirical measure of the sample $Z_1, \dots, Z_N$ defined for any function $g : \mathcal{Z} \to \mathbb{R}$ by $P_N g = N^{-1} \sum_{i=1}^N g(Z_i)$ One way to handle problem (5.1) is to use Empirical Risk Minimizers defined by

$$\hat{f}^{\mathrm{ERM}} \in \operatorname{argmin}_{f \in F} P_N[\ell_f] \ .$$

For multivariate mean estimation, this yields for example, the empirical mean estimator $\hat{f}^{\mathrm{ERM}} = N^{-1}\sum_{i=1}^{N} Z_i$. This estimator is not robust to heavy-tailed data, or the presence of outliers in the dataset.

To build robust alternative, the empirical mean could be replaced by any robust estimator seen in Chapter 1 in the mean problem to get a robust estimator for learning task. As for multivariate mean estimation considered in Chapter 4, this strategy is suboptimal in general. Instead, following the strategy introduced in Section 4.5 of Chapter 4, one can rewrite the min problem (5.1) as follows

$$f^* \in \mathrm{argmin}_{f\in F} P[\ell_f - \ell_{f_0}] = \mathrm{argmin}_{f\in F} \sup_{g\in F} P[\ell_f - \ell_g] \ . \tag{5.2}$$

Then, one can plug into this definition any robust estimator of the increments $P[\ell_f - \ell_g]$. For example, the minmax MOM estimator is defined by

$$\hat{f}_K^{\mathrm{MOM}} \in \mathrm{argmin}_{f\in F} \sup_{g\in F} \mathrm{MOM}_K[\ell_f - \ell_g] \ .$$

The ERM could also be be obtained this way since

$$\hat{f}^{\mathrm{ERM}} \in \mathrm{argmin}_{f\in F} \sup_{g\in F} P_N[\ell_f - \ell_g] \ .$$

Notice also that min MOM and minmax MOM estimators differ in general since MOM processes are not linear.

The ideas that we develop in this chapter intend to analyse the following extension of these minmax strategies. The building blocks of the general construction are *tests statistics* or increment estimators which are random variables $T(f,g)$ where f and g belong to F. For ERM, $T(f,g) = P_N[\ell_f - \ell_g]$ and for minmax MOM estimators $T(f,g) = \mathrm{MOM}_K[\ell_f - \ell_g]$. In both examples, $T(f,g)$ is an estimator of $P[\ell_f - \ell_g]$. An other example is presented later in Section 5.2.4. As this property is satisfied in all examples, it is always assumed that $T(f,g) = -T(g,f)$ and in particular that $T(f,f) = 0$ for any $f \in F$. The heuristic is that $T(f,g) > 0$ means that g is better than f to estimate f^*. The estimator we want to analyse is the minmax estimator

$$\hat{f} \in \mathrm{argmin}_{f\in F} \sup_{g\in F} T(f,g) \ . \tag{5.3}$$

To analyse this estimator, we introduce an evaluation function $\mathcal{E}$. Formally, $\mathcal{E} : F \to \mathbb{R}$ denotes a real valued function such that, for any $f \in F$, $\mathcal{E}(f)$ evaluates the performance of f as an estimator of f^*. As this evaluation function is not involved in the definition of $\hat{f}$, it may perfectly depend on the unknown distribution P. In the examples, $\mathcal{E}$ will usually denote the *excess risk* $\mathcal{E}(f) = P[\ell_f - \ell_{f^*}]$ (which is non negative by definition and obviously null if $f = f^*$) or some distance between f and f^*. In any case, large values of $\mathcal{E}(f)$ indicate that f is *far* from the target f^*, so f is not a desirable estimator of f^*. The evaluation function is used to defined a geometry on F and we introduce, for any $r > 0$,

$$\mathbf{B}(r) = \{f \in F : \mathcal{E}(f) \leqslant r\}, \qquad \mathbf{S}(r) = \{f \in F : \mathcal{E}(f) = r\} \ .$$

We will also need a *centering* function d, which is a function $d : F \times F \to \mathbb{R}$ satisfying $d(f,g) = -d(g,f)$ for any f and g in F. In Vapnik's learning setting, $d(f,g)$ will denote $P(\ell_f - \ell_g)$ but other functions can be considered as in Chapter 8.

5.2 General results

This section gathers the main results of this chapter. The goal is to reduce the analysis of minmax estimators to concentration inequalities for suprema of test processes $\sup_{f\in\mathbf{B}(r)} T(f^*,f)$ for suitably calibrated balls $\mathbf{B}(r)\subset F$ localized around the oracle f^*. When applied to the tests $T(f,g)=P_N[\ell_f-\ell_g]$ defining ERM, these results extend well known localization ideas widely used to prove fast rates of convergence for this estimator, see for example [31] for an overview on this topic.

5.2.1 Link with multiple testing theory

The first result extracts the idea underlying the proof of the risk bound for minmax MOM estimator of multivariate expectations in Theorem 55. Interestingly, it establishes a link between learning or estimation from tests, which is the analysis of estimators built as in (5.3) and multiple testing theory, which is an extension of the classical theory of tests in statistics where one is interested in testing several null hypotheses at the same time.

Lemma 61. *[Link with Multiple Testing] Let Θ and r denote positive real numbers. Let Ω denote the event where the following equations hold.*

$$\sup_{g\in F} T(f^*,g)\leqslant\Theta\ , \tag{5.4}$$

$$\sup_{g\notin\boldsymbol{B}(r)} T(f^*,g)<-\Theta\ . \tag{5.5}$$

Then, on Ω, the minmax estimator $\hat{f}$ defined in (5.3) satisfies $\mathcal{E}(\hat{f})\leqslant r$.

Proof. By definition of $\hat{f}$, on Ω, by (5.4).

$$\sup_{g\in F} T(\hat{f},g)\leqslant\sup_{g\in F} T(f^*,g)\leqslant\Theta\ .$$

On the other hand, by (5.5), any f such that $\mathcal{E}(f)>r$ satisfies

$$\sup_{g\in F} T(f,g)\geqslant T(f,f^*)>\Theta\ .$$

As a consequence, $\mathcal{E}(\hat{f})\leqslant r$ on Ω. □

Conditions (5.4) and (5.5) are intuitively clear. (5.4) means that the tests between f^* and any $g\in F$ should not become too large ($\Theta=0$ in (5.4) for the ideal test $T(f,g)=P[\ell_f-\ell_g]$). Θ controls typically the fluctuations of the process $\sup_{g\in F} T(f^*,g)$ which has negative drift. (5.5) means that f^* is preferred to any g with large drift with a margin larger than the noise level.

Let us clarify in which sense Lemma 61 makes a link with multiple testing theory. This link uses the formalism and definitions borrowed from [21]. Let $\mathcal{P}$ denote a class of probability distributions on $\mathcal{Z}$ and for any $P\in\mathcal{P}$, denote by

$$F_P^*=\operatorname{argmin}_{f\in F} P[\ell_f-\ell_{f_0}]\subset F\ .$$

Define, for any $f\in F$, the hypothesis $H_f=\{P\in\mathcal{P}: f\in F_P^*\}\subset\mathcal{P}$. We want to test simultaneously all assumptions $\mathcal{H}=\{H_f,\ f\in F\}$. A multiple test of

$\mathcal{H}$ is a random subset $\mathcal{R} \subset \mathcal{H}$ of *rejected* hypotheses. To evaluate the multiple testing procedure $\mathcal{R}$, introduce, for any $P \in \mathcal{P}$, the sets

$$\mathcal{F}(P) = \{H_f \in \mathcal{H} : \ f \notin F_P^*\}, \qquad \mathcal{F}_r(P) = \{H_f \in \mathcal{H} : \ f \notin \mathbf{B}(r)\} \ .$$

$\mathcal{F}(P)$ is called the set of *false* hypotheses and $\mathcal{F}_r(P)$ is the set of assumptions that are r-separated from the true assumptions $\mathcal{T}(P) = \mathcal{H} \setminus \mathcal{F}(P)$. The *family-wise error rate* (FWER) of the multiple testing $\mathcal{R}$ is defined by

$$\mathrm{FWER}(\mathcal{R}) := \sup_{P \in \mathcal{P}} P\big(\mathcal{R} \cap \mathcal{T}(P) \neq \emptyset\big) = 1 - \inf_{P \in \mathcal{P}} P(\mathcal{R} \subset \mathcal{F}(P)) \ .$$

It is the (maximal) probability to reject at least one true hypothesis. To understand this definition, consider the situation where $\mathcal{H}$ is reduced to a singleton $\mathcal{H} = \{H\}$. In this case, one can consider a simple test ϕ_H of the assumption H against the complementary $H^c = \mathcal{P}/H$: $\phi_H = 1$ means that H is rejected and $\phi = 0$ that H is not rejected. The multiple test associated with the simple test ϕ_H is $\mathcal{R} = \{H : \phi_H = 1\}$. In words, $\mathcal{R} = \{H\}$ if H is rejected by the simple test ϕ_H and $\mathcal{R} = \emptyset$ if H is not rejected by ϕ_H. In this case, $\mathrm{FWER}(\mathcal{R})$ is the size of the simple test ϕ_H. In particular, ϕ_H has level α iff $\mathrm{FWER}(\mathcal{R}) \leqslant \alpha$. In this sense, $\mathrm{FWER}(\mathcal{R})$ is an extension of the first type error rate for simple tests. The *family-wise separation rate* (FWSR) of the test $\mathcal{R}$ is defined by

$$\mathrm{FWSR}_\beta(\mathcal{R}) = \inf\{r > 0 : \inf_{P \in \mathcal{P}} P(\mathcal{R} \supset \mathcal{F}_r(P)) \geqslant 1 - \beta\} \ .$$

FWSR measures the minimal distance between H_f and $\mathcal{T}(P)$ such that H_f is rejected with confidence level β. FWSR extends the notion of *separation rates* for simple tests, see [5] for a definition of separation rates and [21] for more details on this extension. FWSR is a measure of the second type error rate for multiple testing that allows to define a minimax theory for these tests.

Going back to learning from tests, one can use the family of test statistics $T(f, g)$ to build a multiple testing on $\mathcal{H}$. The idea is to use the score

$$\hat{\mathcal{E}}(f) = \sup_{g \in F} T(f, g)$$

as a test statistic to build a simple test of the assumption H_f. Small values of $\hat{\mathcal{E}}(f)$ indicate that H_f might be true and large values that it seems false. This suggests to consider, for some threshold $\Theta > 0$, the multiple testing

$$\mathcal{R}_\Theta = \{H_f \in \mathcal{H} : \hat{\mathcal{E}}(f) > \Theta\} \ .$$

This test satisfies $\mathrm{FWER}(\mathcal{R}_\Theta) \leqslant \alpha$ if, for any $P \in \mathcal{P}$, $\mathcal{R}_\Theta$ does not contain any $H_{f_P^*}$, where $f_P^* \in F_P^*$, with P-probability larger than α. In words, to bound the FWER, we have to bound from bellow the probability that any $f_P^* \in F_P^*$ is not rejected, that is the probability of the event

$$\sup_{g \in F} T(f_P^*, g) \leqslant \Theta \ .$$

Bounding from above the FWER of the multiple testing $\mathcal{R}_\Theta$ by α is equivalent to bound from bellow the probability of the event (5.4) by $1 - \alpha$.

Consider now the FWSR of $\mathcal{R}_\Theta$. This FWSR is bounded by r if the probability that any $f \in F$ such that $\mathcal{E}(f) > r$ is rejected with probability at least

$1-\beta$. Formally, $\mathrm{FWSR}_\beta(\mathcal{R}_\Theta) \leqslant r$ if, for any $P \in \mathcal{P}$, $P(\Omega_{\Theta,r}) \geqslant 1-\beta$, where $\Omega_{\Theta,r}$ is the event

$$\forall f \in F : \mathcal{E}(f) > r, \qquad \sup_{g \in F} T(f,g) > \Theta \ . \tag{5.6}$$

Remark that $\Omega_{\Theta,r}$ clearly contains the event $\Omega'_{\Theta,r}$ defined by

$$\forall f \notin \mathbf{B}(r), \qquad T(f,f^*) > \Theta \ .$$

Therefore, if (5.5) holds with probability $1-\beta$, then the FWSR of the test $\mathcal{R}_\Theta$ is bounded from above by r.

It transpires from the proof of Lemma 61 that Assumption (5.6) can replace Assumption 5.5 with the same conclusion: $\mathcal{E}(\hat{f}) \leqslant r$. Therefore, if $\mathrm{FWER}(\mathcal{R}_\Theta) \leqslant \alpha$ and $\mathrm{FWSR}_\beta(\mathcal{R}_\Theta) \leqslant r$, then $\mathcal{E}(\hat{f}) \leqslant r$ with probability $1-\alpha-\beta$ for any choice of the probability distribution $P \in \mathcal{P}$. However, besides this application, we will always use the restricted form of this result given by Lemma 61 which is why we presented this version.

5.2.2 The homogeneity lemma

The following result is the most important of this chapter and one the most fundamental tool of these notes. It is called the "homogeneity lemma" and it shows that risk bounds for minmax estimators follow from concentration of suprema of test processes over suitably calibrated balls $\mathbf{B}(r) \subset F$ *localized* around the oracle f^*. This result holds under abstract conditions on the test statistics that can easily be checked in the applications developed in the following chapters. It will be at the heart of the proofs of all risk bounds given afterwards. The idea is to show that conditions (5.4) and (5.5) in Lemma 61 are met if suprema of test processes over localized classes are controled.

Lemma 62 (homogeneity lemma)**.** *Assume that the tests $T(f,g)$ satisfy the homogeneity property.*

(HP)*: There exists $r_0 > 0$ such that, for any $r > r_0$ and any $f \notin \boldsymbol{B}(r)$, there exists $f_r \in \boldsymbol{S}(r)$ and $\alpha \geqslant 1$ such that*

$$T(f,f^*) \geqslant \alpha T(f_r,f^*) \ . \tag{5.7}$$

Let $\theta : \mathbb{R}_+ \to \mathbb{R}_+$ and $d : F^2 \to \mathbb{R}$ such that $d(f,g) = -d(g,f)$. Consider, for any $r > 0$, the event

$$\Omega_r = \left\{ \sup_{f \in \boldsymbol{B}(r)} T(f^*,f) - d(f^*,f) \leqslant \theta(r) \right\} \ .$$

Assume that there exists $r_1 > r_0$ such that

$$\theta(r_1) - \inf_{f \in \boldsymbol{S}(r_1)} d(f,f^*) \leqslant 0 \ , \tag{5.8}$$

Let $\zeta \geqslant \theta(r_1) + \sup_{f \in \boldsymbol{B}(r_1)} d(f^*,f)$. Assume that there exists $r_2 > r_0$ such that

$$\theta(r_2) - \inf_{f \in \boldsymbol{S}(r_2)} d(f,f^*) < -\zeta \ . \tag{5.9}$$

On the event $\Omega_{r_1} \cap \Omega_{r_2}$, (5.4) and (5.5) hold with $\Theta = \zeta$ and $r = r_2$. In particular, $\mathbb{P}(\mathcal{E}(\hat{f}) \leqslant r_2) \geqslant \mathbb{P}(\Omega_{r_1} \cap \Omega_{r_2})$.

Proof. On Ω_{r_1}, by definition, for any $f \in F$ such that $\mathcal{E}(f) \leqslant r_1$,

$$T(f^*, f) = d(f^*, f) + (T(f^*, f) - d(f^*, f)) \leqslant \zeta \ . \tag{5.10}$$

Moreover, for any $r > r_0$ and any $f \notin \mathbf{B}(r)$, there exist $f_r \in \mathbf{S}(r)$ and $\alpha \geqslant 1$ such that

$$T(f^*, f) \leqslant \alpha T(f^*, f_r) \ .$$

It follows that, for any $r > r_0$ and any $f \notin \mathbf{B}(r)$, on Ω_r,

$$\begin{aligned} T(f^*, f) &\leqslant \alpha[-d(f_r, f^*) + (T(f^*, f_r) - d(f^*, f_r))] \\ &\leqslant \alpha[-\inf_{f \in \mathbf{S}(r)} \{d(f, f^*)\} + \theta(r)] \ . \end{aligned}$$

Hence, on Ω_{r_1}, $T(f^*, f) \leqslant 0$ for any $f \notin \mathbf{B}(r_1)$, and by (5.10), (5.4) holds with $\Theta = \zeta$. Moreover, on Ω_{r_2}, for any $f \notin \mathbf{B}(r_2)$,

$$T(f^*, f) \leqslant \alpha[-\inf_{f \in \mathbf{S}(r_2)} \{d(f, f^*)\} + \theta(r_2)] < -\zeta \ .$$

Therefore, (5.5) holds with $r = r_2$. □

Remark 63. *In some applications, the set F is discrete and the requirement $\mathcal{E}(f_r) = r$ may be restrictive. The interested reader can check that a direct adaptation of the proof allows to relax slightly Condition* **(HP)** *into*

(HPr) *there exist $r_0 > 0$ and an absolute constant c such that, for any $r > r_0$ and any $f \notin \boldsymbol{B}(r)$, there exists $f_r \in F$ and $\alpha \geqslant 1$ such that*

$$\mathcal{E}(f_r) \in [cr, r], \qquad T(f, f^*) \geqslant \alpha T(f_r, f^*) \ .$$

Under this relaxed condition, the conclusion of Lemma 62 holds, with the minor modification that $\inf_{f \in \boldsymbol{S}(r)} d(f, f^*)$ *has to be replaced by* $\inf_{f \in F : \mathcal{E}(f) \in [cr, r]} d(f, f^*)$ *in the definition of r_1 and r_2. Remark that Lemma 62 is a particular instance of this extended result in the ideal case where $c = 1$.*

Remark 64. *The function d will be $d(f, g) = P[\ell_f - \ell_g]$ in learning problems. This function clearly satisfies the requirements $d(f, g) = -d(g, f)$. Moreover, in this case, $d(f^*, f) \leqslant 0$ so $\zeta = \theta(r_1)$.*

The following corollary is the typical application of the homogeneity lemma that will be used in the applications. It follows from elementary algebraic computations.

Corollary 65. *Grant the conditions of the homogeneity lemma 62 and assume moreover that there exists constants $a > 0, b < c$ and d such that, for any $r > 0$,*

$$\theta(r) \leqslant ar + br^2, \qquad \inf_{f \in \boldsymbol{S}(r)} d(f^*, f) \geqslant cr^2, \qquad \sup_{f \in \boldsymbol{B}(r)} d(f^*, f) \leqslant dr^2 \ .$$

Then (5.8) *is satisfied with $r_1 = a/(c - b)$, $\zeta = \frac{ca^2}{(c-b)^2}$ and* (5.9) *holds with $r_2 = (1 + \sqrt{\frac{c+d}{c-d}})r_1$.*

5.2.3 Convex losses

When working in Vapnik's learning framework, $\mathcal{E}(f)$ will usually denote a distance between f and f^* derived from a norm $\mathcal{E}(f) = \|f - f^*\|$. In this setting, the concentration inequalities of the previous chapter allow to bound the probability of the events Ω_r. To conclude this section, we show that some assumptions of the homogeneity lemma are met when $T(f,g) = \hat{P}[\ell_f - \ell_g]$ when the process $\hat{P}$ is positive and homogeneous and when the loss ℓ is convex.

Lemma 66 (convex losses)**.** *Assume that F is convex and that*

$$\forall z \in \mathcal{Z}, \qquad f \mapsto \ell_f(z) \text{ is convex } .$$

Assume that there exists a norm $\|\cdot\|$ such that $\mathcal{E}(f) = \|f - f^*\|$. Assume that the estimators $\hat{P}[g]$ are well defined for any real valued function g and satisfy the following requirement:

(i) $\hat{P}$ is non-decreasing: for any $g \leqslant g'$, $\hat{P}[g] \leqslant \hat{P}[g']$,

(ii) $\hat{P}$ is homogeneous: for any $a \in \mathbb{R}$, $\hat{P}[ag] = a\hat{P}[g]$,

Then the tests $T(f,g) = \hat{P}[\ell_f - \ell_g]$ satisfy the homogeneity property of Lemma 62 with $r_0 = 0$: for any $r > 0$ and any $f \notin \boldsymbol{B}(r)$, there exists $f_r \in \boldsymbol{S}(r)$ and $\alpha \geqslant 1$ such that

$$T(f, f^*) \geqslant \alpha T(f_r, f^*) \ .$$

Proof. Let $r > 0$ and $f \notin \mathbf{B}(r)$. Let $\alpha = \mathcal{E}(f)/r > 1$ and $f_r = f^* + \alpha^{-1}(f - f^*) = \alpha^{-1} f + (1 - \alpha^{-1}) f^*$. By convexity of F, $f_r \in F$. Moreover,

$$\mathcal{E}(f_r) = \|f^* - f_r\| = \left\| \frac{f^* - f}{\alpha} \right\| = \frac{\|f^* - f\|}{\alpha} = r \ .$$

Now, by the convexity assumption, for any $z \in \mathcal{Z}$,

$$\ell_{f_r}(z) \leqslant \alpha^{-1} \ell_f(z) - (1 - \alpha^{-1}) \ell_{f^*}(z) \ ,$$

hence,

$$\ell_{f_r}(z) - \ell_{f^*}(z) \leqslant \alpha^{-1} (\ell_f(z) - \ell_{f^*}(z)) \ .$$

It follows that

$$T(f, f^*) = \hat{P}[\ell_f - \ell_{f^*}] \geqslant \hat{P}[\alpha(\ell_{f_r} - \ell_{f^*})] = \alpha \hat{P}[\ell_{f_r} - \ell_{f^*}] = \alpha T(f_r, f^*) \ .$$

□

Examples of operator $\hat{P}$. We will use repeatedly Lemma 66 when $\hat{P}$ denotes the empirical mean P_N or the median-of-means operator $\mathrm{MOM}_K[\cdot]$. As both empirical means and the median satisfy conditions (i) and (ii) of Lemma 66, these estimators can actually safely be used when applying this lemma. It is worth noticing though that neither smoothed median-of-means nor M-estimators in general satisfy the homogeneity condition (ii).

5.2.4 The tests of ρ-estimation.

ρ-estimators have been introduced in [4] and further extended in [6]. The idea is to estimate a distribution P^* on a measurable space $\mathcal{X}$ from an i.i.d. sample $X_1, \dots, X_N$ with common distribution P^*. The risk is measured for any estimator $\hat{P}$ by the squared Hellinger distance between P and $\hat{P}$: $h^2(\hat{P}, P)$, where, for all distributions P and Q, for any measure μ such that $P \ll \mu$, $Q \ll \mu$, denoting by $p = \mathrm{d}P/\mathrm{d}\mu$ and $q = \mathrm{d}P/\mathrm{d}\mu$,

$$h^2(P,Q) = \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 \mathrm{d}\mu \ .$$

It is easy to check that $h^2(P,Q)$ is well defined for any P and Q, that it does not depend on μ and that it always satisfies

$$0 \leqslant h^2(P,Q) \leqslant 1 \ .$$

This problem does not directly falls into Vapnik's learning framework. Nevertheless, the homogeneity lemma in its general form presented in this chapter may be used in this problem. Let μ denote a measure on $\mathcal{X}$ and let F denote a closed convex set of densities with respect to μ, that is, for any $f \in F$, $f \geqslant 0$ μ-a.s. and $\int f \mathrm{d}\mu = 1$. For any $f \in F$, let also P_f denote the distribution with density f w.r.t. μ. To compare elements f and g in F, Baraud and Birgé defined in [6] the following tests:

$$T(f,g) = \sum_{i=1}^{N} \rho\left(\sqrt{\frac{g(Z_i)}{f(Z_i)}}\right) \ . \tag{5.11}$$

Here, the function $\rho = (x-1)/(x+1)$ is non-decreasing $[0, +\infty] \to [-1, 1]$, 2-Lipschitz, it satisfies $\rho(1/x) = -\rho(x)$ for any $x \in [0, +\infty)$. This last property implies that

$$T(f,g) = \sum_{i=1}^{N} \rho\left(\sqrt{\frac{g(Z_i)}{f(Z_i)}}\right) = -\sum_{i=1}^{N} \rho\left(\sqrt{\frac{f(Z_i)}{g(Z_i)}}\right) = -T(g,f) \ .$$

Hence, $T(f,g)$ are test statistics in the sense of Section 5.1. To conclude this chapter, we show that these test statistics satisfy the homogeneity property.

Lemma 67. *The ρ-test defined in* (5.11) *satisfy the homogeneity property of Lemma 62 with the evaluation function $\mathcal{E}(f) = h(P^*, P_f)$ and minimal radius $r_0 = \min_{f \in F} h(P^*, P_f)$: for any $r > r_0$ and any $f \notin \boldsymbol{B}(r)$, there exists $f_r \in \boldsymbol{S}(r)$ and $\alpha \geqslant 1$ such that*

$$T(f, f^*) \geqslant \alpha T(f_r, f^*) \ . \tag{5.12}$$

Remark 68. *In* (5.12)*, it is assumed for simplicity that $f^* \in \operatorname{argmin}_{f \in F} h(P^*, P_f)$ exists. If it does not $T(f, f^*)$ should be replaced by $\inf_{g \in F} T(f,g)$.*

Proof. Let $f^* \in F$ denote an oracle, that is a function such that $r_0 = \mathcal{E}(f^*)$ (assuming that it exists, in general take an approximating sequence). For any $r > r_0$, any $f \in F$ such that $\mathcal{E}(f) > r$ and any $\epsilon \in (0,1)$, let

$$f_\epsilon = \epsilon f^* + (1-\epsilon) f \ .$$

By convexity of F, $f_\epsilon \in F$. Moreover, if P^* is absolutely continuous with respect to μ (otherwise, one can change μ to $\mu + P^*$), and denoting by p^* its density,

$$\mathcal{E}(f_\epsilon) = \frac{1}{2}\int (\sqrt{p^*} - \sqrt{\epsilon f^* + (1-\epsilon)f})^2 \mathrm{d}\mu \ .$$

The map $\epsilon \mapsto \mathcal{E}(f_\epsilon)$ is continuous and takes value $\mathcal{E}(f) > r$ when $\epsilon = 0$, $\mathcal{E}(f^*) < r$ when $\epsilon = 1$. Therefore, there exists $\epsilon \in (0,1)$ such that $\mathcal{E}(f_\epsilon) = r$. Elementary calculus shows that, for any $a \geqslant 0$, the functions $\eta_a(x) = (2a)/(a + \sqrt{x})$ are convex. Therefore, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\rho\left(\sqrt{\frac{f^*(x)}{f_\epsilon(x)}}\right) &= \frac{\sqrt{\frac{f^*(x)}{f_\epsilon(x)}} - 1}{\sqrt{\frac{f^*(x)}{f_\epsilon(x)}} + 1} = \frac{\sqrt{f^*(x)} - \sqrt{f_\epsilon(x)}}{\sqrt{f^*(x)} + \sqrt{f_\epsilon(x)}} \\
&= \frac{2\sqrt{f^*(x)}}{\sqrt{f^*(x)} + \sqrt{f_\epsilon(x)}} - 1 \\
&= \eta_{\sqrt{f^*(x)}}(\epsilon f^*(x) + (1-\epsilon)f(x)) - 1 \\
&\leqslant \epsilon \eta_{\sqrt{f^*(x)}}(f^*(x)) + (1-\epsilon)\eta_{\sqrt{f^*(x)}}(f(x)) - 1 \\
&= (1-\epsilon)(\eta_{\sqrt{f^*(x)}}(f(x)) - 1) = (1-\epsilon)\rho\left(\sqrt{\frac{f^*(x)}{f(x)}}\right) \ .
\end{aligned}$$

It follows that

$$T(f, f^*) = \sum_{i=1}^N \rho\left(\sqrt{\frac{f^*(X_i)}{f(X_i)}}\right) \geqslant \frac{1}{1-\epsilon}\sum_{i=1}^N \rho\left(\sqrt{\frac{f^*(X_i)}{f_\epsilon(X_i)}}\right) = \frac{1}{1-\epsilon}T(f_\epsilon, f^*) \ .$$

In words, T satisfies Eq (5.7) with $\alpha = 1/(1-\epsilon)$. □

5.3 Back to multivariate mean estimation.

As a first example of application of the freshly introduced general methodology, let us go back to the problem of estimating a multivariate expectation discussed in Chapter 4.

Recall that $\|\cdot\|$ denotes the Euclidean norm on $\mathcal{Z} = \mathbb{R}^d$ and $\mathcal{P}_2$ denote the set of distributions on $\mathbb{R}^d$ such that $P[\|X\|^2] < \infty$. For any $P \in \mathcal{P}_2$, let $f_P^* = PX \in \mathbb{R}^d$ and $\Sigma_P = P[(X-\mu_P)(X-\mu_P)^T] \in \mathbb{R}^{d\times d}$. Recall that estimating f_P^* is a learning problem that falls into Vapnik's framework: let $F = \mathbb{R}^d$ and $\ell_f(z) = \|z - f\|^2$. Then, the quadratic loss satisfies the quadratic/multiplier decomposition:

$$\forall f, g \in F, \forall z \in \mathcal{Z}, \qquad \ell_f(z) - \ell_g(z) = -2(f-g)^T(z-g) + \|f-g\|^2 \ .$$

In particular,

$$\ell_f(Z) - \ell_{f_P^*}(Z) = -2(f - f_P^*)^T(Z - f_P^*) + \|f - f_P^*\|^2, \tag{5.13}$$

so

$$P[\ell_f - \ell_{f_P^*}] = \|f - f_P^*\|^2 \ . \tag{5.14}$$

Therefore,

$$\{f_P^*\} = \operatorname{argmin}_{f\in F} P\ell_f \ .$$

The loss satisfies the convexity assumption in Lemma 66. Moreover, as discussed after this lemma, the empirical mean P_N or MOM processes $\mathrm{MOM}_K[\cdot]$ satisfy conditions (i) and (ii) on the mean estimators $\hat{P}$. It follows that Lemma 66 applies to the tests $T(f,g) = \hat{P}[\ell_f - \ell_g]$. In particular, these tests satisfy the homogeneity property in the homogeneity lemma.

To deduce risk bounds for the associated minmax estimators, it remains to compute the function Θ for a choice of evaluation function $\mathcal{E}$ and pseudo-distance function d in the homogeneity lemma. By (5.14), $d(f, f_P^*) := P[\ell_f - \ell_{f_P^*}] = \|f - f_P^*\|^2$. Pick $\mathcal{E}(f) = \|f - f_P^*\|$ so, for any $r > 0$, $\inf_{f\in\mathbf{S}(r)} d(f, f_P^*) = r^2$. It follows therefore from (5.13) that

$$T(f_P^*, f) - d(f, f_P^*) = \hat{P}[2(f - f_P^*)^T(Z - f_P^*) - \|f - f_P^*\|^2] + \|f - f_P^*\|^2 \ .$$

Therefore by homogeneity and translation invariance of $\hat{P}$: for any function g and any $b \in \mathbb{R}$, $\hat{P}[g + b] = \hat{P}[g] + b$,

$$\begin{aligned} T(f_P^*, f) - d(f, f_P^*) &= 2\|f - f_P^*\|\hat{P}\left[\left(\frac{f - f_P^*}{\|f - f_P^*\|}\right)^T (Z - f_P^*)\right] \\ &\leqslant 2\|f - f_P^*\|R \ , \end{aligned} \tag{5.15}$$

where

$$R = \sup_{\mathbf{u}\in\mathbf{S}} \hat{P}[\mathbf{u}^T(Z - f_P^*)] \ , \tag{5.16}$$

where $\mathbf{S} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$.

5.3.1 ERM in the Gaussian case

Start with an application to the Gaussian case. The purpose here is to show that one can recover Hanson-Wright result (up to constants) using our general methodology.

Theorem 69 (ERM)**.** *If Z is Gaussian, the ERM $\hat{f}^{ERM} = N^{-1}\sum_{i=1}^N Z_i$ satisfies*

$$\forall s > 0, \qquad \mathbb{P}\left(\|\hat{f}^{ERM} - f_P^*\| > 4\left(\frac{\sqrt{Tr(\Sigma)} + \sqrt{2\|\Sigma\|_{op}s}}{\sqrt{N}}\right)\right) \leqslant e^{-s} \ .$$

Proof. The proof of Theorem 44 shows that, with probability at least $1 - e^{-s}$, the random variable R defined in (5.16) with $\hat{P} = P_N$ satisfies $R \leqslant r_s$, where

$$r_s = \sqrt{\frac{\mathrm{Tr}(\Sigma_P)}{N}} + \sqrt{\frac{2\|\Sigma_P\|_{\mathrm{op}}s}{N}} \ . \tag{5.17}$$

Hence, if $\theta(r) = 2r_s r$, by (5.15), all events $\{\Omega_r, r > 0\}$, where Ω_r is defined in Lemma 62 are contained in $\Omega_{\mathrm{good}} = \{R \leqslant r_s\}$. Recall that the choices $\mathcal{E}(f) = \|f - f_P^*\|$ and $d(f,g) = P[\ell_f - \ell_g]$ imply $\sup_{f\in\mathbf{B}(r)} d(f^*, f) = 0$ and

$$\inf_{f\in\mathbf{S}(r)} P[\ell_f - \ell_{f_P^*}] = \inf_{f:\|f - f_P^*\| = r} \|f - f_P^*\|^2 = r^2 \ .$$

Therefore, the conditions of Corollary 65 are satisfied with $a = 2r_s$, $b = d = 0$, $c = 1$, thus, thanks to this result, the homogeneity lemma holds with $r_1 = 2r_s$, $\zeta = 4r_s^2$, $r_2 = 4r_s$. □

5.3.2 Minmax MOM estimators

This section shows that the general methodology can easily be used to analyse minmax MOM estimators also. As for Hanson-Wright result, the result obtained via a direct approach can be recovered from the general principles.

Theorem 70. *Assume that $P \in \mathcal{P}_2$ then the minmax MOM estimator*

$$\hat{f}_K \in \operatorname{argmin}_{f\in\mathbb{R}^d} \sup_{g\in\mathbb{R}^d} MOM_K\left[\|X-f\|^2 - \|X-g\|^2\right]$$

satisfies,

$$\mathbb{P}\left(\|\hat{f}_K - f_P^*\| > 16\left(32\sqrt{\frac{Tr(\Sigma)}{N}} \vee \sqrt{\frac{2\|\Sigma\|_{op}K}{N}}\right)\right) \leqslant e^{-K/32} \ .$$

Proof. From Eq (4.3) in Theorem 46, with probability $1-e^{-K/32}$, the random variable R defined in Eq (5.16) satisfies $R \leqslant r_K$, where

$$r_K = 128\sqrt{\frac{\mathrm{Tr}(\Sigma_P)}{N}} \vee 4\sqrt{\frac{2\|\Sigma_P\|_{\mathrm{op}}K}{N}} \ .$$

Choosing $\theta(r) = 2r_K r$, all events $\{\Omega_r, r>0\}$, where Ω_r is defined in Lemma 62 are included in the event $\Omega_{\mathrm{good}} = \{R \leqslant r_K\}$. Recall that the choice of $\mathcal{E}$ and d imply $\sup_{f\in F} d(f^*, f) = 0$ and

$$\inf_{f:\mathcal{E}(f)=r} d(f, f_P^*) = r^2 \ .$$

Therefore, the conditions of Corollary 65 are satisfied with $a = 2r_K$, $b = d = 0$, $c = 1$, thus, thanks to this result, the homogeneity lemma holds with $r_1 = 2r_K$, $\zeta = 4r_K^2$, $r_2 = 4r_K$. □

Chapter 6

Learning from Lipschitz-convex losses

This chapter presents results that have been proved in [18]. Following [1], we first investigate the ERM in a general statistical learning setting where the loss function is assumed to be both convex and Lipschtiz in its first variable, see Assumption (6.1). This setting includes several losses that have been considered for convex relaxation of the $0-1$ loss in classification as the hinge loss that is used in the SVM algorithm and the logistic loss that is used in the Boosting algorithm. It also includes classical losses in robust regression as the famous Huber's loss. This analysis is conducted under sub-Gaussian assumption on the design X. We also provide an analysis of minmax MOM estimators which holds under moment conditions only on the design.

6.1 General setting

Consider the supervized learning framework where one observes a dataset $\mathcal{D}_N = (Z_1, \dots, Z_N)$ of random variables taking values in a measurable space $\mathcal{Z}$. The space $\mathcal{Z}$ is a product space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and a data $z \in \mathcal{Z}$ is a couple $z = (x, y)$, where x, called the input, takes values in a measurable space $\mathcal{X}$ and y, called the output, takes values in $\mathcal{Y} \subset \mathbb{R}$. The goal is to predict the value of the output Y from the input X when $Z = (X, Y)$ is drawn from P, independently of $\mathcal{D}_N$. The parameters $f \in F$ are functions $f : \mathcal{X} \to \mathbb{R}$ and the loss function $\ell_f(z)$ takes the form $\ell_f(z) = c(f(x), y)$ for some *cost* function c measuring the accuracy of the prediction of y by $f(x)$.

All along the chapter, the function c is defined on $\bar{\mathcal{Y}} \times \mathcal{Y}$, where $\mathcal{Y} \subset \bar{\mathcal{Y}} \subset \mathbb{R}$ is a convex set containing all possible values of $f(x)$ for $f \in F$ and $x \in \mathcal{X}$, F is a convex set of functions and the following assumption always holds.

$$\exists L > 0 : \ \forall y \in \mathcal{Y}, \qquad c(\cdot, y) \text{ is convex and } L\text{-Lipschitz} \ . \tag{6.1}$$

6.2 Examples of loss functions

Before analysing estimators based on these losses, we provide a few examples of problems in machine learning where Condition (6.1) is met.

Huber regression Let $\alpha > 0$, the Huber function is defined by

$$h_\alpha(x) = \begin{cases} \frac{x^2}{2} & \text{if } x \leqslant \alpha \ , \\ \alpha|x| - \frac{\alpha^2}{2} & \text{if } x > \alpha \ . \end{cases}$$

This function is convex and continuously differentiable, with derivative bounded by α. It interpolates between the quadratic function $x \mapsto x^2/2$ and the absolute value $x \mapsto |x|$. In the 1960's, to build robust alternatives to least-squares minimizers, Huber proposed to estimate the regression function by

$$\hat{f}_{\mathrm{Hub},\alpha} \in \mathrm{argmin}_{f \in F} \sum_{i=1}^{N} h_\alpha(f(x) - y) \ .$$

This estimator typically interpolates between the (unbiased but non robust) least-squares estimator that would be obtained for the function $h(x) = x^2/2$ and the (robust but biased) empirical median that would be obtained for the function $h(x) = |x|$. It transpires from this definition that $\hat{f}_{\mathrm{Hub},\alpha}$ is the ERM associated to the loss function $\ell_f(x, y) = c(f(x), y)$, with $c(u, y) = h_\alpha(u - y)$. In this case, for any subsets $\mathcal{Y} \subset \bar{\mathcal{Y}} = \mathbb{R}$, this cost function satisfies Assumption (6.1) with $L = \alpha$.

Logistic regression Here $\mathcal{Y} = \{-1, 1\}$. The most classical loss in classification is the $0-1$ loss defined by $\mathbf{1}_{\{y \neq f(x)\}}$, which is used in the work of Vapnik for example. The problem with this loss is that the minimization problem defining the ERM

$$\hat{f} \in \mathrm{argmin}_{f \in F} \sum_{i=1}^{N} \mathbf{1}_{\{Y_i \neq f(X_i)\}}$$

is at best computationally demanding, and cannot even be solved in most interesting cases. The problem is that neither F nor the function $f \mapsto P_N \ell_f$ are convex. To bypass this issue, several convex surrogates to the $0 - 1$ loss have been considered. Logistic loss is among the most famous. Define the logistic function

$$\mathcal{L}(u) = \log_2(1 + e^u) \ . \tag{6.2}$$

The logistic function $\mathcal{L}$ is convex, non-increasing and L-Lipschitz with $L = 1/\log(2)$. It is used to define the logistic loss $\ell_f(x, y) = \mathcal{L}(-yf(x))$. This loss has the form $c(f(x), y)$, with

$$c(u, y) = \mathcal{L}(-yu) \ .$$

It is clear that $c(\cdot, y)$ satisfies Assumption 6.1 with $L = 1/\log 2$.

Hinge loss As in the previous example $\mathcal{Y} = \{-1, 1\}$. The hinge loss is another convex surrogate to the $0 - 1$ loss, which is used for example in the SVM algorithms. Define the hinge function

$$H(u) = (1 + u)_+, \qquad \text{where} \qquad \forall x \in \mathbb{R}, \ x_+ = \max(x, 0) \ . \tag{6.3}$$

The hinge function defines the hinge loss $\ell_f(x, y) = H(-yf(x))$. This loss has the form $c(f(x), y)$ with $c(u, y) = H(-uy)$. It satisfies Assumption (6.1) with $L = 1$.

6.3 Examples of classes of functions

This section presents three classes of functions F.

6.3.1 SVM

Recall the definition of reproducing kernel Hilbert spaces.

Definition 71. *Let W denote a Hilbert space of functions $f : \mathcal{X} \to \bar{\mathcal{Y}}$, with $\mathcal{X}$ separable and endowed with a continuous function $K : \mathcal{X}^2 \to \mathbb{R}$ such that*

(i) K is symmetric $K(x, x') = K(x', x)$, for any $x, x' \in \mathcal{X}$,

(ii) for any $x \in \mathcal{X}$, $K(x, \cdot) \in W$,

(iii) for any $f \in W$ and any $x \in \mathcal{X}$, $\langle f, K(x, \cdot)\rangle_W = f(x)$.

The space W is called reproducing kernel Hilbert space (RKHS) with kernel K.

Let W denote a RKHS with kernel K, $\mathcal{D}_N = ((X_1, Y_1), \ldots, (X_N, Y_N))$ and

$$F = \left\{ f \in W : \|f\|_W \leqslant \theta \right\} \ .$$

The class F is used in the SVM algorithm. Let ℓ_f denote the hinge loss: $\ell_f(z) = H(-yf(x))$ (the function H being defined in (6.3)). The support vector machine (SVM) estimator is defined as

$$\hat{f}_{\text{svm}} \in \operatorname{argmin}_{f \in F} P_N \ell_f \ . \tag{6.4}$$

The SVM estimator $\hat{f}_{\text{svm}}$ is an ERM based on a convex and Lipschitz loss. SVM algorithm (6.4) can be equivalently defined as a solution of the minmax problem: if $T_{\text{emp}}(f, g) = P_N[\ell_f - \ell_g]$ denotes the usual empirical test, then

$$\hat{f}_{\text{svm}} \in \operatorname{argmin}_{f \in F} P_N \ell_f = \operatorname{argmin}_{f \in F} \sup_{f \in F} T_{\text{emp}}(f, g) \ .$$

A natural alternative to SVM would therefore be the MOM SVM estimators: if $T_{\text{mom}}(f, g) = \text{MOM}_K[\ell_f - \ell_g]$ denotes the MOM tests,

$$\hat{f}_{\text{msvm}} \in \operatorname{argmin}_{f \in F} \sup_{f \in F} T_{\text{mom}}(f, g) \ . \tag{6.5}$$

Computational issues To actually compute the SVM estimator, the representer theorem shows that SVM equivalently solves $\min_{f \in F_0} P_N \ell_f$, where

$$F_0 = \left\{ \mathbf{a}^T \mathbf{K}, \ : \mathbf{a}^T \mathbb{K} \mathbf{a} \leqslant \theta^2 \right\}, \qquad \mathbf{K}(x) = \begin{bmatrix} K(X_1, x) \\ \vdots \\ K(X_N, x) \end{bmatrix} \ .$$

Here, $\mathbb{K}$ denotes the (random) $N \times N$ matrix with entries $K(X_i, X_j)$. Likewise, for computational issues, the representer theorem can be used to show that

$$\hat{f}_{\text{momSVM}} \in \operatorname{argmin}_{f \in F_0} \sup_{f \in F_0} T_{\text{mom}}(f, g) \ .$$

6.3.2 Boosting

Let $f_1, \dots, f_d$ denote functions $f_i : \mathcal{X} \to \bar{\mathcal{Y}}$ and let Δ_d denote the simplex in $\mathbb{R}^d$:

$$\Delta_d = \left\{ \mathbf{a} \in \mathbb{R}_+^d : \sum_{i=1}^d a_i = 1 \right\} \ .$$

The Boosting estimator is defined as

$$\hat{f}_{\mathrm{Boost}} = \widehat{\mathbf{a}}_b^T \mathbf{f}, \qquad \text{where} \qquad \widehat{\mathbf{a}}_b \in \mathrm{argmin}_{\mathbf{a} \in \Delta_d} P_N \ell_{\mathbf{a}}, \ \mathbf{f}(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_d(x) \end{bmatrix} \ . \quad (6.6)$$

Here, pick $\varphi \in \{\mathcal{L}, H\}$ where the hinge function H and the logistic function $\mathcal{L}$ have been defined respectively in (6.3) and (6.2) and define

$$\ell_{\mathbf{a}}(z) = \varphi(-y \mathbf{a}^T \mathbf{f}(x)) \ .$$

Clearly $\widehat{\mathbf{a}}_b$ is an ERM based on Lipschitz and convex losses $\ell_{\mathbf{a}}$. Alternatively, one can consider MOM Boosting estimators, simply by considering

$$\hat{f}_{\mathrm{mBoost}} = \widehat{\mathbf{a}}_{\mathrm{mb}}^T \mathbf{f}, \qquad \text{where} \qquad \widehat{\mathbf{a}}_{\mathrm{mb}} \in \mathrm{argmin}_{\mathbf{a} \in \Delta_d} \sup_{\mathbf{b} \in \Delta_d} T_{\mathrm{mom}}(\mathbf{a}, \mathbf{b}) \ . \quad (6.7)$$

6.4 Non-localized bounds

Start with a lemma extending Vapnik's bound for ERM. Recall that this elementary upper bound states that

$$P[\ell_{\hat{f}_{\mathrm{erm}}} - \ell_{f^*}] \leqslant 2 \sup_{f \in F} |(P_N - P)\ell_f| \ .$$

This comes from the following fact.

Lemma 72. *Let $\hat{P}$ denote any estimator of the operator P and let*

$$\hat{f} \in \mathrm{argmin}_{f \in F} \hat{P} \ell_f \ .$$

Then,

$$P[\ell_{\hat{f}} - \ell_{f^*}] \leqslant 2 \sup_{f \in F} |(\hat{P} - P)\ell_f| \ .$$

Proof.

$$P[\ell_{\hat{f}} - \ell_{f^*}] = (P - \hat{P})\ell_{\hat{f}} + (\hat{P} - P)\ell_{f^*} + [\hat{P}\ell_{\hat{f}} - \hat{P}\ell_{f^*}] \ .$$

The third term is non-positive by definition of $\hat{f}$ while the two first terms are upper bounded by $\sup_{f \in F} |(\hat{P} - P)\ell_f|$. □

The following lemma extends this bound for minmax estimators.

Lemma 73. *Let $\hat{f}$ denote a minmax estimator:*

$$\hat{f} \in \mathrm{argmin}_{f \in F} \sup_{g \in F} \hat{P}[\ell_f - \ell_g] \ .$$

Then, almost surely,

$$P[\ell_{\hat{f}} - \ell_{f^*}] \leqslant 2 \sup_{f \in F} (\hat{P} - P)[\ell_{f^*} - \ell_f] \ .$$

Proof. Start with basics:

$$\begin{aligned} P[\ell_{\hat{f}} - \ell_{f^*}] &\leqslant \hat{P}[\ell_{\hat{f}} - \ell_{f^*}] + (\hat{P} - P)[\ell_{f^*} - \ell_{\hat{f}}] \\ &\leqslant \hat{P}[\ell_{\hat{f}} - \ell_{f^*}] + \sup_{f \in F} (\hat{P} - P)[\ell_{f^*} - \ell_f] \ . \end{aligned}$$

Then, by definition of $\hat{f}$,

$$\hat{P}[\ell_{\hat{f}} - \ell_{f^*}] \leqslant \sup_{g \in F} \hat{P}[\ell_{\hat{f}} - \ell_g] \leqslant \sup_{g \in F} \hat{P}[\ell_{f^*} - \ell_g] \ .$$

Finally, by definition of f^*, $P[\ell_{f^*} - \ell_g] \leqslant 0$ for any $g \in F$, so

$$\hat{P}[\ell_{\hat{f}} - \ell_{f^*}] \leqslant \sup_{g \in F} (\hat{P} - P)[\ell_{f^*} - \ell_g] \ .$$

This concludes the first inequality of Lemma 73. □

Together with concentration bounds of Chapter 3, Lemma 73 allows to obtain first basic bounds that can be useful in some examples.

Theorem 74. *Assume that $\ell_f(z) = c(f(x), y)$ where c satisfies Assumption 6.1 and that all $f \in F$ have finite $L^2(P)$-moments. Let $\sigma^2(F) = \sup_{f \in F} Var(f(X))$. Then, the min MOM estimator $\hat{f}_{mom} \in \operatorname{argmin}_{f \in F} MOM_K[\ell_f]$ satisfies*

$$\mathbb{P}\left(P[\ell_{\hat{f}_{mom}} - \ell_{f^*}] \leqslant 16\left(32\sqrt{\frac{D_N(F)}{N}} \vee \sqrt{\frac{2\sigma^2(F)K}{N}} \right) \right) \geqslant 1 - e^{-K/32} \ .$$

If all $f(X)$ are Gaussian random variables, then, the ERM $\hat{f}_{erm} \in \operatorname{argmin}_{f \in F} P_N \ell_f$ satisfies

$$\forall s > 0, \qquad \mathbb{P}\left(P[\ell_{\hat{f}} - \ell_{f^*}] \leqslant 8L\sqrt{\frac{D_N(F)}{N}} + 2L\sqrt{\frac{2\sigma^2(F)s}{N}} \right) \geqslant 1 - e^{-s} \ .$$

Proof. By Lemma 73,

$$P[\ell_{\hat{f}_{\mathrm{mom}}} - \ell_{f^*}] \leqslant 4 \sup_{f \in F} |\mathrm{MOM}_K[\ell_f - P\ell_f]| \ .$$

By Theorem 39,

$$\mathbb{P}\left(\sup_{f \in F} |\mathrm{MOM}_K[\ell_f - P\ell_f]| > 128\sqrt{\frac{D_N(F)}{N}} \vee 4\sqrt{\frac{2\sigma^2(F)K}{N}} \right) \leqslant e^{-K/32} \ .$$

By Lemma 72,

$$P[\ell_{\hat{f}_{\mathrm{erm}}} - \ell_{f^*}] \leqslant 2 \sup_{f \in F} |(P_N - P)[\ell_f]| \ .$$

By Assumption 6.1, for any $f \in F$, $\ell_f(z) = c(f(x), y)$ is a L-Lipschitz function. By Theorem 31, it follows that

$$\mathbb{P}\left(\sup_{f \in F} |(P_N - P)[\ell_f]| \leqslant \mathbb{E}[\sup_{f \in F} |(P_N - P)[\ell_f]|] + \sqrt{\frac{2\sigma^2(F)s}{N}} \right) \geqslant 1 - e^{-s} \ .$$

Moreover, by symmetrization,

$$\mathbb{E}[\sup_{f\in F}|(P_N-P)[\ell_f]|] \leqslant 2\mathbb{E}\bigg[\sup_{f\in F}\frac{1}{N}\sum_{i=1}^N \epsilon_i \ell_f(Z_i)\bigg] = 2\sqrt{\frac{D_N(F)}{N}} \ .$$

□

For example, Theorem 74 applies to SVM and Boosting and yields the following corollaries.

Corollary 75. *Assume that the kernel K is a* trace norm *operator, which means that*

$$P[K(X,X)] := k_2 \leqslant \infty \ . \tag{6.8}$$

Let $\Sigma = P[K\otimes K]$, where $K\otimes K : W \to W$ is the random operator defined by

$$\forall f\in W, \qquad K\otimes K(f) = \langle K(X,\cdot), f\rangle_W K(X,\cdot) = f(X)K(X,\cdot) \ .$$

Then, the min MOM SVM estimator satisfies

$$\mathbb{P}\bigg(P[\ell_{\hat f_{msvm}} - \ell_{f^*}] \leqslant 16L\theta\bigg(64\sqrt{\frac{Tr(\Sigma)}{N}} \vee \sqrt{\frac{2\|\Sigma\|_{op}K}{N}}\bigg)\bigg) \geqslant 1-e^{-K/32} \ .$$

If X is a Gaussian vector in $\mathbb{R}^d$, then, the SVM estimator $\hat f_{svm}$ defined in (6.4) *satisfies*

$$\forall s>0, \qquad \mathbb{P}\bigg(P[\ell_{\hat f} - \ell_{f^*}] \leqslant 8L\theta\sqrt{\frac{Tr(\Sigma)}{N}} + 2L\theta\sqrt{\frac{2\|\Sigma\|_{op}s}{N}}\bigg) \geqslant 1-e^{-s} \ .$$

Remark 76. *Assumption 6.8 relaxes the boundedness assumption $\sup_{x\in\mathcal{X}} K(x,x) := k_\infty < +\infty$ usually considered to analyse SVM. The expectation defining Σ is understood in Bochner sense, see for example [52].*

Proof. The result is a combination of Theorem 74 with the following lemma. □

Lemma 77. *Assume that K is a trace norm operator and let $\Sigma = P[K\otimes K]$. Then,*

$$D_N(F) \leqslant \theta^2 k_2 = \theta^2\, Tr(\Sigma) \ ,$$
$$\sigma^2(F) = \sup_{f\in F} Var(\ell_f(Z)) \leqslant 2L^2 \sup_{f\in F} P[f^2(X)] = 2L^2\theta^2\|\Sigma\|_{op} \ .$$

Proof. Start with the variance. Let Z' denote an independent copy of Z. By Jensen's inequality,

$$\begin{aligned}\mathrm{Var}(\ell_f(Z)) &= \mathbb{E}[(\ell_f(Z) - \mathbb{E}[\ell_f(Z')|Z])^2] \leqslant \mathbb{E}[(\ell_f(Z)-\ell_f(Z'))^2] \\ &\leqslant L^2\mathbb{E}[(f(X)-f(X'))^2] \leqslant 2L^2\mathrm{Var}(f(X)) \leqslant 2L^2P[f^2] \ .\end{aligned}$$

The operator $K\otimes K$ is a.s. symmetric: for any f,g in W,

$$\langle K\otimes K(f), g\rangle_W = \langle K(X,\cdot), f\rangle_W \langle K(X,\cdot), g\rangle_W = \langle f, K\otimes K(g)\rangle_W \ .$$

Therefore, Σ is symmetric and, as W is separable under the assumptions that $\mathcal{X}$ is separable and K continuous, see for example [52, Lemma 4.33], this implies that there exists an orthonormal basis of W made of eigenvectors of Σ. Moreover, for any $f \in W$,

$$P[f^2(X)] = P[\langle f, K \otimes K(f)\rangle_W] = \langle f, \Sigma(f)\rangle_W \ .$$

Therefore,

$$\sup_{f\in F} P[f^2] = \theta^2 \|\Sigma\|_{\mathrm{op}} \ . \tag{6.9}$$

Let us now turn to the Rademacher complexity of F. Using successively the representation property (iii) and Cauchy-Schwarz inequality twice,

$$\begin{aligned}
D_N(F) &= \left(\mathbb{E}\left[\sup_{f\in W:\|f\|_W\leqslant\theta} \frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i f(X_i)\right]\right)^2 \\
&= \left(\mathbb{E}\left[\sup_{f\in W:\|f\|_W\leqslant\theta} \left\langle f, \frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i K(X_i,\cdot)\right\rangle_W\right]\right)^2 \\
&\leqslant \theta^2\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i K(X_i,\cdot)\right\|_W\right]\right)^2 \\
&= \theta^2\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i K(X_i,\cdot)\right\|_W^2\right] \ .
\end{aligned}$$

Moreover, developing the square-norm, using the representation property (iii) shows that

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i K(X_i,\cdot)\right\|_W^2\right] &= \frac{1}{N}\sum_{1\leqslant i,j\leqslant N} \mathbb{E}\big[\epsilon_i\epsilon_j \langle K(X_i,\cdot), K(X_j,\cdot)\rangle_W\big] \\
&= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\langle K(X_i,\cdot), K(X_i,\cdot)\rangle_W\big] = k_2 \ .
\end{aligned}$$

Hence,

$$D_N(F) \leqslant k_2\theta^2 \ .$$

Finally, the random operator $K\otimes K$ has clearly rank 1 with $K(X,X)$ as single singular value. By Fubbini-Tonelli theorem, it yields

$$k_2 = P[K(X,X)] = P[\mathrm{Tr}(K\otimes K)] = \mathrm{Tr}(P[K\otimes K]) = \mathrm{Tr}(\Sigma) \ .$$

The trace-norm assumption therefore states that the trace of $K\otimes K$ is finite. □

Corollary 78. *Consider the boosting class based on a collection of functions satisfying the following assumptions. Let* $\sigma^2 = \max_{1\leqslant i\leqslant d} P[f_i^2]$. *For* $p = \log d$, *there exists a constant* $\gamma > 0$ *such that*

$$\forall j \in \{1,\dots,d\}, \qquad P[f_j^p] \leqslant (\gamma\sigma)^p \ . \tag{6.10}$$

The min MOM estimator satisfies

$$\mathbb{P}\left(P[\ell_{\hat{f}_{mBoost}} - \ell_{f^*}] \leqslant \frac{16L\sigma}{\sqrt{N}}\left(192e\gamma\sqrt{\log d}\vee\sqrt{2K}\right)\right) \geqslant 1 - e^{-K/32} \ .$$

If X is a Gaussian vector in $\mathbb{R}^d$, then, the Boosting estimator $\hat{f}_{Boost}$ defined in (6.6) *satisfies*

$$\forall s > 0, \qquad \mathbb{P}\left(P[\ell_{\hat{f}_{Boost}} - \ell_{f^*}] \leqslant 2L\sqrt{\frac{\|\Sigma\|_\infty}{N}}\left(12e\gamma\sqrt{\log d} + \sqrt{2s}\right)\right) \geqslant 1 - e^{-s} \ .$$

Proof. The result is a combination of Theorem 74 with the following result. □

Lemma 79. *Assume that $P[\|\boldsymbol{f}(X)\|^2] < \infty$ and let*

$$\Sigma = P[\boldsymbol{f}(X)\boldsymbol{f}(X)^T], \qquad \|\Sigma\|_\infty = \max_{1\leqslant i,j\leqslant d} |\Sigma_{i,j}| \ .$$

Then,

$$\sup_{\boldsymbol{a}\in\Delta_d} Var(\boldsymbol{a}^T\boldsymbol{f}(X)) \leqslant P[(\boldsymbol{a}^T\boldsymbol{f}(X))^2] \leqslant \|\Sigma\|_\infty \ .$$

Moreover, for any $p \geqslant 2$ such that $\max_{1\leqslant j\leqslant d} P[|f_j|^p] < \infty$, if $\Theta_p = \sum_{i=1}^d P[|f_j|^p]$, then

$$D_N(F) \leqslant 9p\Theta_p^{2/p} \ . \tag{6.11}$$

In particular, if (6.10) *holds, then*

$$D_N(F) \leqslant 9e^2\gamma^2\|\Sigma\|_\infty \log d \ . \tag{6.12}$$

Proof. Start with the variance. Let $\mathbf{a} \in \Delta_d$,

$$P[(\mathbf{a}^T\mathbf{f}(X))^2] \leqslant \sup_{\mathbf{a}\in\Delta_d} \mathbf{a}^T\Sigma\mathbf{a} \ .$$

It is not hard not see that, for any $\mathbf{a} \in \Delta_d$,

$$\mathbf{a}^T\Sigma\mathbf{a} \leqslant \max_{i=1,\dots,d}(\Sigma\mathbf{a})_i \leqslant \max_{1\leqslant i,j\leqslant d}|\Sigma_{i,j}| = \|\Sigma\|_\infty \ .$$

Hence,

$$\sup_{\mathbf{a}\in\Delta_d} P[(\mathbf{a}^T\mathbf{f}(X))^2] \leqslant \|\Sigma\|_\infty \ .$$

Regarding the Rademacher complexity.

$$\begin{aligned} D_N(F) &= \left(\mathbb{E}\left[\sup_{\mathbf{a}\in\Delta_d} \frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i \mathbf{a}^T\mathbf{f}(X_i)\right]\right)^2 \\ &= \left(\mathbb{E}\left[\sup_{\mathbf{a}\in\Delta_d} \mathbf{a}^T\left(\sum_{i=1}^N \epsilon_i \frac{\mathbf{f}(X_i)}{\sqrt{N}}\right)\right]\right)^2 \\ &= \left(\mathbb{E}\left[\max_{1\leqslant j\leqslant d}\left|\left(\sum_{i=1}^N \epsilon_i \frac{\mathbf{f}(X_i)}{\sqrt{N}}\right)_j\right|\right]\right)^2 \ . \end{aligned} \tag{6.13}$$

Under the assumption $\max_{1\leqslant j\leqslant d} P[|f_j|^p] < \infty$, the random variables

$$Z_j = \left(\sum_{i=1}^N \epsilon_i \frac{\mathbf{f}(X_i)}{\sqrt{N}}\right)_j$$

have finite moments of order p. Moreover, by Jensen's inequality,

$$\mathbb{E}\big[\max_{1\leqslant j\leqslant d}\big|Z_j\big|\big] \leqslant \Big(\mathbb{E}\big[\max_{1\leqslant j\leqslant d}\big|Z_j\big|^p\big]\Big)^{1/p} \leqslant \Big(\sum_{j=1}^{d}\mathbb{E}\big[\big|Z_j\big|^p\big]\Big)^{1/p} \ . \tag{6.14}$$

Now, apply Khinchine's inequality on moments of order p for sums of independent random variables, see for example [10, Chapter 15]. It shows that

$$\begin{aligned}\mathbb{E}\big[\big|Z_j\big|^p\big]^{1/p} &\leqslant 3\sqrt{p\sum_{i=1}^{N}\mathbb{E}\Big[\frac{|f_j(X_i)|^p}{N^{p/2}}\Big]^{2/p}} = 3\sqrt{\frac{p}{N}\sum_{i=1}^{N}\mathbb{E}\big[|f_j(X_i)|^p\big]^{2/p}} \\ &= 3\sqrt{pP\big[|f_j|^p\big]^{2/p}} = 3\sqrt{p}P\big[|f_j|^p\big]^{1/p} \ .\end{aligned}$$

This shows (6.11). By Assumption 6.10, it follows that

$$\mathbb{E}\big[\big|Z_j\big|^p\big]^{1/p} \leqslant 3\gamma\sqrt{pP[f_j^2]} \leqslant 3\gamma\sqrt{p\|\Sigma\|_\infty} \ .$$

Plugging this inequality into (6.14) yields

$$\mathbb{E}\big[\max_{1\leqslant j\leqslant d}\big|Z_j\big|\big] \leqslant 3\gamma\sqrt{p\|\Sigma\|_\infty}d^{1/p} \ .$$

As $p = \log d$, this yields

$$\mathbb{E}\big[\max_{1\leqslant j\leqslant d}\big|Z_j\big|\big] \leqslant 3e\gamma\sqrt{\|\Sigma\|_\infty \log d} \ .$$

Plugging this bound into (6.13) shows (6.12). □

6.5 Localized bounds: preliminary results

Theorem 74 is inoperant when $D(F) = \infty$, which happens for example with classes of linear functions indexed by unbounded subsets of $\mathbb{R}^d$, for example:

$$F = \{\mathbf{f}^T\cdot,\ \mathbf{f}\in\mathbb{R}^d\} \ .$$

The following sections develop a general strategy that allows to deal with these examples. Hereafter, assume that $\mathcal{X} = \mathbb{R}^d$ and F is the set of all linear functions $\mathbf{f}^T\cdot$ with $\mathbf{f}\in\mathbb{R}^d$. Assume also that the distribution P of $Z = (X, Y)$ has a first marginal X satisfying $P[\|X\|^2] < \infty$ and $\mathcal{Y}\subset\mathbb{R}$. Denote by $\Sigma = P[XX^T]$. Both the ERM and minmax MOM estimators will be analysed thanks to the homogeneity lemma, Lemma 62. The convexity of $c(\cdot, y)$ implies the convexity of ℓ_f therefore, Lemma 66 applies and shows that the tests

$$T_{\mathrm{erm}}(f,g) = P_N[\ell_f - \ell_g], \qquad T_{\mathrm{mom}}(f,g) = \mathrm{MOM}_K[\ell_f - \ell_g]$$

satisfy the homogeneity assumption **(HP)** of the homogeneity lemma, provided that the evaluation function $\mathcal{E}$ derives from a (semi-)norm. Hereafter, for any $f\in F$, $\mathcal{E}$ is assumed to derive from the following $\|\cdot\|_\Sigma$ semi-norm, which is defined by

$$\mathcal{E}(f) = \sqrt{P[(f-f^*)^2]} = \sqrt{\mathbb{E}[(\mathbf{X}^T(\mathbf{f}-\mathbf{f}^*))^2]} = \sqrt{(\mathbf{f}-\mathbf{f}^*)^T\Sigma(\mathbf{f}-\mathbf{f}^*)} = \|\mathbf{f}-\mathbf{f}^*\|_\Sigma \ .$$

Finally, as in every learning problem

$$d(f,g) = P[\ell_f - \ell_g] \ ,$$

so $d(f,g) = -d(g,f)$ and $d(f^*,f) \leqslant 0$ so $\zeta = \Theta(r_1)$ in the homogeneity lemma (Lemma 62). The homogeneity lemma will be used under a technical assumption that we introduce and discuss in the following section.

6.6 Bernstein's condition

To check (5.8) and (5.9), the following "local" Bernstein condition will be useful: there exist $A > 0$ and $B > 0$ such that

$$\forall f \in F : \mathcal{E}(f) \leqslant A, \qquad P[\ell_f - \ell_{f^*}] \geqslant B\mathcal{E}(f)^2 \ . \tag{6.15}$$

Relationships between between $\mathcal{E}(f)$ and the excess risk $P[\ell_f - \ell_{f^*}]$ are usually called Bernstein's condition. These are convenient to prove "fast rates" of convergence for ERM with bounded losses, see for example [53] for a discussion on fast and slow rates. To the best of our knowledge, this assumption first appeared in [43, Hyp A2 of Theorem 4.2]. This form of Assumption 6.15 was first introduced in [18]. The relationship between $\mathcal{E}(f)$ and $P[\ell_f - \ell_{f^*}]$ is only assumed in a neighborhood of f^*. This is a necessary constraint to deal with unbounded classes of functions. Actually, by the Lipschitz assumption of c, it holds, by Cauchy-Schwarz inequality,

$$P[\ell_f - \ell_{f^*}] \leqslant LP|f - f^*| \leqslant L\mathcal{E}(f) \ .$$

Hence, the Bernstein's assumption (6.15) can only be true if

$$B\mathcal{E}(f)^2 \leqslant L\mathcal{E}(f), \qquad \text{that is, if} \qquad \mathcal{E}(f) \leqslant \frac{L}{B} \ .$$

Let us present some examples where Assumption (6.15) holds. To proceed, we assume in the remaining of this section that

$$f^* \text{ is a minimizer of } P\ell_f \text{ among all measurable functions } f : \mathcal{X} \to \mathcal{Y} \ . \tag{6.16}$$

This assumption is quite restrictive as it implies that the model F is "exact". It is convenient to make explicite computations. Indeed, it ensures that

$$\forall x \in \mathcal{X}, \qquad f^*(x) \in \operatorname{argmin}_{u\in\mathbb{R}} \mathbb{E}[c(u,Y)|X = x] \ .$$

In particular, it allows to show results on f^* based on assumption on the c.d.f. of Y conditionally on $X = x$.

The second assumption that will be done all along the examples is an hypothesis comparing $L^4(P)$ and $L^2(P)$ norms of functions in F. For any $p \geqslant 1$, for any function $f : \mathcal{X} \to \mathbb{R}$ for which it makes sense, let

$$\|f\|_{L^p(P)} = \left(P[|f|^p]\right)^{1/p} \ .$$

The L^4/L^2 assumption states that there exists $\Delta \geqslant 1$ such that

$$\forall f \in F : \qquad \|f - f^*\|_{L^4(P)} \leqslant \Delta \|f - f^*\|_{L^2(P)} \ . \tag{6.17}$$

Let us comment this assumption. First, by Cauchy-Schwarz inequality

$$\|f - f^*\|_{L^2(P)} \leqslant \|f - f^*\|_{L^4(P)} \ ,$$

hence, the restriction $\Delta \geqslant 1$ in Assumption (6.17) holds without loss of generality. The following proposition gives an example where Assumption (6.17) holds.

Proposition 80. *Assume that $X \in \mathbb{R}^d$ is a vector with centered, independent entries X_i, $i \in \{1, \dots, d\}$ with kurtosis bounded by κ, i.e. such that $P[X_i^4]^{1/4} \leqslant \kappa P[X_i^2]^{1/2}$. Then, any linear function $f(\cdot) = \boldsymbol{f}^T \cdot$ satisfies $\|f\|_{L^4(P)} \leqslant \kappa \|f\|_{L^2(P)}$.*

Proof. One can assume w.l.o.g. that $\kappa \geqslant 1$. Using independence of X_i and the fact that $P[X_i] = 0$,

$$\|f\|_{L^2(P)} = \Big(\sum_{i=1}^d \mathbf{f}_i^2 P[X_i^2] \Big)^{1/2} \ ,$$

$$\|f\|_{L^4(P)} = \Big(\sum_{i=1}^d \mathbf{f}_i^4 P[X_i^4] + \sum_{1 \leqslant i \neq j \leqslant d} \mathbf{f}_i^2 \mathbf{f}_j^2 P[X_i^2] P[X_j]^2 \Big)^{1/4} \ .$$

Using that $P[X_i^4] \leqslant \kappa^4 P[X_i^2]^2$ and $\kappa \geqslant 1$, it yields

$$\begin{aligned} \|f\|_{L^4(P)} &\leqslant \kappa \Big(\sum_{i=1}^d \mathbf{f}_i^4 P[X_i^2]^2 + \sum_{1 \leqslant i \neq j \leqslant d} \mathbf{f}_i^2 \mathbf{f}_j^2 P[X_i^2] P[X_j]^2 \Big)^{1/4} \\ &= \kappa \Big(\sum_{1 \leqslant i,j \leqslant d} \mathbf{f}_i^2 \mathbf{f}_j^2 P[X_i^2] P[X_j]^2 \Big)^{1/4} \\ &= \kappa \Big(\sum_{i=1}^d \mathbf{f}_i^2 P[X_i^2] \Big)^{1/2} = \kappa \|f\|_{L^2(P)} \ . \end{aligned}$$

□

The L^4/L^2 should be used with care as shown by the following example.

Proposition 81. *Let X denote a random variables taking values in a measurable space $\mathcal{X}$. Let $I_1, \dots, I_d$ denote a partition of $\mathcal{X}$ such that $P[I_j] = 1/d$ for any $j \in \{1, \dots, d\}$. Let $\boldsymbol{X} \in \mathbb{R}^d$ denote the vector*

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{\{X \in I_1\}} \\ \vdots \\ \mathbf{1}_{\{X \in I_d\}} \end{bmatrix} \in \mathbb{R}^d \ .$$

Then, for any $\boldsymbol{f} \in \mathbb{R}^d$, $P[(\boldsymbol{f}^T \boldsymbol{X})^4]^{1/4} \leqslant d^{1/4} P[(\boldsymbol{f}^T \boldsymbol{X})^2]^{1/2}$.

Remark 82. *In words, any class of linear functions $f(\cdot) = \boldsymbol{f}^T \cdot$ satisfies Assumption* (6.17)*, but with a parameter Δ that is* not *a constant, but depends on the dimension d.*

Proof. For any $\mathbf{f} \in \mathbb{R}^d$,

$$P[(\mathbf{f}^T \mathbf{X})^4] = \sum_{j=1}^d \mathbf{f}_i^4 P[I_j] \leqslant d \sum_{j=1}^d \mathbf{f}_i^4 P[I_j]^2 \leqslant d \big(\sum_{j=1}^d \mathbf{f}_j^2 P[I_j] \big)^2 = d P[(\mathbf{f}^T \mathbf{X})^2]^2 \ .$$

□

Huber loss Denote by F_x the conditional c.d.f. of Y given $X = x$. Assume that there exists $\nu > 0$ such that

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y} : \ |y - f^*(x)| \leqslant 2A\Delta^2, \qquad F_x(y+\alpha) - F_x(y-\alpha) \geqslant \nu \ . \tag{6.18}$$

For example, Assumption (6.18) holds if the conditional density f_x of Y given $X = x$ is bounded away from 0 in a neighborhood of $f^*(x)$.

Proposition 83. *Assume* (6.16), (6.17) *and* (6.18). *Then,*

$$\forall f \in F : \mathcal{E}(f) \leqslant A, \qquad P[\ell_f - \ell_{f^*}] \geqslant \frac{\nu}{4}\mathcal{E}(f)^2 \ .$$

Proof. Let

$$H_x(u) = \mathbb{E}[h_\alpha(Y-u)|X=x] = \int h_\alpha(y-u)\mathrm{d}F_x(y) \ .$$

The function H_x is differentiable, with

$$\begin{aligned}
H'_x(u) &= -\int h'_\alpha(y-u)F_x(y) \\
&= \alpha \int_{-\infty}^{u-\alpha} \mathrm{d}F_x(y) - \int_{u-\alpha}^{u+\alpha} (y-u)\mathrm{d}F_x(y) - \alpha \int_{u+\alpha}^{+\infty} \mathrm{d}F_x(y) \\
&= \alpha(F_x(u-\alpha) - 1 + F_x(u+\alpha)) - \int_{u-\alpha}^{u+\alpha} (y-u)\mathrm{d}F_x(y) \\
&= \alpha(F_x(u-\alpha) - 1 + F_x(u+\alpha)) - [(y-u)F_x(y)]_{u-\alpha}^{u+\alpha} + \int_{u-\alpha}^{u+\alpha} F_x(y)\mathrm{d}y \\
&= \int_{u-\alpha}^{u+\alpha} F_x(y)\mathrm{d}y - \alpha \ .
\end{aligned}$$

In particular, as $f^*(x) \in \operatorname{argmin}_{u\in\mathbb{R}} H_x(u)$, it follows that $H'_x(f^*(x)) = 0$. Moreover,

$$H''_x(u) = F_x(u+\alpha) - F_x(u-\alpha) \ .$$

Let $\mathcal{X}_{\mathrm{loc}} = \{x \in \mathcal{X} : |f(x) - f^*(x)| \leqslant 2A\Delta^2\}$. For any $x \in \mathcal{X}_{\mathrm{loc}}$, it follows that

$$\begin{aligned}
H_x(f(x)) - H_x(f^*(x)) &= \int_{f^*(x)}^{f(x)} H'_x(u)\mathrm{d}u = \int_{f^*(x)}^{f(x)} (H'_x(u) - H'_x(f^*(x)))\mathrm{d}u \\
&= \int_{f^*(x)}^{f(x)} \int_{f^*(x)}^{u} H''_x(v)\mathrm{d}v\mathrm{d}u \ .
\end{aligned}$$

For any v in the segment with extremities $f^*(x)$ and u, by Assumption (6.18),

$$H''_x(v) = F_x(v+\alpha) - F_x(v-\alpha) \geqslant \nu \ .$$

Therefore, if $f(x) \geqslant f^*(x)$,

$$\begin{aligned}
H_x(f(x)) - H_x(f^*(x)) &\geqslant \int_{f^*(x)}^{f(x)} \int_{f^*(x)}^{u} \nu \mathrm{d}v\mathrm{d}u \\
&= \int_{f^*(x)}^{f(x)} \nu(u - f^*(x))\mathrm{d}u \\
&= \frac{\nu}{2}(f(x) - f^*(x))^2 \ .
\end{aligned}$$

Likewise, if $f(x) \leqslant f^*(x)$,

$$\begin{aligned} H_x(f(x)) - H_x(f^*(x)) &\geqslant \int_{f(x)}^{f^*(x)} \int_u^{f^*(x)} \nu \mathrm{d}v \mathrm{d}u \\ &= \int_{f(x)}^{f^*(x)} \nu(f^*(x) - u) \mathrm{d}u \\ &= \frac{\nu}{2}(f(x) - f^*(x))^2 \ . \end{aligned}$$

Overall, by definition of $f^*(x)$, $H_x(f(x)) - H_x(f^*(x)) \geqslant 0$ for any $x \in \mathcal{X}$ and

$$\forall x \in \mathcal{X}_{\mathrm{loc}}, \qquad H_x(f(x)) - H_x(f^*(x)) \geqslant \frac{\nu}{2}(f(x) - f^*(x))^2 \ .$$

It follows that

$$\begin{aligned} P[\ell_f - \ell_{f^*}] &= \mathbb{E}[H_X(f(X)) - H_X(f^*(X))] \\ &\geqslant \mathbb{E}[\{H_X(f(X)) - H_X(f^*(X))\}\mathbf{1}_{\{X \in \mathcal{X}_{\mathrm{loc}}\}}] \\ &\geqslant \frac{\nu}{2}\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \in \mathcal{X}_{\mathrm{loc}}\}}] \\ &= \frac{\nu}{2}(\mathcal{E}(f) - \mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\mathrm{loc}}\}}]) \ . \end{aligned} \tag{6.19}$$

By Cauchy-Schwarz,

$$\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\mathrm{loc}}\}}] \leqslant \|f - f^*\|_{L^4(P)}^2 \sqrt{\mathbb{P}(X \notin \mathcal{X}_{\mathrm{loc}})} \ . \tag{6.20}$$

By Markov's inequality,

$$\mathbb{P}(X \notin \mathcal{X}_{\mathrm{loc}}) = \mathbb{P}(|f(x) - f^*(x)| > 2A\Delta^2) \leqslant \frac{\|f - f^*\|_{L^2(P)}^2}{4A^2\Delta^4} = \frac{\mathcal{E}(f)^2}{4A^2\Delta^4} \ .$$

If $\mathcal{E}(f) \leqslant A$, it follows that

$$\mathbb{P}(X \notin \mathcal{X}_{\mathrm{loc}}) \leqslant \frac{1}{4\Delta^4} \ .$$

Plugging this into (6.20) shows that, for any $f \in F$ such that $\mathcal{E}(f) \leqslant A$.

$$\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\mathrm{loc}}\}}] \leqslant \frac{\|f - f^*\|_{L^4(P)}^2}{2\Delta^2} \ .$$

Using (6.17), we get

$$\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\mathrm{loc}}\}}] \leqslant \frac{\|f - f^*\|_{L^2(P)}^2}{2} = \frac{\mathcal{E}(f)^2}{2} \ .$$

Plugging this inequality into (6.19) concludes the proof. □

Logistic regression Denote by $\eta : \mathcal{X} \to \mathcal{Y}$ the regression function satisfying $\mathbb{E}[Y\vartheta(X)] = P[\eta\vartheta]$ for any bounded measurable function ϑ. Recall that

$$\log\left[\frac{\eta(x)}{1 - \eta(x)}\right] \in \mathrm{argmin}_{u \in \mathbb{R}} \, \mathbb{E}[\mathcal{L}(-Yu)|X = x] \ .$$

Assume that there exists $\nu > 0$, such that

$$\mathbb{P}\left(\frac{1}{1+e^{\nu}} \leqslant \eta(X) \leqslant \frac{1}{1+e^{-\nu}} \right) \geqslant 1 - \frac{1}{8\Delta^4} \ . \tag{6.21}$$

This is equivalent to

$$\mathbb{P}\left(\log\left[\frac{\eta(X)}{1-\eta(X)} \right] > \nu \right) \leqslant \frac{1}{8\Delta^4} \ .$$

Proposition 84. *Assume* (6.16), (6.17), (6.21). *Then, there exists a constant* $B = \boldsymbol{B}(A, \nu, \Delta) > 0$ *such that, for all* $f \in F$ *such that* $\mathcal{E}(f) \leqslant A$, $P[\ell_f - \ell_{f^*}] \geqslant B\mathcal{E}(f)^2$.

Remark 85. *A value of the constant* B *is given in Eq* (6.22) *in the proof.*

Proof. Let $f \in F$ such that $\mathcal{E}(f) \leqslant A$. Let $H_x(u) = \eta(x)\log_2(1+e^{-u}) + (1-\eta(x))\log_2(1+e^{u})$. The function H_x is continuously twice differentiable with

$$\begin{aligned} H_x'(u) &= \frac{\eta(x)}{\log(2)} \frac{-e^{-u}}{1+e^{-u}} + \frac{1-\eta(x)}{\log 2} \frac{e^{u}}{1+e^{u}} \\ &= \frac{-\eta(x) + (1-\eta(x))e^{u}}{(\log 2)(1+e^{u})} \ . \end{aligned}$$

$$\begin{aligned} H_x''(u) &= \frac{(1-\eta(x))e^{u}(1+e^{u}) - (-\eta(x) + (1-\eta(x))e^{u})e^{u}}{(\log 2)(1+e^{u})^2} \\ &= \frac{e^{u}}{(\log 2)(1+e^{u})^2} \ . \end{aligned}$$

Let $\mathcal{X}_{\text{loc}} = \{x \in \mathcal{X} : |f^*(x)| \leqslant \nu, |f(x) - f^*(x)| \leqslant \sqrt{8}A\Delta^2\}$. For any $x \in \mathcal{X}_{\text{loc}}$, $\max\{|f(x)|, |f^*(x)|\} \leqslant \nu + \sqrt{8}A\Delta^2$. Therefore, as $H_x'(f^*(x)) = 0$, for any $x \in \mathcal{X}$, $H_x(f(x)) - H_x(f^*(x)) \geqslant 0$ and

$$\forall x \in \mathcal{X}_{\text{loc}}, \qquad H_x(f(x)) - H_x(f^*(x)) \geqslant 2B(f(x) - f^*(x))^2 \ ,$$

where

$$B = \frac{e^{-(\nu + \sqrt{8}A\Delta^2)}}{2(\log 2)(1+e^{\nu+\sqrt{8}A\Delta^2})^2} \ . \tag{6.22}$$

It follows that

$$\begin{aligned} P[\ell_f - \ell_{f^*}] &= \mathbb{E}[H_X(f(X)) - H_X(f^*(X))] \\ &\geqslant \mathbb{E}[\{H_X(f(X)) - H_X(f^*(X))\}\mathbf{1}_{\{X \in \mathcal{X}_{\text{loc}}\}}] \\ &\geqslant \frac{\nu}{2}\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \in \mathcal{X}_{\text{loc}}\}}] \\ &= 2B(\mathcal{E}(f) - \mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\text{loc}}\}}]) \ . \end{aligned} \tag{6.23}$$

By Cauchy-Schwarz,

$$\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\text{loc}}\}}] \leqslant \|f - f^*\|_{L^4(P)}^2 \sqrt{\mathbb{P}(X \notin \mathcal{X}_{\text{loc}})} \ . \tag{6.24}$$

By Markov's inequality,

$$\begin{aligned}\mathbb{P}(X \notin \mathcal{X}_{\text{loc}}) &\leqslant \mathbb{P}(|f^*(X)| > \nu) + \mathbb{P}(|f(X) - f^*(X)| > \sqrt{8}A\Delta^2) \\ &\leqslant \frac{1}{8\Delta^4} + \frac{\|f - f^*\|^2_{L^2(P)}}{8A^2\Delta^4} \\ &= \frac{1}{8\Delta^4} + \frac{\mathcal{E}(f)^2}{8A^2\Delta^4} \ .\end{aligned}$$

If $\mathcal{E}(f) \leqslant A$, it follows that

$$\mathbb{P}(X \notin \mathcal{X}_{\text{loc}}) \leqslant \frac{1}{4\Delta^4} \ .$$

Plugging this into (6.24) shows that, for any $f \in F$ such that $\mathcal{E}(f) \leqslant A$.

$$\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\text{loc}}\}}] \leqslant \frac{\|f - f^*\|^2_{L^4(P)}}{2\Delta^2} \ .$$

Using (6.17), we get

$$\mathbb{E}[(f(X) - f^*(X))^2 \mathbf{1}_{\{X \notin \mathcal{X}_{\text{loc}}\}}] \leqslant \frac{\|f - f^*\|^2_{L^2(P)}}{2} = \frac{\mathcal{E}(f)^2}{2} \ .$$

Plugging this inequality into (6.23) concludes the proof. □

Exercise Find conditions sufficient to prove the "local" Bernstein's condition for the Hinge loss.

6.7 ERM in the Gaussian case

We consider a cost function c satisfying Assumption (6.1) and study the estimator

$$\hat{f} \in \text{argmin}_{f \in \mathbb{R}^d} \sum_{i=1}^{N} c(f^T X_i, Y_i) \ . \tag{6.25}$$

Theorem 86. *Assume that X is Gaussian with $\Sigma = P[XX^T]$ positive definite. Assume that the Bernstein assumption* (6.15) *holds for constants A and B such that*

$$AB\sqrt{N} \geqslant 16L\sqrt{d} \ .$$

Then, for any s such that

$$16L(\sqrt{d} + \sqrt{s}) \leqslant AB\sqrt{N} \ , \tag{6.26}$$

the empirical risk minimizer (6.25) *satisfies*

$$\mathbb{P}\left(\mathcal{E}(\hat{f}) \leqslant \frac{16L}{B}\frac{\sqrt{d} + \sqrt{s}}{\sqrt{N}}\right) \geqslant 1 - 2e^{-s} \ .$$

Remark 87. *This result is "robust" as it does not involve assumptions on the outputs Y_i.*

Proof. Recall that the empirical risk minimizer

$$\hat{f} \in \operatorname{argmin}_{f\in F} P_N \ell_f = \operatorname{argmin}_{f\in F} \sup_{g\in F} T_{\mathrm{emp}}(f,g) \ ,$$

with

$$T_{\mathrm{emp}}(f,g) = P_N[\ell_f - \ell_g] \ .$$

As explained in Section 6.5, the test $T_{\mathrm{erm}}(f,g)$ satisfy Assumption **(HP)** of the homogeneity lemma (Lemma 62). Moreover, recall that we want to apply this lemma with $d(f,g) = P[\ell_f - \ell_g]$. It remains to compute the function θ in the homogeneity lemma, and for this, we look for a bound $\theta(r)$ such that, for $\mathbf{B}(r) = \{f \in F : \mathcal{E}(f) \leqslant r\}$, with high probability,

$$\sup_{f\in\mathbf{B}(r)} (P_N - P)[\ell_{f^*} - \ell_f] \leqslant \theta(r) \ .$$

Assume to simplify the argument that $f^* = 0$. This case can be solved with the basic Gaussian concentration inequality. The general case involves more elaborated tools on Gaussian processes, see [1, Lemma 8.1.].

By Theorem 31, with probability larger that $1 - e^{-s}$,

$$\sup_{f\in\mathbf{B}(r)} (P_N - P)[\ell_{f^*} - \ell_f] \leqslant E_N(\mathbf{B}(r)) + \sqrt{\frac{2\sigma^2(\mathbf{B}(r))s}{N}} \ .$$

Here, $\sigma^2(\mathbf{B}(r)) = \sup_{f\in\mathbf{B}(r)} \mathrm{Var}((\ell_f - \ell_{f^*})(Z))$ and

$$E_N(\mathbf{B}(r)) = \mathbb{E}\left[\sup_{f\in\mathbf{B}(r)} (P_N - P)[\ell_{f^*} - \ell_f] \right] \ .$$

Let us first bound the variance.

$$\mathrm{Var}((\ell_f - \ell_{f^*})(Z)) \leqslant P[(\ell_f - \ell_{f^*})^2] \leqslant L^2 P[(f - f^*)^2] = L^2 \mathcal{E}(f)^2 \ .$$

Hence, $\sigma^2(\mathbf{B}(r)) \leqslant L^2 r^2$. Using the symmetrization trick,

$$E_N(\mathbf{B}(r)) \leqslant 2\sqrt{\frac{D_N(\mathbf{B}(r))}{N}} \ ,$$

where

$$D_N(\mathbf{B}(r)) = \left(\mathbb{E}\left[\sup_{f\in\mathbf{B}(r)} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i (\ell_f - \ell_{f^*})(Z_i) \right] \right)^2 \ .$$

By the contraction lemma,

$$D_N(\mathbf{B}(r)) \leqslant 4L^2 \left(\mathbb{E}\left[\sup_{f\in\mathbf{B}(r)} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i (f - f^*)(X_i) \right] \right)^2 \ .$$

Now, $\mathbf{B}(r) = \{f = f^* + rg, \ g \in \mathbf{B}\}$, with $\mathbf{B} = \{f = \mathbf{f}^T \cdot : \ P[(\mathbf{f}^T X)^2] = 1\}$. Hence,

$$D_N(\mathbf{B}(r)) \leqslant 4L^2 r^2 \left(\mathbb{E}\left[\sup_{f=\mathbf{f}^T\cdot\in\mathbf{B}} \mathbf{f}^T \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i \right) \right] \right)^2 \ .$$

Assume that Σ is positive definite. In this case, one can define a positive definite square root $\Sigma^{1/2}$ of Σ. Therefore, for any $\mathbf{a}$, $\mathbf{b}$ in $\mathbb{R}^d$,

$$\mathbf{a}^T\mathbf{b} = (\Sigma^{1/2}\mathbf{a})^T(\Sigma^{-1/2}\mathbf{b}) \leqslant (\mathbf{a}^T\Sigma\mathbf{a})^{1/2}(\mathbf{b}^T\Sigma^{-1}\mathbf{b})^{1/2} \ .$$

Defining, for any positive semi-definite matrix $\mathcal{M}$ and any vector $\mathbf{a} \in \mathbb{R}^d$, $\|\mathbf{a}\|_{\mathcal{M}} = \mathbf{a}^T\mathcal{M}\mathbf{a}$, it follows that

$$D_N(\mathbf{B}(r)) \leqslant 4L^2r^2\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i X_i\right\|_{\Sigma^{-1}}\right]\right)^2 \ .$$

By Cauchy Schwarz inequality,

$$\begin{aligned} D_N(\mathbf{B}(r)) &\leqslant 4L^2r^2\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i X_i\right\|_{\Sigma^{-1}}^2\right] \\ &= \frac{4L^2r^2}{N}\sum_{1\leqslant i,j\leqslant N}\mathbb{E}[\epsilon_i\epsilon_j X_i^T\Sigma^{-1}X_j] \\ &= 4L^2r^2\mathbb{E}[X^T\Sigma^{-1}X] = 4L^2r^2d \ . \end{aligned}$$

Hence, with probability larger that $1 - e^{-s}$,

$$\sup_{f\in\mathbf{B}(r)}(P_N - P)[\ell_{f^*} - \ell_f] \leqslant Lr\frac{4\sqrt{d} + \sqrt{2s}}{\sqrt{N}} \ .$$

This suggests to use in the homogeneity lemma the function

$$\theta(r) = Lr\frac{4\sqrt{d} + \sqrt{2s}}{\sqrt{N}} := rr_s \ .$$

Recall that $\sup_{f\in\mathbf{B}(r)} d(f^*, f) = 0$ and that, by the Bernstein Assumption (6.15), for any $r \leqslant A$,

$$\inf_{f\in\mathbf{S}(r)} d(f^*, f) \geqslant Br^2 \ .$$

By Lemma 65, it follows that (5.8) of the homogeneity lemma is satisfied with $r_1 = r_s/B$, $\zeta = \frac{a^2}{B^2}$ and (5.9) holds with $r_2 = 2r_1$. Of course, these results only holds if $r_2 \leqslant A$, which is true for any s satisfying Assumption (6.26). □

6.8 Minmax MOM estimators

This section extends the previous result to the case where the *design* is not assumed to be sub-Gaussian anymore. Indeed, Lipshitz losses are classically considered in robust statistics. This success, as explained after Theorem 86, is due to the fact that the ERM can be analysed in this framework *without assumptions on the outputs* Y. However, this analysis highly depends on the sub-Gaussian assumption made on the design. The extension is even more important to handle possibly corrupted datasets. Indeed, these data are likely to be corrupted, specially in high dimensional settings.

Consider the minmax MOM estimator

$$\hat{f}_K \in \operatorname{argmin}_{f\in F}\sup_{g\in F}\mathrm{MOM}_K[\ell_f - \ell_g]. \tag{6.27}$$

The main result here is the following.

Theorem 88. *Assume that the Bernstein assumption* (6.15) *holds for constants A and B such that*

$$AB\sqrt{N} \geqslant 1024L\sqrt{d} \ .$$

Then, for any K such that

$$1024L\big(\sqrt{d} \vee \sqrt{K}\big) \leqslant AB\sqrt{N} \ ,$$

the empirical risk minimizer (6.25) *satisfies*

$$\mathbb{P}\left(\mathcal{E}(\hat{f}_K) \leqslant \frac{1024L}{B}\sqrt{\frac{d \vee K}{N}}\right) \geqslant 1 - 2e^{-K/32} \ .$$

Proof. As for the ERM, the key is to compute the function θ in the homogeneity lemma. Let $r > 0$ fixed. Apply the concentration bound for suprema of MOM processes on the class of functions

$$F_r = \{\ell_{f^*} - \ell_f - P[\ell_{f^*} - \ell_f], f \in \mathbf{B}(r)\} \ .$$

With probability at least $1 - e^{-K/32}$,

$$\sup_{f \in \mathbf{B}(r)} \mathrm{MOM}_K[\ell_{f^*} - \ell_f - P[\ell_{f^*} - \ell_f]] \leqslant 128\sqrt{\frac{D(F_r)}{N}} \vee 4\sigma(F_r)\sqrt{2\frac{K}{N}} \ .$$

The computations of the previous proof show that $D(F_r) \leqslant 4L^2r^2d$, $\sigma^2(F_r) \leqslant L^2r^2$, hence, with probability at least $1 - e^{-K/32}$,

$$\sup_{f \in \mathbf{B}(r)} \mathrm{MOM}_K[\ell_{f^*} - \ell_f - P[\ell_{f^*} - \ell_f]] \leqslant 1024Lr\sqrt{\frac{d \vee K}{N}} \ .$$

This suggests to use

$$\theta(r) = rr_K, \qquad \text{with} \qquad r_K = 1024Lr\sqrt{\frac{d \vee K}{N}} \ .$$

The proof is concluded with the same arguments as the previous one. □

Chapter 7

Least-squares regression

This chapter considers the classical least-squares linear regression problem. This problem has attracted a lot of attention recently in the case where both the inputs X and the outputs Y may be heavy-tailed. The first paper proving oracle inequalities in this setting is [2]. The estimator there was derived from M-estimators of univariate expectations. Recent articles, in particular the seminal paper [39], see also [41, 33, 34], also investigate median-of-mean approaches in both small and large dimension least-squares regression. This analysis is reproduced in this chapter in the simplified setting of linear least-squares regression. The pros and cons of these approaches are the same as in the multivariate mean estimation problems, see the discussion in Section 4.4. All these results rely on either a L^4/L^2 or a L^2/L^1 comparison between the functions in the hypothesis class F that should hold uniformly for a constant that should not depend on the dimension of F. This last restriction typically fails in many important classes of functions of interest as explained in [51]. Section 7.4 presents two analyses of minmax MOM estimators, proving the statistical optimality of these estimators in a toy example in small dimension ($d \leqslant \sqrt{N}$) where the uniform L^2/L^1 comparison fails.

7.1 Setting

Consider the supervized statistical learning framework where the data space $\mathcal{Z}$ is a product $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$, so data $z \in \mathcal{Z}$ are couples $z = (x, y)$ and the goal is to predict as best as possible an output Y from an input X when $Z = (X, Y)$ is drawn from an unknown distribution P. For any $f \in \mathbb{R}^d$ and $z \in \mathcal{Z}$, let ℓ denote the square loss

$$\ell_f(z) = (y - x^T f)^2 \ .$$

Hereafter, we assume that P satisfies $P[Y^2] < \infty$ and $P[\|X\|^2] < \infty$ and measure the risk of any $f \in \mathbb{R}^d$ by

$$P\ell_f = P[(Y - X^T f)^2] \ .$$

As usual, $f^* \in \operatorname{argmin} P\ell_f$ denotes an oracle. Let $\Sigma = P[XX^T]$ and assume that

$$\Sigma \text{ is positive definite} \ . \tag{7.1}$$

Let $F \subset \mathbb{R}^d$ denote a convex subset of $\mathbb{R}^d$.

This chapter studies both ERM and minmax MOM estimators. As $f \mapsto \ell_f(x,y)$ is convex for any $z=(x,y)$, Lemma 66 applies and shows that the tests

$$T_{\mathrm{emp}}(f,g) = P_N[\ell_f - \ell_g], \qquad T_{\mathrm{mom}}(f,g) = \mathrm{MOM}_K[\ell_f - \ell_g] \ ,$$

satisfy Assumption **(HP)** of the homogeneity lemma, see Lemma 62, provided that the evaluation function $\mathcal{E}$ derives from a norm. Hereafter in this section, for any f and g in $\mathbb{R}^d$, let $d(f,g) = P[\ell_f - \ell_g]$ which satisfies $d(f,g) = -d(g,f)$ and $d(f^*,f) \leqslant 0$, so the functions B and $\mathcal{B}$ in Lemma 62 are equal. For the evaluation function, for any $f \in F$, let

$$\begin{aligned}\mathcal{E}(f) &= \|f-f^*\|_{L^2(P)} := \sqrt{P[(X^T(f-f^*))^2]} \\ &= \sqrt{(f-f^*)^T\Sigma(f-f^*)} = \|f-f^*\|_\Sigma \ ,\end{aligned}$$

where, for any $d\times d$ matrix $\mathcal{M}$ and any $\mathbf{a}\in\mathbb{R}^d$, $\|\mathbf{a}\|_{\mathcal{M}} = \sqrt{\mathbf{a}^T\mathcal{M}\mathbf{a}}$. Developing the square $((y-x^Tg)-x^T(f-g))^2$ shows the so called quadratic/multiplier decomposition of the square loss:

$$\ell_f(x,y) - \ell_g(x,y) = [x^T(f-g)]^2 - 2x^T(f-g)(y-x^Tg) \ . \tag{7.2}$$

Let $\xi = Y - X^Tf^*$, the process

$$f \mapsto \xi X^T(f-f^*) \ ,$$

is called the *multiplier* process and

$$f \mapsto (X^T(f-f^*))^2$$

the *quadratic* process. Let $t\in(0,1)$ and $f\in F$. By definition of f^*, as $(1-t)f^*+tf \in F$,

$$\begin{aligned}P[(Y-X^Tf^*)^2] &\leqslant P[(Y-X^T((1-t)f^*+tf))^2] \\ &= P[(\xi - tX^T(f-f^*))^2] \\ &= P[\xi^2] - 2tP[\xi X^T(f-f^*)] + t^2P[(X^T(f-f^*))^2] \ .\end{aligned}$$

It follows that, for any $t\in(0,1)$,

$$P[\xi X^T(f-f^*)] \leqslant \frac{t}{2}P[(X^T(f-f^*))^2] \ .$$

Letting $t\to 0$ shows that

$$P[\xi X^T(f-f^*)] \leqslant 0 \ . \tag{7.3}$$

Together with (7.2), this implies in particular that

$$P[\ell_f - \ell_{f^*}] = P[(X^T(f-f^*))^2] - 2P[\xi X^T(f-f^*)] \geqslant \|f-f^*\|_\Sigma^2 \ .$$

In particular, the following "global" Bernstein condition is satisfied for least-squares regression:

$$\forall f\in F, \qquad d(f,f^*) = P[\ell_f - \ell_{f^*}] \geqslant \mathcal{E}(f)^2 \ . \tag{7.4}$$

7.2 ERM in the Gaussian case

To establish the Benchmark, let us first consider the ERM estimator when $Z = (X, Y)$ is a Gaussian vector. Let

$$\overline{\Sigma} = P[(X - P[X])(X - P[X])^T], \quad \sigma^2 = P[(Y - X^T f^*)^2] \ .$$

Theorem 89. *Assume that $F = \mathbb{R}^d$ and $Z = (X, Y)$ is a Gaussian vector such that the covariance matrix of X in $\mathbb{R}^{d\times d}$ is non-degenerate. Assume moreover that $64d \leqslant \gamma N$ for $\gamma = \sqrt{2/\pi e}$. Then, for any $s > 0$ such that*

$$8\sqrt{d} + 2\sqrt{2s} \leqslant \sqrt{\gamma N} \ ,$$

the empirical risk minimizer $\hat{f} \in \operatorname{argmin}_{f\in F} P_N[(Y - f^T X)^2]$ satisfies

$$\mathbb{P}\left(\mathcal{E}(\hat{f}) \leqslant \frac{24\sigma}{\sqrt{\gamma^3 N}}\big(3\sqrt{d} + 4\sqrt{s}\big)\right) \geqslant 1 - 5e^{-s} \ .$$

Proof. The key is to compute the function B in the homogeneity lemma. Start with algebraic computations. Let $r > 0$ and $f \in F$ such that $\mathcal{E}(f) = \|f - f^*\|_\Sigma \leqslant r$. Then $f = f^* + rg$ with $g = (f - f^*)/r$ satisfying $\|g\|_\Sigma \leqslant 1$. Then

$$\begin{aligned}(P_N - P)\big[2\xi X^T(f - f^*) - [X^T(f - f^*)]^2\big] \\ = 2r(P_N - P)\big[\xi X^T g)\big] - r^2(P_N - P)\big[(X^T g)^2\big] \ . \end{aligned} \tag{7.5}$$

Let $\mathbf{B} = \{f \in \mathbb{R}^d : \|f\|_\Sigma \leqslant 1\}$,

$$\begin{aligned} M &= \sup_{f\in\mathbf{B}} (P_N - P)\big[\xi X^T f\big] \\ Q &= \inf_{f\in\mathbf{B}} (P_N - P)\big[(X^T f)^2\big] \ . \end{aligned}$$

With these notations, from (7.5),

$$\sup_{f\in\mathbf{B}(r)} (P_N - P)\big[\ell_f - \ell_{f^*}\big] \leqslant 2rM - r^2 Q \ . \tag{7.6}$$

Lemma 90. *For any $s \in [0, N]$, with probability $1 - 4e^{-s}$,*

$$\mathbb{P}\left(M \leqslant \frac{\sigma}{\sqrt{N}}\big(3\sqrt{d} + 4\sqrt{s}\big)\right) \geqslant 1 - 4e^{-s} \ .$$

Proof. Let $f \in \mathbf{B}$. As $Z = (X, Y)$ is a Gaussian vector and $F = \mathbb{R}^d$, $X^T f^*$ is the projection of Y onto the linear span of X in L^2. Therefore, $X^T f$ is, conditionally on ξ, a Gaussian random variable, with mean $P[X^T f]$ and variance $P[((X - P[X])^T f)^2] = f^T \overline{\Sigma} f = \|f\|^2_{\overline{\Sigma}} \leqslant 1$. Let $\mathcal{F}_N$ denote the σ-algebra spanned by $\xi_1, \ldots, \xi_N$. Conditionally on $\mathcal{F}_N$, the random variables $X_f = P_N[\xi X^T f]$ are Gaussian random variables centered at $P_N[\xi]P[X^T f]$ with variance

$$\sigma_f^2 = \frac{P_N[\xi^2]}{N} f^T \overline{\Sigma} f \leqslant V \ , \tag{7.7}$$

where $V = N^{-1}P_N[\xi^2]$. By concentration of suprema of Gaussian processes, for any $s > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathbf{B}}(X_f - \mathbb{E}[X_f|\mathcal{F}_N]) \leqslant \mathbb{E}[\sup_{f\in F}(X_f - \mathbb{E}[X_f|\mathcal{F}_N])|\mathcal{F}_N] + \sqrt{2Vs}|\mathcal{F}_N \right) \leqslant 1 - e^{-s} \ .$$

Now,

$$\mathbb{E}[\sup_{f\in\mathbf{B}}(X_f - \mathbb{E}[X_f|\mathcal{F}_N])|\mathcal{F}_N] = \mathbb{E}[\sup_{f\in\mathbf{B}} P_N[\xi(X - P[X])^T f]|\mathcal{F}_N]$$

Now, as Σ is non degenerate, $\|\cdot\|_\Sigma$ is a norm whose dual norm is $\|\cdot\|_{\Sigma^{-1}}$. Hence,

$$\begin{aligned}\mathbb{E}[\sup_{f\in\mathbf{B}}(X_f - \mathbb{E}[X_f|\mathcal{F}_N])|\mathcal{F}_N] &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \xi_i(X_i - P[X])\right\|_{\Sigma^{-1}}|\mathcal{F}_N\right] \\ \text{by Cauchy-Schwarz } &\leqslant \sqrt{\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \xi_i(X_i - P[X])\right\|^2_{\Sigma^{-1}}|\mathcal{F}_N\right]} \ .\end{aligned}$$

Now, developing the square-norm and using the independence between ξ and X,

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \xi_i(X_i - P[X])\right\|^2_{\Sigma^{-1}}|\mathcal{F}_N\right] \\ &= \frac{1}{N^2}\sum_{1\leqslant i,j\leqslant N} \xi_i\xi_j\mathbb{E}[(X_i - P[X])^T\Sigma^{-1}(X_j - P[X])] \\ &= \frac{1}{N^2}\sum_{i=1}^N \xi_i^2\mathbb{E}[(X_i - P[X])^T\Sigma^{-1}(X_i - P[X])] \\ &\leqslant \frac{P_N[\xi^2]}{N}P[X^T\Sigma^{-1}X] \ .\end{aligned}$$

Finally,

$$P[X^T\Sigma^{-1}X] = P[\mathrm{Tr}(X^T\Sigma^{-1}X)] = P[\mathrm{Tr}(\Sigma^{-1}XX^T)] = \mathrm{Tr}(\mathbf{I}_d) = d \ . \quad (7.8)$$

Therefore,

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \xi_i(X_i - P[X])\right\|^2_{\Sigma^{-1}}|\mathcal{F}_N\right] \leqslant \frac{P_N[\xi^2]}{N}d \ ,$$

so

$$\mathbb{E}[\sup_{f\in\mathbf{B}}(X_f - \mathbb{E}[X_f|\mathcal{F}_N])|\mathcal{F}_N] \leqslant \sqrt{Vd} \ .$$

Overall, with probability at least $1 - e^{-s}$,

$$M = \sup_{f\in\mathbf{B}} X_f \leqslant |P_N[\xi]| + \sqrt{V}\left(\sqrt{d} + \sqrt{2s}\right) \ . \quad (7.9)$$

Now, with probability $1 - 2e^{-s}$, the centered Gaussian random variable $P_N[\xi]$ satisfies

$$|P_N[\xi]| \leqslant \sigma\sqrt{\frac{2s}{N}} \ . \quad (7.10)$$

Moreover, $\xi \sim \mathrm{N}(0, \sigma^2)$, so $\mathbb{E}[\xi^{2k}] = (2k)!\sigma^{2k}/2^k k!$ and, for any $u < 1/2\sigma^2$,

$$\begin{aligned}\mathbb{E}[e^{u\xi^2}] &= 1 + u\sigma^2 + \sum_{k\geqslant 2} \frac{u^k (2k)! \sigma^{2k}}{2^k (k!)^2} \\ &\leqslant 1 + u\sigma^2 + \sum_{k\geqslant 2} (2u\sigma^2)^k \\ &= 1 + u\sigma^2 + \frac{4u^2\sigma^4}{1 - 2u\sigma^2} \ .\end{aligned}$$

Hence, for any $u < N/2\sigma^2$,

$$\log \mathbb{E}\big[e^{u(NV-\sigma^2)}\big] \leqslant N \log\left(1 + \frac{4(u/N)^2\sigma^4}{1 - 2(u/N)\sigma^2}\right) \leqslant \frac{u^2 8\sigma^4/N}{2(1 - u2\sigma^2/N)} \ .$$

It follows therefore from Bernstein's inequality that, for any $s > 0$,

$$\mathbb{P}\left(NV - \sigma^2 > 2\sigma^2\left(2\sqrt{\frac{s}{N}} + \frac{s}{N}\right)\right) \leqslant e^{-s} \ .$$

Plugging this bound and (7.10) into (7.9) shows that, for any $s \leqslant N$, with probability at least $1 - 4e^{-s}$,

$$M \leqslant \sigma\sqrt{\frac{2s}{N}} + \left(\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\right)\sigma\sqrt{1 + 4\sqrt{\frac{s}{N}} + \frac{2s}{N}} \leqslant \sigma\left(3\sqrt{\frac{d}{N}} + 4\sqrt{\frac{s}{N}}\right) \ .$$

□

Let us now bound the quadratic process.

Lemma 91.

$$\forall s > 0, \qquad \mathbb{P}\left(Q < \frac{\gamma}{2} - 1 - \left(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\right)^2\right) \leqslant e^{-s} \ .$$

Proof. Elementary calculus shows that, for any real valued Gaussian random variable Z,

$$\mathbb{E}[|Z|] \geqslant \gamma\sqrt{\mathbb{E}[Z^2]} \ ,$$

where $\gamma = \sqrt{2/\pi e}$. From this remark follows that the class of linear functions satisfies the *small ball assumption* of Mendelson: as $X^T f$ is Gaussian for any $f \in \mathbb{R}^d$,

$$\forall f \in F, \qquad P[|X^T f|] \geqslant \gamma\sqrt{P[(X^T f)^2]} = \gamma\|f\|_\Sigma \ . \tag{7.11}$$

By Jensen's inequality,

$$P_N\big[(X^T f)^2\big] \geqslant \big(P_N[|X^T f|]\big)^2 \ .$$

Let $f \in \mathbf{B}$,

$$\mathrm{Var}(X^T f) = f^T \overline{\Sigma} f \leqslant 1 \ .$$

Now, by Borel's concentration inequality, with probability at least $1 - e^{-s}$,

$$\sup_{f\in\mathbf{B}} \big|(P_N - P)|X^T f|\big| \leqslant \mathbb{E}\left[\sup_{f\in\mathbf{B}} \big|(P_N - P)|X^T f|\big|\right] + \sqrt{\frac{2s}{N}}$$

By symmetrization and contraction,

$$\sup_{f\in\mathbf{B}} \big|(P_N - P)|X^T f|\big| \leqslant 4\mathbb{E}\Big[\sup_{f\in\mathbf{B}} \big|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_i^T f\big|\Big] + \sqrt{\frac{2s}{N}}$$
$$= 4\mathbb{E}\Big[\sup_{f\in\mathbf{B}} \big|f^T\Big(\frac{1}{N}\sum_{i=1}^N \epsilon_i X_i\Big)\big|\Big] + \sqrt{\frac{2s}{N}} \ .$$

Using that $\|\cdot\|_\Sigma$ is a norm with dual norm $\|\cdot\|_{\Sigma^{-1}}$,

$$\sup_{f\in\mathbf{B}} \big|(P_N - P)|f^T X|\big| = 4\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_i\Big\|_{\Sigma^{-1}}\Big] + \sqrt{\frac{2s}{N}}$$
$$\text{by Cauchy-Schwarz } \leqslant 4\sqrt{\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N \epsilon_i X_i\Big\|_{\Sigma^{-1}}^2\Big]} + \sqrt{\frac{2s}{N}} \ .$$

Developing the square, using independence between the ϵ_i and X_i and that $\mathbb{E}[\epsilon_i] = 0$,

$$\sup_{f\in\mathbf{B}} \big|(P_N - P)|X^T f|\big| = 4\sqrt{\frac{1}{N^2}\sum_{1\leqslant i,j\leqslant N} \mathbb{E}\Big[\epsilon_i\epsilon_j X_i^T\Sigma^{-1}X_j\Big]} + \sqrt{\frac{2s}{N}}$$
$$= 4\sqrt{\frac{1}{N}P\big[X^T\Sigma^{-1}X\big]} + \sqrt{\frac{2s}{N}} \ .$$

By (7.8), it follows that $\mathbb{P}(\Omega_s) \geqslant 1 - e^{-s}$, where

$$\Omega_s = \Big\{\sup_{f\in\mathbf{B}} \big|(P_N - P)|X^T f|\big| \leqslant 4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\Big\} \ .$$

On Ω_s, for any $f \in F$,

$$P_N\big[(X^T f)^2\big] \geqslant \Big(P[|X^T f|] - 4\sqrt{\frac{d}{N}} - \sqrt{\frac{2s}{N}}\Big)^2$$
$$\geqslant \frac{1}{2}(P[|X^T f|])^2 - \Big(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\Big)^2$$
$$\geqslant \frac{\gamma}{2}\|f\|_\Sigma^2 - \Big(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\Big)^2 \ .$$

It follows that

$$(P_N - P)\big[(X^T f)^2\big] \geqslant \Big(\frac{\gamma}{2} - 1\Big)\|f\|_\Sigma^2 - \Big(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\Big)^2 \ .$$

Therefore, as $\gamma/2 - 1 < 0$,

$$\forall s > 0, \qquad \mathbb{P}\Big(Q \geqslant \frac{\gamma}{2} - 1 - \Big(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\Big)^2\Big) \geqslant 1 - e^{-s} \ .$$

□

Let

$$m_s = \frac{\sigma}{\sqrt{N}}\big(3\sqrt{d} + 4\sqrt{s}\big), \qquad q_s = \frac{\gamma}{2} - \left(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\right)^2 \ .$$

It follows from Lemmas 90 and 91 that the event $\Omega = \{M \leqslant m_s\} \cap \{Q \geqslant q_s - 1\}$ has probability larger than $1 - 5e^{-s}$. Moreover, from (7.6), Ω contains $\cap_{r>0}\Omega_r$, where

$$\Omega_r = \left\{ \sup_{f \in F : \mathcal{E}(f) \leqslant r} (P_N - P)\big[\ell_f - \ell_{f^*}\big] \leqslant B(r) \right\} \ ,$$

with $B(r) = 2rm_s - r^2(q_s - 1)$. With this choice of function B, by (7.4), it follows that (5.8) holds if $q_s > 0$ and

$$2r_1 m_s - r_1^2 q_s \leqslant 0, \quad \text{i.e.} \quad r_1 \geqslant \frac{2m_s}{q_s} \ .$$

Let

$$r_1 = \frac{2m_s}{q_s} \quad \text{so} \quad B(r_1) = \frac{4m_s^2}{q_s} - \frac{4m_s^2(q_s - 1)}{q_s^2} = \frac{4m_s^2}{q_s^2} \ .$$

Then, (5.9) holds if $q_s > 0$ and

$$\frac{4m_s^2}{q_s^2} + 2r_2 m_s - r_2^2 q_s = \frac{4m_s^2}{q_s^2} + \frac{m_s^2}{q_s} - q_s(r_2 - \frac{m_s}{q_s})^2 \leqslant 0 \ ,$$

that is if

$$r_2 = \frac{2m_s}{q_s}\left(1 + \frac{1}{\sqrt{q_s}}\right) \ .$$

As

$$\left(4\sqrt{\frac{d}{N}} + \sqrt{\frac{2s}{N}}\right)^2 \leqslant \frac{\gamma}{4}, \qquad q_s \geqslant \frac{\gamma}{4} \ .$$

Therefore,

$$r_2 \leqslant \frac{24}{\gamma^{3/2}} m_s = \frac{24\sigma}{\sqrt{\gamma^3 N}}\big(3\sqrt{d} + 4\sqrt{s}\big) \ .$$

The proof is concluded by Lemma 62. □

7.3 Minmax MOM estimators

In the previous section, we used several features of Gaussian distributions to prove the deviation bound in least-squares regression. The first of these properties is that, since (X, Y) was Gaussian, the vector $f^* \in \operatorname{argmin}_{f \in F} P[\|Y - X^T f\|^2]$ satisfies $X^T f^* = \mathbb{E}[Y|X]$ and one can write

$$Y = X^T f^* + \xi \ ,$$

where ξ/σ is a standard Gaussian *independent from* X. It follows that

$$\forall f \in \mathbb{R}^d, \qquad \operatorname{Var}\big((X^T f)\xi\big) = \sigma^2 \operatorname{Var}\big(X^T f\big) = \sigma^2 \|f\|_\Sigma^2 \ .$$

Therefore

$$\sigma^2 \geqslant \sup_{f \in F : \|f\|_\Sigma = 1} P\big(\xi^2 (X^T f)^2\big) \ . \tag{7.12}$$

The independence between ξ and X also allows to show that

$$P[\xi^2 \sup_{f\in F:\|f\|_\Sigma\leqslant 1}(X^T f)^2] = P[\xi^2\|X\|^2_{\Sigma^{-1}}] = \sigma^2 P[\|X\|^2_{\Sigma^{-1}}] = \sigma^2 d \ .$$

The last inequality comes from (7.8). Hence,

$$\sigma^2 \geqslant \frac{P[\xi^2\|X\|^2_{\Sigma^{-1}}]}{d} \ . \tag{7.13}$$

It turns out that independence between the noise ξ and the inputs X can be removed provided that $\sigma = P[\xi^2]$ is replaced by the adequate quantity in conditions (7.12) and (7.13). Hereafter, denote by $\overline{\sigma}$ a positive real number such that

$$\overline{\sigma}^2 \geqslant \sup_{f\in F:\|f\|_\Sigma=1} P\big(\xi^2(X^T f)^2\big) \vee \frac{P[\xi^2\|X\|^2_{\Sigma^{-1}}]}{d} \ . \tag{7.14}$$

The parameter $\overline{\sigma}$ does not appear in the construction of the minmax estimator and may therefore be unknown from the statistician. Notice that, as ξ and X may not be independent, it is implicitly assumed that $\overline{\sigma} < +\infty$ in the following.

7.3.1 The small ball hypothesis

The second property of Gaussian distributions was the small ball property [32, 45], see Eq (7.11) in the previous proof. To extend the Gaussian case, we will *assume* that this property holds for the distribution of the vector X. Formally, there exists an absolute constant $\gamma > 0$ such that

$$\forall f \in F, \qquad P[|X^T f|] \geqslant \gamma\sqrt{P[(X^T f)^2]} = \|f\|_\Sigma \ . \tag{7.15}$$

This assumption is checked in the following example.

Lemma 92. *Assume that the random vector X has coordinates $(X^{(i)})_{1\leqslant i\leqslant N}$ satisfying the following property. There exist constants $C_1, C_2 > 0$ such that, $\forall 1 \leqslant i, j \leqslant N$,*

$$\mathbb{E}[X^{(i)}X^{(j)}] \leqslant C_1\mathbb{E}[|X^{(i)}|]\mathbb{E}[|X^{(j)}|] \ , \tag{7.16}$$

$$\sum_{i=1}^N |f_i|\mathbb{E}[|X^{(i)}|] \leqslant C_2 P[|X^T f|] \ . \tag{7.17}$$

Then, (7.15) holds with $\gamma = 1/\sqrt{C_1}C_2$.

Proof. Let $f \in \mathbb{R}^d$.

$$\begin{aligned}
\|f\|^2_\Sigma &= \sum_{1\leqslant i,j\leqslant N} f_i f_j \mathbb{E}[X^{(i)}X^{(j)}] \\
&\leqslant C_1\bigg(\sum_{i=1}^N |f_i|\mathbb{E}[|X^{(i)}|]\bigg)^2 \text{ by (7.16)} \\
&\leqslant C_1 C_2^2 (P[|X^T f|])^2 \text{ by (7.17)} \ .
\end{aligned}$$

Therefore, (7.15) holds with $\gamma = 1/\sqrt{C_1}C_2$. □

Another example where one can check the small ball property is the following.

Lemma 93. *Assume that the L^4/L^2 comparison holds.*

$$\exists C > 0 : \forall f \in F, \qquad P[(X^T f)^4] \leqslant C P[(X^T f)^2]^2 \ . \tag{7.18}$$

Then, (7.15) *holds with* $\gamma = \sqrt{2}/8C$.

Remark 94. *Assumption is discussed in Section 6.6 of the previous chapter. It was used there to check the Bernstein assumption.*

Proof. The proof relies on the following simple Paley-Zigmund argument. Let $f \in F$,

$$\begin{aligned}
P[(X^T f)^2] &= P[(X^T f)^2 \mathbf{1}_{|X^T f| \leqslant 4CP[|X^T f|]}] + P[(X^T f)^2 \mathbf{1}_{|X^T f| > 4CP[|X^T f|]}] \\
\text{by Cauchy-Schwarz} &\leqslant 16C^2 P[|X^T f|]^2 + \sqrt{P[(X^T f)^4] \mathbb{P}\big(|X^T f| > 4CP[|X^T f|]\big)} \\
\text{by Markov} &\leqslant 16C^2 P[|X^T f|]^2 + \sqrt{\frac{P[(X^T f)^4]}{4C}} \\
\text{by (7.18)} &\leqslant 16C^2 P[|X^T f|]^2 + \frac{P[(X^T f)^2]}{2} \ .
\end{aligned}$$

It follows that

$$P[(f^T X)^2] \leqslant 32C^2 P[|f^T X|]^2 \ .$$

In other words, (7.15) holds with $\gamma = \sqrt{2}/8C$. □

7.3.2 Main results

Recall that we observe $(X_1, Y_1), \ldots, (X_N, Y_N)$ i.i.d. copies of (X, Y), a random vector taking values in $\mathbb{R}^d \times \mathbb{R}$, that $\Sigma = P[XX^T]$, $\mathbb{E}[Y^2] < \infty$, $f^* \in \operatorname{argmin}_{f \in F} P\ell_f$, where $\ell_f(x, y) = (y - f^T x)^2$ and $\xi = Y - X^T f^*$.

Theorem 95. *Let $\overline{\sigma}$ be defined in* (7.14) *and assume that* (7.15) *holds. There exists an absolute constant C such that, if $Cd \leqslant \gamma^2 N$, then, for any K such that $CK \leqslant \gamma^2 N$, the minmax MOM estimator*

$$\widehat{f}_K \in \operatorname{argmin}_{f \in F} \sup_{g \in F} MOM_K[\ell_f - \ell_g]$$

satisfies

$$\mathbb{P}\left(\|\widehat{f}_K - f^*\|_\Sigma \leqslant \frac{C\overline{\sigma}}{\gamma^3 \sqrt{N}} \big(\sqrt{d} \vee \sqrt{K} \big) \right) \geqslant 1 - 4e^{-K/C} \ .$$

Proof. The key is to compute the function B in the homogeneity lemma. Let $r > 0$ and $F_r = \{f \in F : \|f - f^*\|_\Sigma \leqslant r\}$. By the quadratic/multiplier decomposition of the quadratic loss (7.2), one wants to bound from above

$$\mathcal{M}_r = \sup_{f \in F_r} \mathrm{MOM}_K \big[2\xi X^T (f - f^*) - (X^T (f - f^*))^2 + P[(X^T (f - f^*))^2] \big] \ .$$

As $F_r = \{f = f^* + ru, \ u \in \mathbf{B}\}$, with $\mathbf{B} = \{u \in \mathbb{R}^d : \|u\|_\Sigma \leqslant 1\}$, it holds

$$\mathcal{M}_r = \sup_{u \in \mathbf{B}} \mathrm{MOM}_K \big[2r\xi [X^T u] + r^2 (\|u\|_\Sigma^2 - (X^T u)^2) \big] \ .$$

To bound $\mathcal{M}_r$, the following lemmas will prove useful.

Lemma 96. *There exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where*

$$\sup_{u \in \boldsymbol{B}} (P_{B_k} - P)[\xi X^T u] = \sup_{u \in \boldsymbol{B}} P_{B_k}[\xi X^T u] \leqslant \frac{c^* \overline{\sigma}}{\sqrt{N}} \left(\sqrt{d} \vee \sqrt{K} \right) \ .$$

Proof. Consider the set of functions $\mathcal{F}_M = \{(x, y) \mapsto (y - x^T f^*)(x^T u),\ u \in \mathbf{B}\}$. By definition of $\overline{\sigma}$, see (7.14),

$$\forall u \in \mathbf{B}, \qquad \sigma_u^2 = P[\xi^2 (X^T u)^2] \leqslant \overline{\sigma}^2 \ .$$

Hence, $\sigma^2(\mathcal{F}_M) = \sup_{f \in \mathcal{F}_M} \mathrm{Var}(f(Z)) \leqslant \overline{\sigma}^2$. Moreover, as $\| \cdot \|_{\Sigma^{-1}}$ is the dual norm of $\| \cdot \|_\Sigma$, one can bound the Rademacher complexity of $\mathcal{F}_M$ as follows.

$$\begin{aligned} D(\mathcal{F}_M) &= \left(\mathbb{E} \left[\sup_{u \in \mathbf{B}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i \xi_i (X_i^T u) \right] \right)^2 \\ &= \left(\mathbb{E} \left[\sup_{u \in \mathbf{B}} u^T \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i \xi_i X_i \right) \right] \right)^2 \\ &= \left(\mathbb{E} \left[\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i \xi_i X_i \right\|_{\Sigma^{-1}} \right] \right)^2 \ . \end{aligned}$$

By Cauchy-Schwarz,

$$D(\mathcal{F}_M) \leqslant \mathbb{E} \left[\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i \xi_i X_i \right\|_{\Sigma^{-1}}^2 \right] \ .$$

Developing the square norm and using independence and centering,

$$D(\mathcal{F}_M) \leqslant \mathbb{E} \big[\big\| \xi X \big\|_{\Sigma^{-1}}^2 \big] = P \big[\xi^2 \big\| X \big\|_{\Sigma^{-1}}^2 \big] \leqslant \overline{\sigma}^2 d \ . \tag{7.19}$$

The last inequality uses the definition of $\overline{\sigma}$, see (7.14). By the general concentration result for quantile of means processes, see Theorem 40, there exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where

$$\sup_{u \in \mathbf{B}} (P_{B_k} - P)[\xi X^T u] \leqslant \frac{c^* \overline{\sigma}}{\sqrt{N}} \left(\sqrt{d} \vee \sqrt{K} \right) \ .$$

□

Lemma 97. *There exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where, for any $u \in \boldsymbol{B}$,*

$$P_{B_k}[(X^T u)^2] \geqslant \left(\gamma \|u\|_\Sigma - \frac{c^* \overline{\sigma}}{\sqrt{N}} \big(\sqrt{d} \vee \sqrt{K} \big) \right)_+^2 \ .$$

Proof. Consider $\mathcal{F}_Q = \{x \mapsto |x^T u|,\ u \in \mathbf{B}\}$.

$$\forall u \in \mathbf{B}, \qquad \sigma_u^2 = P[(X^T u)^2] \leqslant 1 \ .$$

Hence, $\sigma^2(\mathcal{F}_Q) = \sup_{f\in\mathcal{F}_Q} \mathrm{Var}(f(X)) \leqslant 1$. Moreover, by the contraction principle, the Rademacher complexity of $\mathcal{F}_Q$ can be upper bounded as follows.

$$D(\mathcal{F}_Q) = \left(\mathbb{E}\left[\sup_{u\in\mathbf{B}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i |X_i^T u|\right]\right)^2 \leqslant 4\left(\mathbb{E}\left[\sup_{u\in\mathbf{B}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i^T u\right]\right)^2 .$$

As $\|\cdot\|_{\Sigma^{-1}}$ is the dual norm of $\|\cdot\|_\Sigma$,

$$D(\mathcal{F}_Q) \leqslant 4\left(\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i\right\|_{\Sigma^{-1}}\right]\right)^2 \leqslant 4\mathbb{E}\left[\left\|\frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i\right\|_{\Sigma^{-1}}^2\right] .$$

Developing the square and using independence yields

$$D(\mathcal{F}_Q) \leqslant 4P[\|X\|_{\Sigma^{-1}}^2] = 4d \ . \tag{7.20}$$

The last equality comes from (7.8). By Theorem 40, there exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, at least $9K/10$ blocks B_k satisfy

$$\sup_{u\in\mathbf{B}} |(P_{B_k} - P)[|X^T u|]| \leqslant \frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big) \ .$$

Moreover, $P[|u^T X|] \geqslant \gamma\|u\|_\Sigma$, therefore, with probability at least $1 - e^{-K/c^*}$, for any $u \in \mathbf{B}$, there exist at least $9K/10$ blocks B_k where

$$P_{B_k}[(X^T u)^2] \geqslant (P_{B_k}[|X^T u|])^2 \geqslant \left(\gamma\|u\|_\Sigma - \frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big)\right)_+^2 .$$

□

Denote by c^* the largest of the absolute constants appearing in Lemmas 96 and 97. Define Ω as the event where, simultaneously, there exist $9K/10$ blocks B_k where

$$\sup_{u\in\mathbf{B}} P_{B_k}[\xi X^T u] \leqslant \frac{c^*\overline{\sigma}}{\sqrt{N}}\left(\sqrt{d} \vee \sqrt{K}\right) =: m_K \ ,$$

and $9K/10$ blocks B_k where, for any $u \in \mathbf{B}$,

$$P_{B_k}[(X^T u)^2] \geqslant \left(\gamma\|u\|_\Sigma - \frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big)\right)_+^2 .$$

By Lemmas 96 and 97, $\mathbb{P}(\Omega) \geqslant 1 - 2e^{-K/c^*}$. On Ω, there exist at least $9K/10$ blocks where, for any $u \in \mathbf{B}$,

$$\|u\|_\Sigma^2 - P_{B_k}[(X^T u)^2] \leqslant \|u\|_\Sigma^2 - \left(\gamma\|u\|_\Sigma - \frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big)\right)_+^2 .$$

Assume that

$$\frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big) \leqslant \frac{\gamma}{2} \ ,$$

As the functions $u \mapsto u^2 - (\alpha u - \beta)_+^2$, for $\alpha < 1$ are non-decreasing on $[0,1]$, it follows that, on Ω, there exist at least $9K/10$ blocks where, for any $u \in \mathbf{B}$,

$$\|u\|_\Sigma^2 - P_{B_k}[(X^T u)^2] \leqslant \|u\|_\Sigma^2 - \left(\gamma\|u\|_\Sigma - \frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big)\right)_+^2$$
$$\leqslant 1 - \left(\gamma - \frac{c^*}{\sqrt{N}}\big(\sqrt{d} \vee \sqrt{K}\big)\right)_+^2 \leqslant 1 - \frac{\gamma^2}{4} \ .$$

It follows that, on Ω, there exists at least $8K/10$ blocks where, simultaneously, for any $u \in \mathbf{B}$,

$$P_{B_k}[\xi X^T u] \leqslant m_K, \qquad \|u\|_\Sigma^2 - P_{B_k}[(X^T u)^2] \leqslant 1 - \frac{\gamma^2}{4} \ .$$

On these blocks,

$$\forall r > 0, \qquad P_{B_k}\big[2r\xi[X^T u] + r^2(\|u\|_\Sigma^2 - (X^T u)^2)\big] \leqslant 2rm_K + (1 - \gamma^2/4)r^2 \ .$$

As this relationship holds on more than $K/2$ blocks, it holds for the median, so Ω contains $\cap_{r>0}\Omega_r$, where

$$\forall r > 0, \qquad \Omega_r = \{\mathcal{M}_r \leqslant B(r)\}, \qquad B(r) = 2rm_K + \left(1 - \frac{\gamma^2}{4}\right)r^2 \ .$$

With this choice of function B, by (7.4), it follows that (5.8) holds if

$$2r_1 m_K - r_1^2 \frac{\gamma^2}{4} \leqslant 0, \quad \text{i.e.} \quad r_1 \geqslant \frac{8m_K}{\gamma^2} \ .$$

Let

$$r_1 = \frac{8m_K}{\gamma^2} \quad \text{so} \quad B(r_1) = \frac{16m_K^2}{\gamma^2} + \left(1 - \frac{\gamma^2}{4}\right)\frac{64m_K^2}{\gamma^4} = \frac{64m_K^2}{\gamma^4} \ .$$

Then, (5.9) holds if

$$\frac{64m_K^2}{\gamma^4} + 2r_2 m_K - r_2^2\frac{\gamma^2}{4} = \frac{64m_K^2}{\gamma^4} + \frac{4m_K^2}{\gamma^2} - \frac{\gamma^2}{4}\left(r_2 - \frac{4m_K}{\gamma^2}\right)^2 \leqslant 0 \ ,$$

that is if

$$r_2 = \frac{8m_K}{\gamma^2}\left(1 + \frac{1}{\gamma}\right) \leqslant \frac{16m_K}{\gamma^3} = \frac{16c^*\overline{\sigma}}{\gamma^3\sqrt{N}}\left(\sqrt{d} \vee \sqrt{K}\right) \ .$$

The proof is concluded by Lemma 62. □

7.4 Saumard's problem

This section discusses the problem of least-squares regression in the case where the (elegant as it only involves L^1 and L^2 moments) Assumption 7.15 does not hold uniformly. Let us first get convinced that this problem naturally arises in important examples. Consider the following toy-model where the observations

$(\tilde{X}, Y)$ take values in $\mathcal{X} \times \mathcal{Y}$ and denote by $I_1, \dots, I_d$ a partition of $\mathcal{X}$ such that, for any $i \in \{1, \dots, d\}$, $P[I_i] = 1/d$. Let $\varphi_i(x) = \mathbf{1}_{x \in I_i}$, for any $x \in \mathcal{X}$, $i \in \{1, \dots, d\}$. Let

$$X = \begin{bmatrix} \varphi_1(\tilde{X}) \\ \vdots \\ \varphi_d(\tilde{X}) \end{bmatrix} \in \mathbb{R}^d \ . \tag{7.21}$$

Let $\mathcal{D}_N = (Z_1, \dots, Z_N)$ denote a dataset of i.i.d. copies of $Z = (\tilde{X}, Y)$ and, for any $i \in \{1, \dots, N\}$, let

$$X_i = \begin{bmatrix} \varphi_1(\tilde{X}_i) \\ \vdots \\ \varphi_d(\tilde{X}_i) \end{bmatrix} \in \mathbb{R}^d \ .$$

For any $f \in \mathbb{R}^d$, denoting by $\|f\|_p$ its ℓ_p norm,

$$P[|X^T f|] = P[|\sum_{i=1}^d f_i \varphi_i|] = \sum_{i=1}^d |f_i| P\varphi_i = \frac{\|f\|_1}{d} \ ,$$
$$P[(X^T f)^2] = P[(\sum_{i=1}^d f_i \varphi_i)^2] = P[\sum_{i=1}^d f_i^2 \varphi_i] = \frac{\|f\|_2^2}{d} \ .$$

As $\|f\|_1^2 \geqslant \|f\|_2^2$ (this bound is tight if f is the first element of the canonical basis of $\mathbb{R}^d$), Assumption 7.15 holds with $\gamma = 1/\sqrt{d}$ and $\delta = 1$. Therefore, Theorem 95 does not provide optimal rates of convergence in this example. In [51], Saumard showed that this problem does not hold only on histogram or localized basis, but basically on any space generated by functions $\varphi_1, \dots, \varphi_d$ with reasonable approximation propoerties. The reason is that these spaces are naturally designed to be able to reproduce many functions, in particular "spiky ones" for which the L^2/L^1 comparison does not hold uniformly.

7.4.1 First least-squares analysis of histograms.

The suboptimality in the rates provided in Theorem 95 comes from the analysis of the quadratic process. Improving these rates require modifications of Lemma 97 using properties of histogram spaces that will be the subject of this section. Start with the following rough alternative. The vector X defined in (7.21) satisfies

$$\Sigma = P[XX^T] = \frac{1}{d} \mathbf{I}_d \ .$$

Therefore, for any $u \in \mathbb{R}^d$, $\|u\|_\Sigma = \|u\|/\sqrt{d}$. Let

$$\mathbf{B} = \{u \in \mathbb{R}^d : \|u\|_\Sigma \leqslant 1\} = \{u \in \mathbb{R}^d : \|u\| \leqslant \sqrt{d}\} \ .$$

Lemma 98. *Consider the design vector X defined in* (7.21)*. There exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where, for any $u \in \boldsymbol{B}$,*

$$P_{B_k}[(X^T u)^2] \geqslant P[(X^T u)^2] - \frac{c^* \overline{\sigma}}{\sqrt{N}} \big(d \vee \sqrt{dK}\big) \ .$$

Proof. Consider $\mathcal{F}_Q = \{x \mapsto (x^T u)^2,\ u \in \mathbf{B}\}$. The sup norm of any function in $\mathcal{F}_Q$ can be bounded from above as follows:

$$\|(x^T u)^2\|_\infty = \sup_{\tilde{x}\in\mathcal{X}} \Big(\sum_{i=1}^d u_i \varphi_i(\tilde{x}) \Big)^2 = \sup_{\tilde{x}\in\mathcal{X}} \sum_{i=1}^d u_i^2 \varphi_i(\tilde{x}) = \max_{1\leqslant i\leqslant d} u_i^2 \ .$$

As $u \in \mathbf{B}$, $\max_{i\in\{1,\dots,d\}} u_i^2 \leqslant \|u\|^2 \leqslant d$. Hence,

$$\forall u \in \mathbf{B}, \qquad \sigma_u^2 \leqslant P[(X^T u)^4] \leqslant \|(x^T u)^2\|_\infty P[(X^T u)^2] \leqslant d \ .$$

Hence, $\sigma^2(\mathcal{F}_Q) = \sup_{f\in\mathcal{F}_Q} \mathrm{Var}(f(X)) \leqslant d$. Moreover, the functions $x \mapsto x^T u$ take values in $[-\sqrt{d}, \sqrt{d}]$ and the function $x \mapsto x^2$ is $2\sqrt{d}$ Lipschitz on $[-\sqrt{d}, \sqrt{d}]$. Therefore, by the contraction principle, the Rademacher complexity of $\mathcal{F}_Q$ can be upper bounded as follows.

$$\begin{aligned} D(\mathcal{F}_Q) &= \left(\mathbb{E}\left[\sup_{u\in\mathbf{B}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i (X_i^T u)^2 \right] \right)^2 \\ &\leqslant 8d \left(\mathbb{E}\left[\sup_{u\in\mathbf{B}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i X_i^T u \right] \right)^2 \ . \end{aligned}$$

By (7.20), it follows that

$$D(\mathcal{F}_Q) \leqslant 32 d^2 \ .$$

By Theorem 40, there exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, at least $9K/10$ blocks B_k satisfy

$$\sup_{u\in\mathbf{B}} |(P_{B_k} - P)[(X^T u)^2]| \leqslant \frac{c^*}{\sqrt{N}} \big(d \vee \sqrt{Kd} \big) \ .$$

□

Lemma 98 implies the following corollary.

Corollary 99. *Consider the design vector X defined in* (7.21)*. There exists an absolute constant c^* such that, if*

$$c^* \overline{\sigma} \big(d \vee \sqrt{dK} \big) \leqslant \sqrt{N} \ ,$$

the minmax MOM estimator $\widehat{f}_K$ satisfies, with probability larger than $1 - 2e^{-K/c^*}$,

$$\|\widehat{f}_K - f^*\|_\Sigma \leqslant \frac{c^* \overline{\sigma}}{\sqrt{N}} \left(\sqrt{d} \vee \sqrt{K} \right) \ .$$

Proof. Denote by c^* the largest of the absolute constants appearing in Lemmas 96 and 98. Define Ω as the event where, simultaneously, there exist $9K/10$ blocks B_k where

$$\sup_{u\in\mathbf{B}} P_{B_k}[\xi X^T u] \leqslant \frac{c^* \overline{\sigma}}{\sqrt{N}} \left(\sqrt{d} \vee \sqrt{K} \right) =: m_K \ ,$$

and $9K/10$ blocks B_k where, for any $u \in \mathbf{B}$,

$$P_{B_k}[(X^T u)^2] \geqslant P[(X^T u)^2] - \frac{c^* \overline{\sigma}}{\sqrt{N}} \big(d \vee \sqrt{dK} \big) \ .$$

By Lemmas 96 and 98, $\mathbb{P}(\Omega) \geqslant 1 - 2e^{-K/c^*}$. On Ω, there exist at least $9K/10$ blocks where, for any $u \in \mathbf{B}$,

$$\|u\|_\Sigma^2 - P_{B_k}[(X^T u)^2] \leqslant \frac{c^*\overline{\sigma}}{\sqrt{N}}\big(d \vee \sqrt{dK}\big) \ .$$

Assume that

$$\frac{c^*\overline{\sigma}}{\sqrt{N}}\big(d \vee \sqrt{dK}\big) \leqslant \frac{1}{2} \ .$$

It follows that, on Ω, there exists at least $8K/10$ blocks where, simultaneously, for any $u \in \mathbf{B}$,

$$P_{B_k}[\xi X^T u] \leqslant m_K, \qquad \|u\|_\Sigma^2 - P_{B_k}[(X^T u)^2] \leqslant \frac{1}{2} \ .$$

On these blocks,

$$\forall r > 0, \qquad P_{B_k}\big[2r\xi[X^T u] + r^2(\|u\|_\Sigma^2 - (X^T u)^2)\big] \leqslant 2rm_K + \frac{r^2}{2} \ .$$

As this relationship holds on more than $K/2$ blocks, it holds for the median, so Ω contains $\cap_{r>0}\Omega_r$, where

$$\forall r > 0, \qquad \Omega_r = \{\mathcal{M}_r \leqslant B(r)\}, \qquad B(r) = 2rm_K + \frac{r^2}{2} \ .$$

With this choice of function B, by (7.4), it follows that (5.8) holds if

$$2r_1 m_K - \frac{r_1^2}{2} \leqslant 0, \quad \text{i.e.} \quad r_1 \geqslant 4m_K \ .$$

Let

$$r_1 = 4m_K \quad \text{so} \quad B(r_1) = 16m_K^2 \ .$$

Then, (5.9) holds if

$$16m_K^2 + 2r_2 m_K - \frac{r_2^2}{2} = 18m_K^2 - \frac{1}{2}\Big(r_2 - 2m_K\Big)^2 \leqslant 0 \ ,$$

that is if

$$r_2 = 8m_K = \frac{8c^*\overline{\sigma}}{\sqrt{N}}\left(\sqrt{d} \vee \sqrt{K}\right) \ .$$

The proof is concluded by Lemma 62. □

7.4.2 An alternative analysis

To conclude this section, let us provide an alternative analysis that can be used on histograms too. All along this section $F = \mathbb{R}^d$ and $\mathbf{S} = \{f \in F : \|f\|_\Sigma = 1\}$. Here, the evaluation function $\mathcal{E}$ is defined as $\mathcal{E}(f) = P[\ell_f - \ell_{f*}] = \|f - f^*\|_\Sigma^2$ in the linear regression problem. Let γ denote a constant such that

$$\forall f \in \mathbf{S}, \qquad P[[X^T f]^4] \leqslant \gamma^2 \ . \tag{7.22}$$

The minmax MOM estimator is studied

$$\widehat{f}_K \in \operatorname{argmin}_{f \in F} \sup_{g \in F} \mathrm{MOM}_K[\ell_f - \ell_g] \ .$$

To express the results, the following complexity is used.

$$\mathcal{C}_Q(F) := \mathbb{E}\left[\sup_{f\in\mathbf{S}} \sum_{i=1}^N \epsilon_i (X_i^T f)^2\right] \ . \tag{7.23}$$

Recall also that

$$\mathbb{E}\left[\sup_{f\in\mathbf{S}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i (\xi X - P[\xi X])^T f\right] = \sqrt{D_N(\mathbf{S})} \ ,$$

where $D_N(\mathbf{S})$ is the Rademacher complexity computed in the proof of Lemma 96. By (7.19), it holds that

$$\mathbb{E}\left[\sup_{f\in\mathbf{S}} \sum_{i=1}^N \epsilon_i (\xi X - P[\xi X])^T f\right] = \overline{\sigma}\sqrt{dN} \ . \tag{7.24}$$

This shows that that $\mathcal{C}_M(F)$ is a measure of complexity that extends the dimension used in the previous sections.

Theorem 100. *Assume* (7.22). *There exists an absolute constant $c > 0$ such that the following holds. If*

$$c\mathcal{C}_Q(F) \leqslant N \quad \textit{and} \quad c\gamma\sqrt{K} \leqslant \sqrt{N} \ ,$$

then, the minmax MOM estimator satisfies

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}_K) \leqslant c\overline{\sigma}^2 \frac{d\vee K}{N}\right)\right) \geqslant 1 - 2e^{-K/c} \ .$$

Proof. By (7.24),

$$\mathbb{E}\left[\sup_{f\in\mathbf{S}} \frac{1}{\sqrt{N}} \sum_{i=1}^N \epsilon_i (\xi X - P[\xi X])^T f\right] \leqslant \overline{\sigma}\sqrt{d} \ .$$

By Theorem 40, there exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where

$$\forall f \in F, \qquad (P_{B_k} - P)[\xi X^T f] \leqslant c^*\overline{\sigma}\sqrt{\frac{d\vee K}{N}}\|f\|_\Sigma \ .$$

By (7.3), for any $f \in F$, $P[\xi X^T(f - f^*)] \leqslant 0$. This implies that, with probability larger than $1 - e^{-K/c^*}$, for any $f \in F$, there exist more than $9K/10$ blocks B_k where

$$P_{B_k}[\xi X^T(f - f^*)] \leqslant c^*\overline{\sigma}\sqrt{\frac{d\vee K}{N}}\|f - f^*\|_\Sigma \ . \tag{7.25}$$

For any $f \in \mathbf{S}$,

$$\mathrm{Var}((X^T f)^2) \leqslant \gamma^2 \ .$$

Moreover,

$$\mathbb{E}\left[\sup_{f\in\mathbf{S}} \sum_{i=1}^N \epsilon_i (X^T f)^2\right] \leqslant \mathcal{C}(Q) \ .$$

By Theorem 40, there exists an absolute constant c^* such that, with probability larger than $1 - e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where, for any $f \in F$,

$$P_{B_k}[(X^T f)^2] \geqslant \|f\|_\Sigma^2(1-\theta_K), \qquad \theta_K = c^*\left(\frac{\mathcal{C}(Q)}{N} \vee \gamma\sqrt{\frac{K}{N}}\right) .$$

Combined with (7.25), this shows that, with probability larger than $1-2e^{-K/c^*}$, there exist at least $9K/10$ blocks B_k where

$$\forall f \in F : P_{B_k}[\xi X^T f] \leqslant \|f\|_\Sigma m_K, \qquad \text{where} \qquad m_K = c^*\overline{\sigma}\sqrt{\frac{d \vee K}{N}}\Bigg)$$

and at least $9K/10$ blocks B_k where

$$P_{B_k}[(X^T f)^2] \geqslant \|f\|_\Sigma^2(1-\theta_K)\big) .$$

Let $m'_K = m_K \vee \mathcal{C}(Q) \geqslant \mathcal{C}(Q) \vee \mathcal{C}(M)$. If c in the theorem is chosen such that $\theta_K \leqslant 1/2$, on this event, there is at least $8K/10$ blocks where, for any $f \in F$

$$P_{B_k}[2\xi X^T f - (X^T f)^2] \leqslant 2\|f\|_\Sigma m_K - \|f\|_\Sigma^2(1-\theta_K) \leqslant 2m_K^2 .$$

It follows that

$$\sup_{f\in F} \mathrm{MOM}_K[\ell_{f^*} - \ell_f] \leqslant 2m_K^2 .$$

The proof terminates with the non-localized bound Lemma 73. □

A second proof of Corollary 99. In the histogram example, for any vector $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\|_\Sigma = \|\mathbf{u}\|/\sqrt{d}$. Moreover,

$$P[(X^T\mathbf{u})^4] = P[(\sum_{i=1}^d u_i\varphi_i(\tilde{X}))^4] = \sum_{i=1}^d u_i^4 P\varphi_i = \frac{1}{d}\|\mathbf{u}\|_4^4 \leqslant \frac{\|\mathbf{u}\|^4}{d} = d\|\mathbf{u}\|_\Sigma^4 .$$

Hence, (7.22) holds with $\gamma = \sqrt{d}$. Moreover, for any $f \in \mathbf{S}$,

$$\begin{aligned}\sum_{i=1}^N \epsilon_i (X_i^T f)^2 &= \sum_{i=1}^N \epsilon_i (\sum_{j=1}^d u_j\varphi_j(\tilde{X}_i))^2 \\ &= \sum_{i=1}^N \epsilon_i \sum_{j=1}^d u_j^2\varphi_j(\tilde{X}_i) \\ &\leqslant d\|\mathbf{u}\|_\Sigma^2 \max_{j\in\{1,\dots,d\}} \left|\sum_{i=1}^N \epsilon_i\varphi_j(\tilde{X}_i)\right| .\end{aligned}$$

Hence,

$$\mathcal{C}_Q(F) \leqslant d\mathbb{E}\left[\max_{j\in\{1,\dots,d\}} \left|\sum_{i=1}^N \epsilon_i\varphi_j(\tilde{X}_i)\right|\right] .$$

By (4.6),

$$\mathcal{C}_Q(F) \leqslant 5d\sqrt{N} .$$

Therefore, the conditions of Theorem 100 reduce to those of Corollary 99. It follows from Theorem 100 that Corollary 99 holds. □

Chapter 8

Density estimation with Hellinger loss

This chapter presents basic properties of ρ-estimators that have been introduced in [3, 4, 6]. The purpose is not to make a complete presentation of this rich theory, the interested reader is invited to read the mentioned references for this. Instead, I try to stress some links between robust learning theory and this extension of Le Cam and Birgé's works on estimation from robust tests, see [8] for an account on this theory and references. In particular, one can see that these estimators are built from the minmax principle presented in Section 5.1 of Chapter 5 and can be analysed with Talagrand's inequality and the homogeneity lemma instead of the peeling argument used in the original proofs of the main result of this chapter. It provides an example of estimation problem that does not fall into Vapnik's theory presented in the introduction where the homogeneity lemma in its general form is useful. Besides this minor modification, all the material presented here is borrowed from [6].

8.1 Setting

This chapter deals with a particular instance of *unsupervised learning* where the dataset $\mathcal{D}_N = (Z_1, \dots, Z_N)$ is a set of i.i.d. random variables taking values in a measurable space $\mathcal{Z}$, with common distribution P^*. Let μ denote a measure on $\mathcal{Z}$. The parameters $f \in F$ are real valued functions defined on $\mathcal{Z}$. These functions are densities with respect to μ and define the measures P_f on $\mathcal{Z}$, P_f being the distribution with density f with respect to μ. To measure distances between probability distributions and evaluate the distributions P_f as estimators of P^*, we use the Hellinger distance h. Let P and Q denote two probability measures and let λ denote a measure dominating both P and Q, the Hellinger distance between P and Q is defined by

$$h(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\int (\sqrt{p} - \sqrt{q})^2 \mathrm{d}\lambda} \ .$$

It is clear that $0 \leqslant h(P,Q) \leqslant 1$ for any probability measures P and Q and that $h(P,Q)$ does not depend on the dominating measure λ. The evaluation function

$\mathcal{E}$ is defined on F as $\mathcal{E}(f) = h(P_f, P^*)$.

The purpose of this chapter is to analyse ρ-estimators of P^* introduced in [4] and defined by $P_{\widehat{f}}$, where

$$\widehat{f} \in \operatorname{argmin}_{f\in F} \sup_{g\in F} T(f,g), \qquad \text{where} \qquad T(f,g) = \sum_{i=1}^{N} \rho\left(\sqrt{\frac{g(Z_i)}{f(Z_i)}}\right) \ . \quad (8.1)$$

Here, the function $\rho = (x-1)/(x+1)$ is non-decreasing $[0,+\infty] \to [-1,1]$, 2-Lipschitz, it satisfies $\rho(1/x) = -\rho(x)$ for any $x \in [0,+\infty)$.

8.2 Preliminary results

This section presents the first results on the tests defining ρ-estimators. The goal is to understand the intuition behind the construction of these estimators. The choice of function ρ is justified by the following remarkable property. The material of this section is borrowed from [6].

Theorem 101. *For any $f \in F$, let P_f denote the probability distribution with density f w.r.t. the measure μ, then*

$$\int \rho\left(\sqrt{\frac{f}{g}}\right) \mathrm{d}R \leqslant 4h^2(R,P_g) - (3/8)h^2(R,P_f) \ ,$$
$$\int \rho^2\left(\sqrt{\frac{f}{g}}\right) \mathrm{d}R \leqslant 3\sqrt{2}[h^2(R,P_g) + h^2(R,P_f)] \ .$$

Remark 102. *The strength of this result is that it is valid for any distributions P_f, P_g and R. It implies in particular that the sign of the expectation $\mathbb{E}[T(f,g)]$, where $T(f,g)$ is defined in (8.1), provides relevant informations regarding which distribution between P_f and P_g is the closest to P^*.*

Proof. The proof proceeds in two steps.

Lemma 103. *Theorem 101 holds for any R absolutely continuous w.r.t. μ.*

Proof. The proof is quite technical and not very intuitive. It uses repeatedly the following relation: for any distributions P, Q and any measure λ dominating P and Q,

$$h^2(P,Q) = \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 \mathrm{d}\lambda = 1 - \int \sqrt{pq}\mathrm{d}\lambda \ .$$

Let r denote the density $r = \delta^{-2}(\sqrt{f} + \sqrt{g})^2$, where

$$\delta^2 = \int (\sqrt{f} + \sqrt{g})^2 \mathrm{d}\mu = 4\left(1 - \frac{h^2(P_f,P_g)}{2}\right) \ .$$

As $h^2(P_f,P_g) \in [0,1]$, this implies that $\sqrt{2} \leqslant \delta \leqslant 2$. Moreover, by convexity of the map $\vartheta : u \mapsto 1/\sqrt{1-u}$ on $(0,1)$,

$$\frac{2}{\delta} = \vartheta\left(\frac{h^2(P_f,P_g)}{2}\right) \geqslant \vartheta(0) + \vartheta'(0)\frac{h^2(P_f,P_g)}{2} = 1 + \frac{h^2(P_f,P_g)}{4} \ . \quad (8.2)$$

Denote by s the density of R with respect to μ. It follows that

$$h^2(R,P_r) = 1 - \int \sqrt{sr}\mathrm{d}\mu = 1 - \frac{1}{\delta}\Big(\int \sqrt{sf}\mathrm{d}\mu + \int \sqrt{sg}\mathrm{d}\mu\Big) \tag{8.3}$$

$$= 1 - \frac{2}{\delta} + \frac{h^2(R,P_f)+h^2(R,P_g)}{\delta}$$

$$\leqslant \frac{h^2(R,P_f)+h^2(R,P_g)}{\delta} - \frac{h^2(P_f,P_g)}{4} \ . \tag{8.4}$$

Elementary calculus shows that

$$\int \rho^2\Big(\sqrt{\frac{f}{g}}\Big)s\mathrm{d}\mu = \int_{r>0}\left(\frac{\sqrt{f}-\sqrt{g}}{\sqrt{f}+\sqrt{g}}\right)^2(\sqrt{s}-\sqrt{r}+\sqrt{r})^2\mathrm{d}\mu \ .$$

Using the inequality $(a+b)^2 \leqslant (1+\alpha)a^2+(1+\alpha^{-1})b^2$, valid for any real numbers a and b and any $\alpha > 0$ to $a = \sqrt{s}-\sqrt{r}$ and $b = \sqrt{r}$ shows that, for any $\alpha > 0$,

$$\int \rho^2\Big(\sqrt{\frac{f}{g}}\Big)s\mathrm{d}\mu \leqslant (1+\alpha)\int_{r>0}\left(\frac{\sqrt{f}-\sqrt{g}}{\sqrt{f}+\sqrt{g}}\right)^2(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu$$

$$+ (1+\alpha^{-1})\int_{r>0}\left(\frac{\sqrt{f}-\sqrt{g}}{\sqrt{f}+\sqrt{g}}\right)^2\left(\frac{\sqrt{f}+\sqrt{g}}{\delta}\right)^2\mathrm{d}\mu \ .$$

In this expression, as $\big((\sqrt{f}-\sqrt{g})(\sqrt{f}+\sqrt{g})\big)^2 \leqslant 1$, the first item in the right hand side is bounded from above by $2(1+\alpha)h^2(R,P_r)$. The second item in the right hand side is equal to $(1+\alpha^{-1})(2/\delta^2)h^2(P_f,P_g)$. Combining these upper bounds yields

$$\int \rho^2\Big(\sqrt{\frac{f}{g}}\Big)s\mathrm{d}\mu \leqslant 2(1+\alpha)h^2(R,P_r) + \frac{2(1+\alpha^{-1})}{\delta^2}h^2(P_f,P_g) \ .$$

Then by (8.4),

$$\int \rho^2\Big(\sqrt{\frac{f}{g}}\Big)s\mathrm{d}\mu$$

$$\leqslant \frac{2(1+\alpha)}{\delta}\big(h^2(R,P_f)+h^2(R,P_g)\big) - \frac{\delta^2(1+\alpha)-4(1+\alpha^{-1})}{2\delta^2}h^2(P_f,P_g) \ .$$

If $(1+\alpha)\delta^2 = 4(1+\alpha^{-1})$, it implies

$$\int \rho^2\Big(\sqrt{\frac{f}{g}}\Big)s\mathrm{d}\mu = \frac{2(1+\alpha)}{\delta}\big(h^2(R,P_f)+h^2(R,P_g)\big) \ .$$

Solving the equation $(1+\alpha)\delta^2 = 4(1+\alpha^{-1})$ in α gives $\alpha = 4/\delta^2$, thus $2(1+\alpha)/\delta = 2/\delta + 4/\delta^3 \leqslant 3\sqrt{2}$ since $\delta \geqslant \sqrt{2}$. This proves the second item of Theorem 101 when R is absolutely continuous with respect to μ.

Moving to the first item, define, for any $f \in F$,

$$\rho_r(R,P_f) = \frac{1}{2}\Big[\int \sqrt{fr}\mathrm{d}\mu + \int \sqrt{\frac{f}{r}}s\mathrm{d}\mu\Big] \ .$$

The increments of $\rho_r(R,\cdot)$ are intimately related to the expectation of T: for any f and g in F,

$$\begin{aligned}\rho_r(R,P_f)-\rho_r(R,P_g)&\\ &= \frac{1}{2}\left[\frac{1}{\delta}\int(\sqrt{f}-\sqrt{g})(\sqrt{f}+\sqrt{g})\mathrm{d}\mu+\delta\int\frac{\sqrt{f}-\sqrt{g}}{\sqrt{f}+\sqrt{g}}s\mathrm{d}\mu\right]\\ &= \frac{\delta}{2}\int\rho\left(\sqrt{\frac{f}{g}}\right)s\mathrm{d}\mu=\frac{\delta}{2N}\mathbb{E}[T(f,g)]\ . \end{aligned}\tag{8.5}$$

Moreover,

$$\begin{aligned}\int\sqrt{\frac{f}{r}}s\mathrm{d}\mu &= \int\sqrt{\frac{f}{r}}(\sqrt{s}-\sqrt{r}+\sqrt{r})^2\mathrm{d}\mu\\ &= \int\sqrt{\frac{f}{r}}(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu+\int\sqrt{fr}\mathrm{d}\mu+2\int\sqrt{f}(\sqrt{s}-\sqrt{r})\mathrm{d}\mu\\ &= \int\sqrt{\frac{f}{r}}(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu-\int\sqrt{fr}\mathrm{d}\mu+2\int\sqrt{fs}\mathrm{d}\mu\ .\end{aligned}$$

As $f=0$ on the event $r=0$, it follows that

$$\begin{aligned}\rho_r(R,P_f) &= \int\sqrt{fs}\mathrm{d}\mu+\frac{1}{2}\int_{r>0}\sqrt{\frac{f}{r}}(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu\\ &= \int\sqrt{fs}\mathrm{d}\mu+\frac{\delta}{2}\int_{r>0}\frac{\sqrt{f}}{\sqrt{f}+\sqrt{g}}(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu\end{aligned}$$

Thus (8.5) implies that

$$\begin{aligned}\frac{\delta}{2}\int\rho\left(\sqrt{\frac{f}{g}}\right)s\mathrm{d}\mu &= \int(\sqrt{f}-\sqrt{g})\sqrt{s}\mathrm{d}\mu+\frac{\delta}{2}\int_{r>0}\frac{\sqrt{f}-\sqrt{g}}{\sqrt{f}+\sqrt{g}}(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu\\ &= \int(\sqrt{f}-\sqrt{g})\sqrt{s}\mathrm{d}\mu+\frac{\delta}{2}\int_{r>0}\rho\left(\sqrt{\frac{f}{g}}\right)(\sqrt{s}-\sqrt{r})^2\mathrm{d}\mu\ .\end{aligned}$$

As ρ takes values in $[-1,1]$,

$$\frac{\delta}{2}\int\rho\left(\sqrt{\frac{f}{g}}\right)s\mathrm{d}\mu\leqslant\int(\sqrt{f}-\sqrt{g})\sqrt{s}\mathrm{d}\mu+\delta h^2(R,P_r)\ .$$

By (8.3),

$$\frac{\delta}{2}\int\rho\left(\sqrt{\frac{f}{g}}\right)s\mathrm{d}\mu\leqslant\delta-2\int\sqrt{gs}\mathrm{d}\mu\ .$$

By (8.2),

$$\begin{aligned}\int\rho\left(\sqrt{\frac{f}{g}}\right)s\mathrm{d}\mu &\leqslant 2\left[1-\int\sqrt{gs}\mathrm{d}\mu\left(1+\frac{h^2(P_f,P_g)}{4}\right)\right]\\ &\leqslant 2\left[h^2(R,P_g)\left(1+\frac{h^2(P_f,P_g)}{4}\right)-\frac{h^2(P_f,P_g)}{4}\right]\\ &\leqslant\frac{1}{2}[5h^2(R,P_g)-h^2(P_f,P_g)]\ .\end{aligned}$$

By the triangular inequality, $h(P_f, P_g) \geqslant |h(R, P_g) - h(R, P_f)|$, hence,

$$\begin{aligned} h^2(P_f, P_g) &\geqslant h^2(R, P_g) + h^2(R, P_f) - 2h(R, P_g)h(R, P_f) \\ &\geqslant \frac{3}{4}h^2(R, P_f) - 3h^2(R, P_g) \ . \end{aligned}$$

Therefore,

$$\int \rho\left(\sqrt{\frac{f}{g}}\right) s \mathrm{d}\mu \leqslant 4h^2(R, P_g) - \frac{3}{8}h^2(R, P_f) \ .$$

The first item of Theorem 101 is established in the case where R is absolutely continuous with respect to μ. This concludes the proof of Lemma 103. □

The second result shows that it is sufficient to show Theorem 101 when R is absolutely continuous with respect to μ to prove it in general.

Lemma 104. *If Theorem 101 holds for any R absolutely continuous with respect to μ, it holds for any R.*

Proof. Write $R = \delta^2 R' + (1-\delta^2)R''$, with R' absolutely continuous with respect to μ, R'' orthogonal to μ and $\delta \in (0,1)$. Let $\overline{\mu} = R + P_f$ which dominates both R and P_f. As R'' is orthogonal to μ, it holds that $(\mathrm{d}R''/\mathrm{d}\overline{\mu})(\mathrm{d}P_f/\mathrm{d}\overline{\mu}) = 0$. Therefore, the following fundamental relationship between the Hellinger distances $h^2(R, P_f)$ and $h^2(R', P_f)$ holds,

$$1 - h^2(R, P_f) = \int \sqrt{\left(\delta^2 \frac{\mathrm{d}R'}{\mathrm{d}\overline{\mu}} + (1-\delta^2)\frac{\mathrm{d}R''}{\mathrm{d}\overline{\mu}}\right)\frac{\mathrm{d}P_f}{\mathrm{d}\overline{\mu}}}\mathrm{d}\overline{\mu} = \delta(1 - h^2(R', P_f)) \ .$$

As this holds for any $f \in F$, in particular,

$$\begin{aligned} h^2(R, P_f) &= 1 - \delta + \delta h^2(R', P_f) \geqslant 1 - \delta \ . \\ h^2(R, P_g) &= 1 - \delta + \delta h^2(R', P_g) \geqslant 1 - \delta \ . \end{aligned} \tag{8.6}$$

By hypothesis, the second item of Theorem 101 applies to R' that is absolutely continuous with respect to μ, so

$$\begin{aligned} \int \rho^2\left(\sqrt{\frac{f}{g}}\right)\mathrm{d}R &\leqslant \delta^2 \int \rho^2\left(\sqrt{\frac{f}{g}}\right)\mathrm{d}R' + 1 - \delta^2 \\ &\leqslant 3\sqrt{2}\delta^2[h^2(R', P_g) + h^2(R', P_f)] + 1 - \delta^2 \ . \end{aligned}$$

Applying the fundamental relations (8.6) yields

$$\begin{aligned} \int \rho^2\left(\sqrt{\frac{f}{g}}\right)\mathrm{d}R &\leqslant 3\sqrt{2}\delta(h^2(R, P_g) + h^2(R, P_f) - 2(1-\delta)) + (1-\delta^2) \\ &= 3\sqrt{2}(h^2(R, P_g) + h^2(R, P_f)) + \mathrm{Rem}(\delta) \ , \end{aligned}$$

where the remainder term satisfies, according to the fundamental relations (8.6),

$$\begin{aligned} \mathrm{Rem}(\delta) &= (1-\delta^2) - 3\sqrt{2}(2\delta(1-\delta) + (1-\delta)(h^2(R, P_g) + h^2(R, P_f))) \\ &\leqslant (1-\delta^2) - 3\sqrt{2}(2\delta(1-\delta) + 2(1-\delta)^2) \\ &= (1-\delta)(1 + \delta - 6\sqrt{2}(2-\delta)) \leqslant (1-\delta)(2 - 6\sqrt{2}) \leqslant 0 \ . \end{aligned}$$

This proves the second item of Theorem 101 for R.

By hypothesis, the first item of Theorem 101 applies to R' that is absolutely continuous with respect to μ, so

$$\begin{aligned}\int \rho\left(\sqrt{\frac{f}{g}}\right)\mathrm{d}R &\leqslant \delta^2 \int \rho\left(\sqrt{\frac{f}{g}}\right)\mathrm{d}R' + (1-\delta^2) \\ &\leqslant \delta^2(4h^2(R', P_g) - (3/8)h^2(R', P_f)) + (1-\delta^2) \\ &= \delta(4h^2(R, P_g) - (3/8)h^2(R, P_f) - (29/8)(1-\delta)) + (1-\delta^2) \\ &= 4h^2(R, P_g) - (3/8)h^2(R, P_f) + \mathrm{Rem}(\delta) \ ,\end{aligned}$$

where the remainder term

$$\begin{aligned}\mathrm{Rem}(\delta) &= (1-\delta^2) - (29/8)\delta(1-\delta) - (1-\delta)(4h^2(R, P_g) - (3/8)h^2(R, P_f)) \\ &\leqslant (1-\delta)(11/8 + \delta - 29/8\delta - 4(1-\delta)) \\ &= (1-\delta)(-21/8 + 11/8\delta) \leqslant 0 \ .\end{aligned}$$

This concludes the proof of the first item of Theorem 101. Therefore, Lemma 104 is proved. □

Theorem 101 is a direct consequence of Lemmas 103 and 104. □

8.3 Main result

The remaining of the chapter is devoted to the proof of the following theorem.

Theorem 105. *Let $f^* \in \mathrm{argmin}_{f\in F} h(P^*, P_f)$ and, for any $f \in F$, let $U_{i,f} = \rho(\sqrt{f(X_i)/f^*(X_i)})$. Define the complexity of the model F as a fixed point of the following local Rademacher complexity of F:*

$$D(F) = 1 \vee N\left(\sup\left\{r > 0 : \mathbb{E}\left[\sup_{f\in F:\mathcal{E}(f)\leqslant r} \frac{1}{N}\sum_{i=1}^{N} \epsilon_i U_{i,f}\right] > \frac{r^2}{80}\right\}\right)^2 \ .$$

There exists an absolute constant C such that any ρ-estimator $\widehat{f}$ defined in (8.1) *satisfies, with probability larger than $1 - 2e^{-t}$,*

$$h^2(P^*, P_{\widehat{f}}) \leqslant C\left(\inf_{f\in F} h^2(P^*, P_f) + \frac{D(F)+t}{N}\right) \ .$$

Remark 106. *Again, the remarkable feature here is that Theorem 105 holds without assumptions on P^* or the set F of densities.*

Proof. Recall that the evaluation function is defined in this chapter, for any $f \in F$, by $\mathcal{E}(f) = h(P^*, P_f)$ and that f^* is defined as a density in F such that

$$\forall f \in F, \qquad \mathcal{E}(f^*) \leqslant \mathcal{E}(f) \ .$$

Hereafter, define also

$$\forall f, g \in F^2, \qquad d(f, g) = N\mathbb{E}_{P^*}\left[\rho\left(\sqrt{\frac{g}{f}}\right)\right] \ .$$

Theorem 101 shows in particular that, for any $f \in F$,

$$(3/8)\mathcal{E}^2(f) - 4\mathcal{E}^2(f^*) \leqslant \frac{d(f,f^*)}{N} \leqslant 4\mathcal{E}^2(f) - 3/8\mathcal{E}^2(f^*) \ . \tag{8.7}$$

Let $r_0 = \mathcal{E}(f^*)$. By Lemma 67, the test T fulfils Condition **(HP)** of the homogeneity lemma (Lemma 62). To bound the Hellinger distance between the associated minmax estimator and the unknown density P^* of the observations, it remains to compute the function B in the homogeneity Lemma.

Fix $r > r_0$. Recall that $U_{i,f} = \rho(\sqrt{f(X_i)/f^*(X_i)})$ are independent random variables, bounded by 1 and that

$$T(f^*, f) = \sum_{i=1}^{N} U_{i,f} \ .$$

Moreover, Theorem 101 shows that, for any $f \in F$ such that $\mathcal{E}(f) \leqslant r$,

$$\mathrm{Var}(U_{i,f}) \leqslant 6\sqrt{2}Nr^2 \ .$$

Therefore, it follows from Talagrand's concentration inequality (Theorem 34) that, for any $t > 0$, the random variable $Z_r = \sup_{f \in F : \mathcal{E}(f) \leqslant r} \sum_{i=1}^{N} (U_{i,f} - \mathbb{E}[U_{i,f}]$ satisfies

$$\mathbb{P}\left(Z_r \leqslant 2\mathbb{E}\big[Z_r\big] + \frac{N}{20}r^2 + (2 + 20\sqrt{6})t \right) \geqslant 1 - e^{-t} \ .$$

By the symmetrization trick, $\mathbb{E}\big[Z_r\big] \leqslant 2\mathbb{E}[Z_{\epsilon,r}]$, where

$$Z_{\epsilon,r} = \sup_{f \in F : \mathcal{E}(f) \leqslant r} \sum_{i=1}^{N} \epsilon_i U_{i,f} \ .$$

Hence, with probability at least $1 - e^{-t}$,

$$Z_r \leqslant 4\mathbb{E}\big[Z_{\epsilon,r}\big] + \frac{N}{20}r^2 + (2 + 20\sqrt{6})t \leqslant 4\mathbb{E}\big[Z_{\epsilon,r}\big] + \frac{N}{20}r^2 + 51t \ .$$

By definition of $D(F)$, for any $r > \sqrt{D(F)/N}$,

$$\mathbb{E}[Z_{\epsilon,r}] \leqslant \frac{Nr^2}{80} \ .$$

Hence, for any $r > r_0 \vee \sqrt{D(F)/N}$, it follows that, with probability at least $1 - e^{-t}$,

$$Z_r \leqslant \frac{N}{10}r^2 + 51t \ .$$

As a consequence, for any $t > 0$ and $r > \sqrt{D(F)/N} \vee r_0$, one can choose

$$B(r) = \frac{Nr^2}{10} + 51t$$

in the homogeneity lemma and get that the event Ω_r in Lemma 62 holds with probability at least $1 - e^{-t}$.

With this value of $B(r)$, from (8.7) that

$$B(r) - \inf_{f \in F : \mathcal{E}(f) = r} d(f, f^*) \leqslant \frac{Nr^2}{10} + 51t - \frac{3N}{8} r^2 + 4N\mathcal{E}^2(f^*)$$
$$\leqslant 51t + 4N\mathcal{E}^2(f^*) - \frac{N}{4} r^2 \ .$$

From this upper bound, one can choose $r_1 = \sqrt{204t/N + 16\mathcal{E}^2(f^*)} \vee \sqrt{D(F)/N}$ in (5.8). Then,

$$B(r_1) = (20.4t + 1.6N\mathcal{E}^2(f^*) + 51t) \vee \left(\frac{D(F)}{10} + 51Nt \right)$$
$$\leqslant (2N\mathcal{E}^2(f^*) + 72t) \vee \left(\frac{D(F)}{10} + 51t \right) \ .$$

By (8.7), $\sup_{f \in F} d(f^*, f) \leqslant 4N\mathcal{E}(f^*)$. Hence, one can choose the following upper bound $\mathcal{B}$ in Lemma 62:

$$\mathcal{B} = 6N\mathcal{E}^2(f^*) + \frac{D(F)}{10} + 72t \ .$$

Hence, (5.9) holds for any r such that

$$51t + 4N\mathcal{E}^2(f^*) - \frac{N}{4} r^2 \leqslant -\left(6N\mathcal{E}^2(f^*) + \frac{D(F)}{10} + 72t \right) \ ,$$

i.e. for any

$$r^2 \geqslant 40\mathcal{E}^2(f^*) + \frac{2D(F)}{5N} + 492t \ .$$

□

Chapter 9

Estimators computable in polynomial time

In the previous chapters, we studied minmax MOM estimators in various contexts and showed that they achieved interesting theoretical performance under weak assumptions on the data.

For example, for multivariate mean estimation, they are proved to satisfy a sub-Gaussian deviation inequality

$$\mathbb{P}\left(\|\widehat{\mu}_K - \mu_P\| > C\frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}} K}}{\sqrt{N}} \right) \leqslant e^{-K/C} \ , \tag{9.1}$$

where C is some absolute constant, assuming only that $P[\|X\|^2] < \infty$.

The first estimator that was shown to achieve this bound was proposed in [42]. The procedure there was closely related to minmax MOM, several other procedures achieving similar bounds have been proposed since then. Some of them are presented in Chapter 4 for example, see also [38] for a proof that a clever extension of the classical trimmed mean estimator on $\mathbb{R}$ has sub-Gaussian deviations and [40] for a review on the subject. The problem with the minmax MOM construction or the one based on Le Cam's aggregation of tests in [42] is that these estimators cannot be computed in polynomial time.

In this chapter, we consider the problem of building estimators achieving sub-Gaussian deviations (9.1) that can be computed in polynomial time. This problem was solved first in [27] using an estimator solving a semidefinite program (SDP). Recall that these take the form of finding the minimizer in $\mathbf{X} \in \mathbb{R}^{d\times d}$ of the functional $\langle \mathbf{X}, \mathbf{C}\rangle = \mathrm{Tr}(\mathbf{X}\mathbf{C}^T)$, subject to the constraints $\langle \mathbf{A}_1, \mathbf{X}\rangle \geqslant 0, \dots, \langle \mathbf{A}_k, \mathbf{X}\rangle \geqslant 0$ and $\mathbf{X}$ ranges over the symmetric positive semi-definite matrices $\mathbf{X} \succeq 0$. Under mild conditions on $\mathbf{C}$ and $\mathbf{A}_1, \dots, \mathbf{A}_k$, semidefinite programs (SDP) can be solved in polynomial time. To find a SDP whose solution achieves (9.1), [27] uses the sum-of-squares (SoS) method. Let $p, q_1, \dots, q_m$ denote multivariate polynomials in $\mathbb{R}[x_1, \dots, x_n]$, the SoS method produces a SDP relaxation of the problem of finding a minimizer of $P(\mathbf{x})$ under the constraints $q_1(\mathbf{x}) \geqslant 0, \dots, q_m(\mathbf{x}) \geqslant 0$. This relaxation depends on an even integer $r \geqslant \max\{\deg(p), \deg(q_i), i = 1, \dots, m\}$. The relaxation is solvable in $O((Nm)^{O(r)})$ operations and, of course, the quality of the approximation improves with r. The solution in [27] uses $r = 8$ and produces an algorithm that

runs in $O(N^{24})$ operations. While this is actually polynomial time algorithm, it can still not be used in practice.

Using ideas related to [27], [17] proposed an alternative SDP relaxation that improved considerably the running time. The method goes as follows. They first considered the problem $P0$ of finding the vectors $\mathbf{b} \in \{0,1\}^K$ and $\mathbf{v} \in \mathbf{S}$ such that $\sum_{k=1}^K b_k$ is as large as possible under the constraint that, for all $k = 1, \dots K$, $b_k \mathbf{v}^T (P_{B_k} X - \mathbf{x}) \geqslant b_k^2 r$. If this problem could be solved, it could be used to estimate first the distance between $\mathbf{x}$ and μ_P by $d_{\mathbf{x}}$ the largest r such that $\sum b_k \geqslant (1-\alpha)K$ and then an estimation $g_{\mathbf{x}}$ of the direction $\mathbf{x} - \mu_P$ by the optimal vector $\mathbf{v}$ given for $r = d_{\mathbf{x}}$. Then, using these estimations, one can build a descent algorithm that moves from $\mathbf{x}$ to $T(\mathbf{x}) = \mathbf{x} - \gamma d_{\mathbf{x}} g_{\mathbf{x}}$ and that stops when $d_{T(\mathbf{x})} > d_{\mathbf{x}}$.

They first proved that this descent algorithm produces in $O(\log \|\widehat{\mu}^{(0)}\|/\epsilon)$ an estimator that is, with probability larger than $1 - e^{-K/C}$ at distance from μ_P bounded from above by

$$\epsilon \vee C \frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}} K}}{\sqrt{N}}$$

The key is thus to find an approximate solution of the basic problem $P0$. For this, they used a SDP relaxation of $P0$. They looked for a positive semidefinite matrix $\mathbf{X} \succeq 0$ of size $K + d + 1$ with entries $x_{i,j}$ such that $\sum_{k=1}^K x_{1,b_k}$ is as large as possible under the constraints that $x_{1,b_k} = x_{b_k,b_k}$, $X_{1,1} = 1$, $\sum_{j=1}^d x_{v_j,v_j} = 1$ and, for any $k = 1, \dots, K$, $\mathbf{X}_{b_k,\mathbf{v}}^T (P_{B_k} X - \mathbf{x}) \geqslant x_{b_k,b_k} r$. Here, the vectors

$$\mathbf{X}_{b_k,\mathbf{v}} = \begin{bmatrix} x_{b_k,v_1} \\ \vdots \\ x_{b_k,v_d} \end{bmatrix} .$$

This SDP can be solved using an interior point method that runs in $O(k^{3.5})$ operations. They also proved that a solution of this problem can be used to build a descent algorithm that, overall, runs in $\tilde{O}(kd + k^{3.5})$ operations. Here and in the following $m = \tilde{O}(f(N,d,k))$ mean that there exists absolute constants c_1, c_2 such that

$$m \leqslant c_1 f(N,d,k) (\log(dN))^{c_2} .$$

The method detailed in this chapter comes from [19]. It uses a convex relaxation of the problem that is closely related to a construction proposed in [16], to build estimators that are robust to a large number of outliers. The key technical tool to solve this problem comes from [49], it is reproduced here without a proof. The main idea in [19] is an extension of Theorem 40 that is provided in Lemma 109. The material of Lemma 108 is a simplification of the Geometric-MOM algorithm of [47] also due to J. Depersin and G. Lecué that provides a particularly simple and elegant construction that yields performance similar to [47], which are sadly slightly sub-optimal. A competitive method, with similar complexity but using a spectral algorithm instead of a SDP relaxation was also proposed in [37], it will be included in a future version of these notes. The main result of the chapter is the following.

Theorem 107. *There exists a numerical constant C and an algorithm that runs in $\tilde{O}(uK + Kd)$ operations and outputs an estimator $\widehat{\mu}$ of μ_P such that*

$$\mathbb{P}\left(\|\widehat{\mu} - \mu_P\| \leqslant C\left(\sqrt{\frac{Tr(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} K}{N}} \right) \right) \geqslant 1 - e^{-u \wedge (K/C)} \ .$$

All along the chapter, we consider the problem of estimating the multivariate expectation $\mu_P \in \mathbb{R}^d$ of a distribution P from a data-set $X_1, \dots, X_N$ of independent random vectors with common expectations μ_P and common covariance matrix $\Sigma = P[(X - \mu_P)(X - \mu_P)^T]$. Hereafter, K denotes an integer, smaller than N. All results can be extended to allow for a proportion γK of outliers, for some $\gamma \in (0, 1/3)$. These outliers may be adversarial, they may not be independent nor independent from the inliers, without affecting the results.

9.1 Initialization of the algorithm

Let $\mathcal{M} = \{P_{B_k} X, k \in \{1, \dots, K\}\}$ denote the set of means. For any $k \in \{1, \dots, K\}$, let

$$\mathcal{C}_k = \text{median}\{\|P_{B_k} X - m\|, m \in \mathcal{M}\}, \qquad \widehat{k} \in \text{argmin}_{k \in \{1,\dots,K\}} \mathcal{C}_k \ .$$

We initialize the algorithm with

$$\widehat{\mu}^{(0)} = P_{B_{\widehat{k}}} X \ .$$

Lemma 108. *The estimator $\widehat{\mu}^{(0)}$ is computed in $O((N + K^2)d)$ operations and satisfies*

$$\mathbb{P}\left(\|\widehat{\mu}^{(0)} - \mu_P\| \leqslant 12\sqrt{\frac{Tr(\Sigma)K}{N}} \right) \geqslant 1 - e^{-K/128} \ .$$

Proof. To compute $\widehat{\mu}^{(0)}$, we need at most Nd operations to compute each $P_{B_k} X$, $K^2 d$ operations to compare all differences $\|P_{B_k} X - P_{B_j} X\|$ and $O(K \log K)$ operations to rank the $\mathcal{C}_k$.

We have

$$P[\|P_{B_k} X - \mu_P\|^2] = \frac{1}{|B_k|^2} \sum_{(i,j) \in B_k} P[(X_i - \mu_P)^T (X_j - \mu_P)] = \frac{\text{Tr}(\Sigma) K}{N} \ .$$

Define $r_K = \text{Tr}(\Sigma)K/N$. By Markov's inequality, it follows that, for any $k \in \{1, \dots, K\}$,

$$\forall x > 0, \qquad \mathbb{P}\big(\|P_{B_k} X - \mu_P\| > x r_K \big) \leqslant \frac{1}{x^2} \ .$$

By Hoeffding's inequality,

$$\forall \alpha > 0, \qquad \mathbb{P}\left(\frac{1}{K} \sum_{k=1}^{K} \left(\mathbf{1}_{\{\|P_{B_k} X - \mu_P\| > x r_K\}} - \frac{1}{x^2} \right) > \alpha \right) \leqslant e^{-2K\alpha^2} \ .$$

In words, for any $x > 0$ and $\alpha > 0$, the probability that there exist at least $(1 - \alpha - 1/x^2)K$ blocks B_k where $\|P_{B_k} X - \mu_P\| \leqslant x r_K$ is larger than $1 - e^{-2K\alpha^2}$.

Choosing $\alpha = 1/16$ and $x = 4$ shows that, with probability at least $1 - e^{-K/128}$, $|\mathcal{K}| \geqslant 7K/8$, where $\mathcal{K}$ is the set of indices k of the blocks B_k such that

$$\|P_{B_k} X - \mu_P\| \leqslant 4r_K \ .$$

For any k in $\mathcal{K}$, by the triangular inequality, for any $j \in \mathcal{K}$,

$$\|P_{B_k} X - P_{B_j} X\| \leqslant 8r_K \ .$$

Therefore, since $K \geqslant 3$, $7K/8 - 1 \geqslant K/2$ and thus, on the event $|\mathcal{K}| \geqslant 7K/8$,

$$\forall k \in \mathcal{K}, \qquad \mathcal{C}_k \leqslant 8r_K \ .$$

In particular thus, $\mathcal{C}_{\widehat{k}} \leqslant 8r_K$. Therefore, there is more than $K/2$ blocks where $\|P_{B_k} X - P_{B_{\widehat{k}}} X\| \leqslant 8r_K$, and $7K/8$ blocks where $\|P_{B_k} X - \mu_P\| \leqslant 4r_K$. As $K \geqslant 3$, it follows that there is at least one blocks B_k such that both inequalities hold. Therefore, on the event $|\mathcal{K}| \geqslant 7K/8$,

$$\|\widehat{\mu}^{(0)} - \mu_P\| \leqslant \|P_{B_k} X - P_{B_{\widehat{k}}} X\| + \|P_{B_k} X - \mu_P\| \leqslant 12r_K \ .$$

□

9.2 Technical tools

Before going to the iteration step of the algorithm, we need a series of results that allow to understand the algorithm.

Let $\mathcal{S}_1$ denote the set of matrices $\mathbf{M} \in \mathbb{R}^{d \times d}$ which are symmetric positive semi-definite and satisfy $\mathrm{Tr}(\mathbf{M}) = 1$. For any $\mathbf{M} \in \mathcal{S}_1$, denote by $\mathbf{M}^{1/2}$ a symmetric, positive semi-definite square-root of $\mathbf{M}$. The following result is the main new insight from [19] that allows to apply the machinery in [16]. It is a non trivial consequence of the deviation theorem on suprema of median-of-means processes, in its general version (see Theorem 40).

Lemma 109. *For any $\alpha \in (0,1)$, there exists a constant C_α such that, for all $K \geqslant 1/\alpha$, with probability larger than $1 - e^{-K/C_\alpha}$, for any $\boldsymbol{M} \in \mathcal{S}_1$, there exist more than $(1-\alpha)K$ blocs B_k satisfying*

$$\|\boldsymbol{M}^{1/2}(P_{B_k} X - \mu_P)\| \leqslant C_\alpha \frac{\sqrt{Tr(\Sigma)} + \sqrt{\|\Sigma\|_{op} K}}{\sqrt{N}} \ .$$

Remark 110. *Notice that $\mathcal{S}_1$ contains all matrices of the form $\boldsymbol{M} = \boldsymbol{v}\boldsymbol{v}^T$, with $\boldsymbol{v} \in \boldsymbol{S}$, hence Lemma 109 implies that, with probability larger than $1 - e^{-K/C_\alpha}$, for any $\boldsymbol{v} \in \boldsymbol{S}$, there exist more than $(1-\alpha)K$ blocs B_k satisfying*

$$[\boldsymbol{v}^T(P_{B_k} X - \mu_P)]^2 = \|\boldsymbol{M}^{1/2}(P_{B_k} X - \mu_P)\|^2 \leqslant C_\alpha \frac{Tr(\Sigma) + \|\Sigma\|_{op} K}{N} \ .$$

It is therefore an extension of Corollary 41.

Proof. Let

$$r = c_\alpha \frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}} K}}{\sqrt{N}} \ .$$

Let $\beta \in (0, \overline{\Phi}(1))$ and let Ω denote the event where, for any $\mathbf{v} \in \mathbf{S}$, $|\mathcal{K}_{\mathbf{v}}| \geqslant (1-\beta\alpha)K$, where $\mathcal{K}_{\mathbf{v}}$ denotes the set of indices k such that

$$|\mathbf{v}^T(P_{B_k}X - \mu_P)| \leqslant r \ .$$

Assume that c_α is chosen such that $\mathbb{P}(\Omega) \geqslant 1 - e^{-K/c_\alpha}$. This is possible thanks to Corollary 41.

Fix $\mathbf{M} \in \mathcal{S}_1$, $a > 0$ and let

$$\mathcal{A}_{\mathbf{M}} = \{k \in \{1, \dots, K\} : \|\mathbf{M}^{1/2}(P_{B_k}X - \mu_P)\| > ar\} \ .$$

Suppose that $|\mathcal{A}_{\mathbf{M}}| \geqslant \alpha K$. Let $b \in (1, a)$, let G denote a Gaussian vector with covariance matrix $\mathbf{M}$, independent from $X_1, \dots, X_N$ and let

$$Z = \sum_{k \in \{1,\dots,K\}} \mathbf{1}_{\{|G^T(P_{B_k} - \mu_P)| > br\}} \ .$$

For any $k \in \{1, \dots, K\}$, conditionally on $\mathcal{D}_N$, $G^T(P_{B_k} - \mu_P)$ is a Gaussian random variable, centered, with variance $\sigma_k^2 = \|M^{1/2}(P_{B_k}X - \mu_P)\|^2$. It follows that, for any $k \in \mathcal{A}_M$,

$$\mathbb{P}(|G^T(P_{B_k} - \mu_P)| > br | \mathcal{D}_N) \geqslant \mathbb{P}(|N| > b/a) \geqslant \overline{\Phi}(1) \ ,$$

where N denote a standard Gaussian random variable. Therefore,

$$\mathbb{E}[Z|\mathcal{D}_N] \geqslant \overline{\Phi}(1)|\mathcal{A}_M| \geqslant \overline{\Phi}(1)\alpha K \ .$$

The Paley-Zygmund inequality grants that, for any $\theta \in [0, 1]$, any non-negative random variable Y with finite variance satisfies (**Exercise:** Prove it!)

$$\mathbb{P}(Y > \theta\mathbb{E}[Y]) \geqslant (1-\theta)^2 \frac{\mathbb{E}[Y]^2}{\mathbb{E}[Y^2]} \ .$$

As $0 \leqslant Z \leqslant K$ almost surely, $\mathbb{E}[Z^2|\mathcal{D}_N] \leqslant K^2$, so, for $\theta = \beta/\overline{\Phi}(1)$,

$$\mathbb{P}(Z > \beta\alpha K | \mathcal{D}_N) \geqslant (1-\theta)^2 \frac{(\alpha\overline{\Phi}(1))^2 K^2}{\mathbb{E}[K^2]} = \big(\overline{\Phi}(1) - \beta\big)^2 \alpha^2 \ .$$

The Gaussian concentration inequality implies also that, with probability larger than $1 - e^{-x}$,

$$\|G\| \leqslant \mathbb{E}[\|G\|] + \sqrt{2\|\mathbf{M}\|_{\mathrm{op}} x} \ .$$

As $\mathbb{E}[\|G\|] \leqslant \mathrm{Tr}(\mathbf{M}) \leqslant 1$ and $\|\mathbf{M}\|_{\mathrm{op}} \leqslant \mathrm{Tr}(\mathbf{M}) \leqslant 1$, this implies that

$$\mathbb{P}(\|G\| \leqslant 1 + \sqrt{2x}) \geqslant 1 - e^{-x} \ .$$

For any $x > -2\log\big[\alpha\big(\overline{\Phi}(1) - \beta\big)\big]$, if $|\mathcal{A}_{\mathbf{M}}| \geqslant \alpha K$, the event

$$\{Z > \beta\alpha K\} \cap \{(\|G\| \leqslant 1 + \sqrt{2x}\} \neq \emptyset \ .$$

Hence, if $|\mathcal{A}_{\mathbf{M}}| \geqslant \alpha K$, there exists a vector g such that $\|g\| = 3\log\big[\alpha\big(\overline{\Phi}(1) - \beta\big)\big]$ and $\beta\alpha K$ blocks such that $|g^T(P_{B_k}X - \mu_P)| > br$. Fix $b = 3\log\big[\alpha\big(\overline{\Phi}(1) - \beta\big)\big]$ and $a = 2b$, the vector $\mathbf{v} = g/b \in \mathbf{S}$ satisfies $|\mathcal{K}_{\mathbf{v}}| < (1 - \beta\alpha)K$ on $|\mathcal{A}_{\mathbf{M}}| \geqslant \alpha K$. Therefore, the event $|\mathcal{A}_{\mathbf{M}}| \geqslant \alpha K$ is by definition contained in Ω^c, it has probability smaller than e^{-K/c_α}. □

Fix C_α as in Lemma 109 and in the remaining of this section, fix

$$r = C_\alpha \frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}} K}}{\sqrt{N}} \ .$$

Let Ω_α denote the event where there exist more than $(1-\alpha)K$ blocs B_k satisfying

$$\sup_{\mathbf{M} \in \mathcal{S}_1} \|\mathbf{M}^{1/2}(P_{B_k} X - \mu_P)\| \leqslant r \ .$$

The triangular inequality gives the following corollary of Lemma 109.

Corollary 111. *On Ω_α, for any $\boldsymbol{M} \in \mathcal{S}_1$, there are more than $(1-\alpha)K$ blocks such that, for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\boldsymbol{M}^{1/2}(\mu_P - x)\| - r \leqslant \|\boldsymbol{M}^{1/2}(P_{B_k} X - \boldsymbol{x})\| \leqslant \|\boldsymbol{M}^{1/2}(\mu_P - x)\| + r \ .$$

9.3 Toward a convex relaxation

The section introduces an optimization problem whose solutions are proved in Section 9.4 are used in the iteration step of the algorithm. This problem is solved in Section 9.5.

For any $\mathbf{w} \in \mathbb{R}^K$ and $\mathbf{x} \in \mathbb{R}^d$, let

$$\widehat{\mathbf{M}}_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{K} w_k (P_{B_k} X - \mathbf{x})(P_{B_k} X - \mathbf{x})^T \ .$$

Let

$$\Delta_K = \{w \in [0, 1/[(1-\alpha)K]]^K : \sum_{k=1}^{K} w_k = 1\} \ .$$

Denote by

$$\mathrm{OPT}(\mathbf{x}) = \sup_{\mathbf{M} \in \mathcal{S}_1} \inf_{\mathbf{w} \in \Delta_K} \mathrm{Tr}(\mathbf{M}\widehat{\mathbf{M}}_{\mathbf{w}}(\mathbf{x})) \ .$$

Let also $h_{\mathbf{x}}$ denote the following function.

$$h_{\mathbf{x}} : \mathbf{M} \mapsto \inf_{\mathbf{w} \in \Delta_K} \mathrm{Tr}(\mathbf{M}\widehat{\mathbf{M}}_{\mathbf{w}}(\mathbf{x}))$$

Let $I \subset \{1, \dots, K\}$ denote the set of indices k such that $(P_{B_k} X - \mathbf{x})^T \mathbf{M} (P_{B_k} X - \mathbf{x})$ is one of the $(1-\alpha)K$ smallest values among the $((P_{B_j} X - \mathbf{x})^T \mathbf{M} (P_{B_j} X - \mathbf{x}))_{j \in \{1,\dots,K\}}$. The infimum in the definition of $h_{\mathbf{x}}(\mathbf{M})$ is achieved (it is therefore a minimum) by the vector $\mathbf{w}$ such that

$$w_k = \begin{cases} 1/[(1-\alpha)K] & \text{if } k \in I \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{9.2}$$

The following lemma bounds $\mathrm{OPT}(\mathbf{x})$ when $\mathbf{x}$ is far from μ_P.

Lemma 112. *On Ω_α, for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$OPT(\boldsymbol{x}) \leqslant (\|\boldsymbol{x} - \mu_P\| + r)^2 \ .$$

Moreover, for any $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x} - \mu_P\| \geqslant r$,

$$OPT(\boldsymbol{x}) \geqslant \frac{1-2\alpha}{1-\alpha} (\|\boldsymbol{x} - \mu_P\| - r)^2 \ .$$

Proof. Fix $\mathbf{M} \in \mathcal{S}_1$ and $\mathbf{x} \in \mathbb{R}^d$. Let

$$\mathcal{K}_\mathbf{M} = \{k \in \{1, \dots, K\} : \|\mathbf{M}^{1/2}(P_{B_k}X - \mu_P)\| \leqslant r\} \ .$$

On Ω_α, $|\mathcal{K}_\mathbf{M}| \geqslant (1-\alpha)K$. By the triangular inequality, for any $k \in \mathcal{K}_\mathbf{M}$,

$$\|\mathbf{M}^{1/2}(\mu_P - \mathbf{x})\| - r \leqslant \|\mathbf{M}^{1/2}(P_{B_k}X - \mathbf{x})\| \leqslant \|\mathbf{M}^{1/2}(\mu_P - \mathbf{x})\| + r \ . \quad (9.3)$$

Define the vector $\mathbf{w} \in \mathbb{R}^K$ as follows:

$$w_k = \begin{cases} \frac{1}{|\mathcal{K}_\mathbf{M}|} & \text{if } k \in \mathcal{K}_\mathbf{M} \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

On Ω_α, $\mathbf{w} \in \Delta_K$, so, by definition of $h_\mathbf{x}$,

$$\begin{aligned} h_\mathbf{x}(\mathbf{M}) &\leqslant \mathrm{Tr}(\mathbf{M}\widehat{\mathbf{M}}_\mathbf{w}(\mathbf{x})) = \sum_{k=1}^K w_k (P_{B_k}X - \mathbf{x})^T \mathbf{M} (P_{B_k}X - \mathbf{x}) \\ &= \frac{1}{|\mathcal{K}_\mathbf{M}|} \sum_{k \in \mathcal{K}_\mathbf{M}} \|\mathbf{M}^{1/2}(P_{B_k}X - \mathbf{x})\|^2 \ . \end{aligned}$$

By (9.3), this implies that

$$h_\mathbf{x}(\mathbf{M}) \leqslant (\|\mathbf{M}^{1/2}(\mu_P - \mathbf{x})\| + r)^2 \ . \quad (9.4)$$

Taking the supremum over all $\mathbf{M} \in \mathcal{S}_1$ in this inequality shows that

$$\mathrm{OPT}(\mathbf{x}) \leqslant (\|\mathbf{x} - \mu_P\| + r)^2 \ .$$

Let now $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x} - \mu_P\| > r$. Fix also $\mathbf{M} \in \mathcal{S}_1$ and define I as in the definition of the optimal weights $\mathbf{w}$ in (9.2). On Ω_α, both $|I|$ and $|\mathcal{K}_\mathbf{M}|$ are larger than $(1-\alpha)K$, so $|I \cap \mathcal{K}_\mathbf{M}| \geqslant (1-2\alpha)K$, so

$$\begin{aligned} h_\mathbf{x}(\mathbf{M}) &= \frac{1}{(1-\alpha)K} \sum_{k \in I} \|\mathbf{M}^{1/2}(P_{B_k}X - \mathbf{x})\|^2 \\ &\geqslant \frac{1}{(1-\alpha)K} \sum_{k \in I \cap \mathcal{K}_M} \|\mathbf{M}^{1/2}(P_{B_k}X - \mathbf{x})\|^2 \ . \end{aligned}$$

By (9.3), this implies that

$$h_\mathbf{x}(\mathbf{M}) \geqslant \frac{1-2\alpha}{1-\alpha} (\|\mathbf{M}^{1/2}(\mu_P - \mathbf{x})\| - r)^2 \ .$$

Taking the supremum over all $\mathbf{M} \in \mathcal{S}_1$ in this inequality shows that

$$\mathrm{OPT}(\mathbf{x}) \geqslant \frac{1-2\alpha}{1-\alpha} (\|\mathbf{x} - \mu_P\| - r)^2 \ .$$

□

Lemma 113. *Let $\beta \in [1/\sqrt{2}, 1]$. On Ω_α, for any $\boldsymbol{M} \in \mathcal{S}_1$ such that $h_{\boldsymbol{x}}(\boldsymbol{M}) \geqslant (\beta\|\boldsymbol{x} - \mu_P\| + r)^2$, the (normalized) top eigenvector $\boldsymbol{v}$ of $\boldsymbol{M}$ satisfies*

$$\left| \frac{\boldsymbol{v}^T(\boldsymbol{x} - \mu_P)}{\|\boldsymbol{x} - \mu_P\|} \right| > \sqrt{2\beta^2 - 1} \ .$$

Proof. Let $\mathbf{M}$ satisfying the assumptions of the lemma. From (9.4),

$$h_{\mathbf{x}}(\mathbf{M}) \leqslant (\|\mathbf{M}^{1/2}(\mu_P - \mathbf{x})\| + r)^2 \ .$$

Let $\mathbf{u} = (\mathbf{x} - \mu_P)/\|\mathbf{x} - \mu_P\|$. This implies that

$$\beta\|\mathbf{x} - \mu_P\| \leqslant \|\mathbf{M}^{1/2}(\mu_P - \mathbf{x})\|, \qquad \text{i.e.} \qquad \|\mathbf{M}\|_{\text{op}} \geqslant \mathbf{u}^T\mathbf{M}\mathbf{u} \geqslant \beta^2 \ .$$

Moreover, $\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}$ and $\mathbf{M}[\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}]$ are orthogonal to $\mathbf{v}$, hence,

$$\mathbf{u}^T\mathbf{M}\mathbf{u} = (\mathbf{u}^T\mathbf{v})^2\mathbf{v}^T\mathbf{M}\mathbf{v} + [\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}]^T\mathbf{M}[\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}] \ .$$

First, $\mathbf{v}^T\mathbf{M}\mathbf{v} = \|\mathbf{M}\|_{\text{op}} \leqslant 1$. Second, as $\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}$ is orthogonal to $\mathbf{v}$, $[\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}]^T\mathbf{M}[\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}]$ does not exceed the second largest eigenvalue λ of $\mathbf{M}$. As $\lambda + \|\mathbf{M}\|_{\text{op}} \leqslant \text{Tr}(\mathbf{M}) \leqslant 1$, it follows that

$$[\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}]^T\mathbf{M}[\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}] \leqslant 1 - \|\mathbf{M}\|_{\text{op}} \leqslant 1 - \beta^2 \ .$$

Hence,

$$\beta^2 \leqslant (\mathbf{u}^T\mathbf{v})^2 + 1 - \beta^2 \ ,$$

which proves Lemma 113. □

9.4 The iteration step

We are now in position to show that a solution of the optimization problem defined in Section 9.3 can be used to define the iteration step of our algorithm.

Given $\mathbf{x} \in \mathbb{R}^d$, assume that we are given an approximation $\widehat{\mathbf{M}}_{\mathbf{x}}$ of

$$\widehat{\mathbf{M}}_* \in \text{argmax}_{\mathbf{M}\in\mathcal{S}_1} \inf_{\mathbf{w}\in\Delta_K} \text{Tr}(\mathbf{M}\widehat{\mathbf{M}}_{\mathbf{w}}(\mathbf{x})) \ .$$

This approximation should satisfy the following requirement: There exists an absolute constant A such that, if $\|\mathbf{x} - \mu_P\| \geqslant Ar$, then, there exists $\beta \geqslant 0.8$ such that

$$h_{\mathbf{x}}(\widehat{\mathbf{M}}_{\mathbf{x}}) \geqslant (\beta\|\mathbf{x} - \mu_P\| + r)^2 \ .$$

Let then $\widehat{\mathbf{v}}_{\mathbf{x}} \in \mathbf{S}$ denote a top eigenvector of $\widehat{\mathbf{M}}_{\mathbf{x}}$ and let

$$\theta_{\mathbf{x}} = -\text{median}(\widehat{\mathbf{v}}_{\mathbf{x}}^T(P_{B_k}X - \mathbf{x}), k = 1, \dots, K) \ .$$

The algorithm moves at each iteration from $\mathbf{x}$ to $T(\mathbf{x})$, where,

$$T(\mathbf{x}) = \mathbf{x} - \theta_{\mathbf{x}}\widehat{\mathbf{v}}_{\mathbf{x}} \ . \tag{9.5}$$

Remark 114. *The vector $\boldsymbol{v}_x$ is defined up to its sign. Whatever this sign, the function T is well defined.*

Proposition 115. *If $\alpha \leqslant 1/2$, on Ω_α,*

$$\|T(\boldsymbol{x}) - \mu_P\|^2 \leqslant \frac{3}{4}\|\boldsymbol{x} - \mu_P\|^2 + (A^2 + 1)r^2 \ .$$

Proof. The idea of the proof is the following decomposition of the distance between $T(\mathbf{x})$ and μ_P. Let $\mathbf{v} = (\mathbf{x} - \mu_P)/\|\mathbf{x} - \mu_P\|$ denote the optimal descent direction and decompose $\mathbf{v} = a\widehat{\mathbf{v}}_\mathbf{x} + b\widehat{\mathbf{v}}_\mathbf{x}^\perp$, where $\widehat{\mathbf{v}}_\mathbf{x}^\perp \in \mathbf{S} \cap \{\widehat{\mathbf{v}}_\mathbf{x}\}^\perp$ and, therefore $a^2 + b^2 = 1$. We have, by Pythagoras relation

$$\begin{aligned}\|T(\mathbf{x}) - \mu_P\|^2 &= \|\mathbf{x} - \mu_P - \theta_\mathbf{x}\widehat{\mathbf{v}}_\mathbf{x}\|^2 \\ &= \|\|\mu_P - \mathbf{x}\|(a\widehat{\mathbf{v}}_\mathbf{x} + b\widehat{\mathbf{v}}_\mathbf{x}^\perp) - \theta_\mathbf{x}\widehat{\mathbf{v}}_\mathbf{x}\|^2 \\ &= \|(a\|\mu_P - \mathbf{x}\| - \theta_\mathbf{x})\widehat{\mathbf{v}}_\mathbf{x} + b\|\mu_P - \mathbf{x}\|\widehat{\mathbf{v}}_\mathbf{x}^\perp\|^2 \\ &= (a\|\mu_P - \mathbf{x}\| - \theta_\mathbf{x})^2 + b^2\|\mu_P - \mathbf{x}\|^2 \ .\end{aligned}$$

Since $a\|\mu_P - \mathbf{x}\| = (\mathbf{x} - \mu_P)^T\widehat{\mathbf{v}}_\mathbf{x}$ and $b = \mathbf{v}^T\widehat{\mathbf{v}}_\mathbf{x}^\perp$, this relation can be written

$$\|T(\mathbf{x}) - \mu_P\|^2 = [(\mathbf{x} - \mu_P)^T\widehat{\mathbf{v}}_\mathbf{x} - \theta_\mathbf{x}]^2 + [\mathbf{v}^T\widehat{\mathbf{v}}_\mathbf{x}^\perp]^2\|\mathbf{x} - \mu_P\|^2 \ . \tag{9.6}$$

We bound separately each term in this decomposition. On Ω_α, there are more than $(1-\alpha)K$ blocks such that

$$|\widehat{\mathbf{v}}_\mathbf{x}^T(P_{B_k}X - \mu_P)| \leqslant r \ .$$

On the same blocks

$$|\widehat{\mathbf{v}}_\mathbf{x}^T(P_{B_k}X - \mathbf{x}) - \widehat{\mathbf{v}}_\mathbf{x}^T(\mu_P - \mathbf{x})| \leqslant r \ .$$

Therefore, if $\alpha < 1/2$, on Ω_α,

$$|\widehat{\mathbf{v}}_\mathbf{x}^T(\mathbf{x} - \mu_P) - \theta_\mathbf{x}| \leqslant r \ . \tag{9.7}$$

If $\|\mathbf{x} - \mu_P\| > Ar$, there exists $\beta \geqslant 0.8$ such that

$$h_\mathbf{x}(\widehat{\mathbf{M}}_\mathbf{x}) \geqslant (\beta\|\mathbf{x} - \mu_P\| + r)^2 \ .$$

By Lemma 113, this implies that, on Ω_α, $|\mathbf{v}^T\widehat{\mathbf{v}}_\mathbf{x}| \geqslant \sqrt{2\beta^2 - 1}$, so $|a| \geqslant \sqrt{2\beta^2 - 1}$ and $b^2 = 1 - a^2 \leqslant 2(1 - \beta^2) \leqslant 3/4$. Plugging this inequality and (9.7) into (9.6) yields the result when $\|\mathbf{x} - \mu_P\| > Ar$.

If $\|\mathbf{x} - \mu_P\| \leqslant Ar$, as $|\mathbf{v}_x^T\mathbf{v}| \leqslant 1$, from (9.7) and (9.6),

$$\|T(\mathbf{x}) - \mu_P\|^2 \leqslant (A^2 + 1)r^2 \ .$$

This proves the result when $\|\mathbf{x} - \mu_P\| \leqslant Ar$. □

9.5 Computation of $\widehat{\mathbf{M}}_\mathbf{x}$.

It remains to compute an approximation $\widehat{\mathbf{M}}_\mathbf{x}$ of

$$\widehat{\mathbf{M}}_* \in \mathrm{argmax}_{\mathbf{M}\in\mathcal{S}_1} \inf_{\mathbf{w}\in\Delta_K} \mathrm{Tr}(\mathbf{M}\widehat{\mathbf{M}}_\mathbf{w}(\mathbf{x})) = \mathrm{argmax}_{\mathbf{M}\in\mathcal{S}_1} h_\mathbf{x}(\mathbf{M}) \ ,$$

satisfying the requirement that there exists an absolute constant A such that, if $\|\mathbf{x} - \mu_P\| \geqslant Ar$, then, there exists $\beta \geqslant 0.8$ such that

$$h_\mathbf{x}(\widehat{\mathbf{M}}_\mathbf{x}) \geqslant (\beta\|\mathbf{x} - \mu_P\| + r)^2 \ .$$

We proceed in several steps. Section 9.5.1 presents an equivalent convex problem and Section 9.5.2 presents a convex problem whose solutions approximate those of the equivalent problem in the desired way. This approximating problem is solved using the algorithms of [49, 16] in Section 9.5.3.

The approximating depends on a parameter ρ that should be carefully chosen. Section 9.5.4 gathers technical lemmas that are used in Section 9.5.5 to calibrate ρ so as to ensure that the solution of the approximating satisfies the requirement that there exists an absolute constant A such that, if $\|\mathbf{x}-\mu_P\| \geqslant Ar$, then, there exists $\beta \geqslant 0.8$ such that

$$h_{\mathbf{x}}(\widehat{\mathbf{M}}_{\mathbf{x}}) \geqslant (\beta\|\mathbf{x}-\mu_P\|+r)^2 \ .$$

9.5.1 An equivalent problem

Consider the following convex maximization problem. The set of constraint $\mathcal{C}$ is the set of triplets $z \geqslant 0$, $\mathbf{y} \in \mathbb{R}_+^K$ and $\mathbf{M} \in \mathcal{S}_1$ satisfying, for any $k \in \{1,\dots,K\}$, $(P_{B_k}X-\mu_P)^T\mathbf{M}(P_{B_k}X-\mu_P)+y_k \geqslant z$. The goal is to find $(z,\mathbf{y},\mathbf{M}) \in \mathcal{C}$ maximizing the objective function

$$z - \|\mathbf{y}\|_1/[(1-\alpha)K] \ .$$

The link with $\widehat{\mathbf{M}}_*$ is provided by the following lemma.

Lemma 116. *Fix $\boldsymbol{M} \in \mathcal{S}_1$ and define $\mathcal{C}_{\boldsymbol{M}}$, the set of couples $z \geqslant 0$ and $\boldsymbol{y} \in \mathbb{R}_+^K$ such that $(z,\boldsymbol{y},\boldsymbol{M}) \in \mathcal{C}$ that is, such that, for any $k \in \{1,\dots,K\}$, $(P_{B_k}X-\mu_P)^T\boldsymbol{M}(P_{B_k}X-\mu_P)+y_k \geqslant z$. Then*

$$\max_{(z,y)\in\mathcal{C}_{\boldsymbol{M}}}\{z-\|\boldsymbol{y}\|_1/[(1-\alpha)K]\} = h_{\boldsymbol{x}}(\boldsymbol{M}) \ .$$

This maximal value is achieved when $z = z_{\boldsymbol{M}}$, the $(1-\alpha)K$ largest values among the set $\{(P_{B_k}X-\mu_P)^T\boldsymbol{M}(P_{B_k}X-\mu_P), k=\{1,\dots,K\}\}$ and $\boldsymbol{y}=\boldsymbol{y}_{\boldsymbol{M}}$, where the coordinates of $\boldsymbol{y}_{\boldsymbol{M}}$ are defined by

$$y_k = \big(z_{\boldsymbol{M}} - (P_{B_k}X-\mu_P)^T\boldsymbol{M}(P_{B_k}X-\mu_P)\big)_+ \ .$$

Proof. To compute $\max_{(z,y)\in\mathcal{C}_{\mathbf{M}}}\{z-\|\mathbf{y}\|_1/[(1-\alpha)K]\}$, each $y_k \geqslant 0$ should be chosen as small as possible, thus the constraints imply that the maximum in $\mathbf{y}$ is achieved, for each given z, by

$$y_k = \big(z - (P_{B_k}X-\mu_P)^T\mathbf{M}(P_{B_k}X-\mu_P)\big)_+ \ .$$

For this value of $\mathbf{y}$, one gets

$$z - \frac{\|\mathbf{y}\|_1}{(1-\alpha)K} = z - \frac{1}{(1-\alpha)K}\sum_{k=1}^{K}\big(z-(P_{B_k}X-\mu_P)^T\mathbf{M}(P_{B_k}X-\mu_P)\big)_+ \ .$$

Assume that $(P_{B_k}X-\mu_P)^T\mathbf{M}(P_{B_k}X-\mu_P)$ are arranged in non-decreasing order. Then, if $z \in [(P_{B_k}X-\mu_P)^T\mathbf{M}(P_{B_k}X-\mu_P), (P_{B_{k+1}}X-\mu_P)^T\mathbf{M}(P_{B_{k+1}}X-\mu_P))$,

$$z-\frac{\|\mathbf{y}\|_1}{(1-\alpha)K} = z\left(1-\frac{k}{(1-\alpha)K}\right)+\frac{1}{(1-\alpha)K}\sum_{j=1}^{k}(P_{B_j}X-\mu_P)^T\mathbf{M}(P_{B_j}X-\mu_P) \ .$$

This function of k is

$$\begin{cases} \text{non-decreasing on the interval } [0, (P_{B_{(1-\alpha)K}}X - \mu_P)^T \mathbf{M} (P_{B_{(1-\alpha)K}}X - \mu_P)] \ , \\ \text{non-increasing on } [(P_{B_{(1-\alpha)K}}X - \mu_P)^T \mathbf{M} (P_{B_{(1-\alpha)K}}X - \mu_P), +\infty) \ . \end{cases}$$

It maximal value is achieved for $z = z_{\mathbf{M}}$ and is equal to

$$\frac{1}{(1-\alpha)K} \sum_{j=1}^{(1-\alpha)K} (P_{B_j}X - \mu_P)^T \mathbf{M} (P_{B_j}X - \mu_P) = h_{\mathbf{x}}(\mathbf{M}) \ .$$

□

A first consequence of Lemma 116 is that $\widehat{\mathbf{M}}_*$ satisfies

$$(z_{\widehat{\mathbf{M}}_*}, \mathbf{y}_{\widehat{\mathbf{M}}_*}, \widehat{\mathbf{M}}_*) \in \mathrm{argmax}_{(z,\mathbf{y},\mathbf{M})\in\mathcal{C}} \{z - \|\mathbf{y}\|_1/[(1-\alpha)K]\} \ . \tag{9.8}$$

9.5.2 An approximating problem.

Consider now the following convex optimization problem. Let $\rho > 0$ and define the constraint set $\mathcal{C}_\rho$ as the set of couples $\mathbf{M} \in \mathbb{R}^{K\times K}, \mathbf{y} \in \mathbb{R}^K$ where $\mathbf{M} \succeq 0$ and, for any $k \in \{1, \ldots, K\}$,

$$y_k \geqslant 0, \qquad \rho(P_{B_k}X - \mathbf{x})^T \mathbf{M} (P_{B_k}X - \mathbf{x}) + (1-\alpha)K y_k \geqslant 1 \ .$$

The problem is to find a minimizer on the constraint set $\mathcal{C}_\rho$ of the objective function

$$\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \ ,$$

A useful link between this problem and the one of Section 9.5.1 is provided in the following lemma.

Lemma 117. *If we have built $(\boldsymbol{M}, \boldsymbol{y}) \in \mathcal{C}_\rho$ such that*

$$Tr(\boldsymbol{M}) + \|\boldsymbol{y}\|_1 \leqslant 1 \ ,$$

then one can build in $O(K)$ operations $(z, \boldsymbol{y}', \boldsymbol{M}') \in \mathcal{C}$ such that

$$z - \frac{\|\boldsymbol{y}'\|_1}{(1-\alpha)K} \geqslant \frac{1}{\rho} \ .$$

Conversely, if we have built $(z, \boldsymbol{y}', \boldsymbol{M}') \in \mathcal{C}$ such that

$$z - \frac{\|\boldsymbol{y}'\|_1}{(1-\alpha)K} \geqslant \frac{1}{\rho} \ ,$$

then one can build in $O(1)$ operations $(\boldsymbol{M}, \boldsymbol{y}) \in \mathcal{C}_\rho$ such that

$$Tr(\boldsymbol{M}) + \|\boldsymbol{y}\|_1 \leqslant 1 \ .$$

In both cases, the top eigenvectors of the matrices $\boldsymbol{M}$ and $\boldsymbol{M}'$ are equal.

Proof. Suppose that we have built $(\mathbf{M}, \mathbf{y}) \in \mathcal{C}_\rho$ such that

$$\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant 1 \ .$$

Then, one can define

$$\mathbf{M}' = \frac{\mathbf{M}}{\mathrm{Tr}(\mathbf{M})}, \qquad \mathbf{y}' = \frac{(1-\alpha)K}{\rho \mathrm{Tr}(\mathbf{M})}\mathbf{y} \ .$$

Then $\mathbf{M}' \in \mathcal{S}_1$ and, for all $k \in \{1, \dots, K\}$,

$$y'_k \geqslant 0, \qquad (P_{B_k}X - \mathbf{x})^T \mathbf{M}' (P_{B_k}X - \mathbf{x}) + y'_k \geqslant \frac{1}{\rho \mathrm{Tr}(\mathbf{M})} \ .$$

Therefore, $(z, \mathbf{y}') \in \mathcal{C}_{\mathbf{M}'}$, where $z = 1/(\rho \mathrm{Tr}(\mathbf{M}))$ and

$$z - \frac{\|\mathbf{y}'\|_1}{(1-\alpha)K} = \frac{1 - \|\mathbf{y}\|_1}{\rho \mathrm{Tr}(\mathbf{M})} \geqslant \frac{1}{\rho} \ .$$

Conversely, if we have found $\mathbf{M}$ such that $z_{\mathbf{M}} - \|\mathbf{y}_{\mathbf{M}}\|_1/[(1-\alpha)K] \geqslant 1/\rho$, one can define

$$\mathbf{M}' = \frac{1}{\rho z_{\mathbf{M}}}\mathbf{M}, \qquad \mathbf{y}' = \frac{1}{(1-\alpha)K z_{\mathbf{M}}}\mathbf{y}_{\mathbf{M}} \ .$$

We have $\mathbf{M}' \succeq 0$, and, for all $k \in \{1, \dots, K\}$,

$$\rho (P_{B_k}X - \mathbf{x})^T \frac{1}{\rho z_{\mathbf{M}}} \mathbf{M} (P_{B_k}X - \mathbf{x}) + (1-\alpha)K \frac{y_k}{(1-\alpha)K z_{\mathbf{M}}} \geqslant 1 \ ,$$

that is $(\mathbf{M}', \mathbf{y}') \in \mathcal{C}_\rho$. Moreover,

$$\mathrm{Tr}(\mathbf{M}') + \|\mathbf{y}'\|_1 = \frac{1}{\rho z_{\mathbf{M}}} + \frac{1}{(1-\alpha)K z_{\mathbf{M}}} \|\mathbf{y}_{\mathbf{M}}\|_1 \leqslant \frac{1}{\rho z_{\mathbf{M}}} + \frac{z_{\mathbf{M}} - 1/\rho}{z_{\mathbf{M}}} = 1 \ .$$

□

9.5.3 Solving the approximating problem in nearly linear time

The following Lemma comes from [49], see also [16, Section 4].

Lemma 118. *For every $\rho > 0$ and $\eta > 0$, there exists an algorithm $\mathcal{A} : (\mathbb{R}^d)^n \times [0,1] \to \mathcal{C}_\rho$ such that, if U is a uniform random variable on $[0,1]$ independent of $\mathcal{D}_N$ and $\mathcal{A}(\mathcal{D}_N, U) = (\boldsymbol{M}, \boldsymbol{y})$,*

$$\mathbb{P}(\mathit{Tr}(\boldsymbol{M}) + \|\boldsymbol{y}\|_1 \leqslant (1+\eta) g(\rho) | \mathcal{D}_N) \geqslant 1 - \frac{1}{e} \ .$$

Moreover, $\mathcal{A}(\mathcal{D}_N, U)$ can be evaluated in $O(Kd)$ operations and a top eigenvector of $\boldsymbol{M}$ can be computed in $\tilde{O}(Kd)$ operations.

A consequence of Lemma 118 is the following result.

Lemma 119. *For every $\rho > 0$, $\eta > 0$, $R > 0$, every positive integer u and every $\boldsymbol{x}$ in $\mathbb{R}^d$, one can build in $O((u + \log R)Kd)$ operations $\mathit{Alg}(\mathcal{D}_N, \rho, \eta, u, \boldsymbol{x}) = (\boldsymbol{M}, \boldsymbol{y}) \in \mathcal{C}_\rho$ such that*

$$\mathbb{P}(\mathit{Tr}(\boldsymbol{M}) + \|\boldsymbol{y}\|_1 \leqslant (1+\eta) g(\rho)) \geqslant 1 - \frac{e^{-u}}{R} \ .$$

Proof. Let $t = u + \log R$ and let $U_1, \ldots, U_t$ denote independent random variables with uniform distribution on $[0,1]$, independent from $\mathcal{D}_N$. Let $\mathcal{A}(\mathcal{D}_N, U_1) = (\mathbf{M}_1, \mathbf{y}_1), \ldots, \mathcal{A}(\mathcal{D}_N, U_t) = (\mathbf{M}_t, \mathbf{y}_t)$ denote the random variables built with the algorithm $\mathcal{A}$ of Lemma 118 and let $\hat{t}$ such that $\mathrm{Tr}(\mathbf{M}_{\hat{t}}) + \|\mathbf{y}_{\hat{t}}\|_1$ is minimal. As the random variables $\mathcal{A}(\mathcal{D}_N, U_1), \ldots, \mathcal{A}(\mathcal{D}_N, U_t)$ are i.i.d. conditionally on $\mathcal{D}_N$,

$$\mathbb{P}(\mathrm{Tr}(\mathbf{M}_{\hat{t}}) + \|\mathbf{y}_{\hat{t}}\|_1 > (1+\eta)g(\rho)) = (\mathbb{P}(\mathrm{Tr}(\mathbf{M}_1) + \|\mathbf{y}_1\|_1 > (1+\eta)g(\rho)))^t \ .$$

The result then follows from Lemma 118. □

9.5.4 The optimal solution of the approximating problem

For any $\rho > 0$, let

$$g(\rho) = \min_{(\mathbf{M},\mathbf{y}) \in \mathcal{C}_\rho} \{\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1\} \ .$$

Lemma 120. *For any $\rho' > \rho > 0$,*

$$g(\rho) \geqslant g(\rho') \geqslant \frac{\rho}{\rho'} g(\rho) \ .$$

Proof. Clearly, $\mathcal{C}_\rho \subset \mathcal{C}_{\rho'}$, so $g(\rho) \geqslant g(\rho')$. Moreover, if $(\mathbf{M}, \mathbf{y}) \in \mathcal{C}_{\rho'}$, and $r = \rho'/\rho$, $(r\mathbf{M}, r\mathbf{y}) \in \mathcal{C}_\rho$, so $g(\rho') \geqslant r g(\rho)$. □

It follows from Lemma 120 that g is non-increasing and continuous. Moreover, from Lemma 117, it satisfies $g(\rho) \leqslant 1$ iff $1/\mathrm{OPT}(\mathbf{x}) \geqslant \rho$, so

$$g(1/\mathrm{OPT}(\mathbf{x})) = 1 \ .$$

Lemma 121. *On the event Ω_α, for all $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x} - \mu_P\| > r$,*

$$g(\rho) \leqslant \frac{1}{\rho Opt(\boldsymbol{x})} + \frac{9(\|\boldsymbol{x} - \mu_P\| + r)^2}{8(\|\boldsymbol{x} - \mu_P\| - r)^2} - 1 \ .$$

Proof. For any $\nu > 0$, there exists $\mathbf{M}_0 \in \mathcal{S}_1$ such that $z_{\mathbf{M}_0}, \mathbf{y}_{\mathbf{M}_0}$ (as defined in Lemma 116) satisfy

$$\begin{aligned} z_{\mathbf{M}_0} - \frac{\|\mathbf{y}_{\mathbf{M}_0}\|_1}{(1-\alpha)K} &> \sup_{(z,\mathbf{y},\mathbf{M}) \in \mathcal{C}} \left\{ z - \frac{\|\mathbf{y}\|_1}{(1-\alpha)K} \right\} - \nu \\ &= \sup_{\mathbf{M} \in \mathcal{S}_1} \sup_{(z,\mathbf{y}) \in \mathcal{C}_\mathbf{M}} \left\{ z - \frac{\|\mathbf{y}\|_1}{(1-\alpha)K} \right\} - \nu \ . \end{aligned}$$

By Lemma 116,

$$\sup_{(z,\mathbf{y}) \in \mathcal{C}_\mathbf{M}} \left\{ z - \frac{\|\mathbf{y}\|_1}{(1-\alpha)K} \right\} = h_\mathbf{x}(\mathbf{M}) \ .$$

Therefore,

$$z_{\mathbf{M}_0} - \frac{\|\mathbf{y}_{\mathbf{M}_0}\|_1}{(1-\alpha)K} > \sup_{\mathbf{M} \in \mathcal{S}_1} h_\mathbf{x}(\mathbf{M}) - \nu = \mathrm{OPT}(\mathbf{x}) - \nu \ .$$

Since $\|\mathbf{x} - \mu_P\| > r$, from Lemma 112, on Ω_α,

$$\mathrm{OPT}(\mathbf{x}) \geqslant \frac{1-2\alpha}{1-\alpha} (\|\mathbf{x} - \mu_P\| - r)^2 \ .$$

From Lemma 116, $z_{\mathbf{M}_0}$ is the $(1-\alpha)K$ largest value in the set $\{(P_{B_k}X - \mu_P)^T\mathbf{M}(P_{B_k}X - \mu_P), k = \{1,\dots,K\}\}$. It follows from Corollary 111 that, on Ω_α,

$$z_{\mathbf{M}_0} \leqslant (\|\mathbf{M}_0^{1/2}(\mathbf{x}-\mu_P)\| + r)^2 \leqslant (\|\mathbf{x}-\mu_P\| + r)^2 \ .$$

Define

$$\mathbf{M}' = \frac{1}{\rho z_{\mathbf{M}_0}}\mathbf{M}_0, \qquad \mathbf{y}' = \frac{1}{(1-\alpha)Kz_{\mathbf{M}_0}}\mathbf{y}_{\mathbf{M}_0} \ .$$

We have proved that

$$\begin{aligned} g(\rho) &\leqslant \mathrm{Tr}(\mathbf{M}') + \|\mathbf{y}'\|_1 = \frac{1}{\rho z_{\mathbf{M}_0}} + \frac{z_{\mathbf{M}_0} + \nu - \mathrm{Opt}(\mathbf{x})}{z_{\mathbf{M}_0}} \\ &\leqslant \frac{1}{\rho(\mathrm{Opt}(\mathbf{x}) - \nu)} + \frac{\nu + \mathrm{Opt}(\mathbf{x})\left(\frac{(1-\alpha)\|\mathbf{x}-\mu_P\|+r)^2}{(1-2\alpha)\|\mathbf{x}-\mu_P\|-r)^2} - 1\right)}{\mathrm{Opt}(\mathbf{x}) - \nu} \ . \end{aligned}$$

As the result holds for any $\nu > 0$, this concludes the proof. □

9.5.5 Calibration of the approximating algorithm

Assume that $\|\mathbf{x} - \mu_P\| > Ar$. On Ω_α, by Lemma 121,

$$g(\rho) \leqslant \frac{1}{\rho\mathrm{Opt}(\mathbf{x})} + \frac{9(A+1)^2}{8(A-1)^2} - 1 \leqslant \frac{1}{\rho\mathrm{Opt}(\mathbf{x})} + b \ . \tag{9.9}$$

Here $b = 2\alpha/(1-2\alpha)$. The last inequality holds for any $A \geqslant A_\alpha$.

Lemma 122. *Fix $\epsilon \in (2\alpha/(1-\alpha), 0.4(1-\alpha)/(1-2\alpha))$. Assume that ρ satisfies $g(\rho) \geqslant 1 - \epsilon + b$. There exist constants A_α and $\beta > 0.8$ such that, on Ω_α, for any $\boldsymbol{x} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{x} - \mu_P\| > A_\alpha r$,*

$$h_{\boldsymbol{x}}(\boldsymbol{M}') \geqslant (\beta\|\boldsymbol{x} - \mu_P\|^2 + r)^2 \ .$$

Proof. For any ρ such that $g(\rho) \geqslant 1 - \epsilon + b$, it follows from (9.9) that

$$\frac{1}{\rho} \geqslant (1-\epsilon)\mathrm{Opt}(\mathbf{x}) \ .$$

In this case, by Lemma 117, one can build in $O(K)$ operations $(z, \mathbf{y}', \mathbf{M}') \in \mathcal{C}$ such that

$$z - \frac{\|\mathbf{y}'\|_1}{(1-\alpha)K} \geqslant \frac{1}{\rho} \geqslant (1-\epsilon)\mathrm{Opt}(\mathbf{x}) \ .$$

By Lemma 116, the matrix $\mathbf{M}'$ satisfies

$$h_{\mathbf{x}}(\mathbf{M}') \geqslant z - \frac{\|\mathbf{y}'\|_1}{(1-\alpha)K} \geqslant (1-\epsilon)\mathrm{Opt}(\mathbf{x}) \ .$$

By Lemma 112, it follows that, on Ω_α,

$$h_{\mathbf{x}}(\mathbf{M}') \geqslant (1-\epsilon)\frac{1-2\alpha}{1-\alpha}(\|\mathbf{x} - \mu_P\| - r)^2 \ .$$

The condition on ϵ implies the result provided that A is large enough. □

It remains to find ρ such that $g(\rho) \geqslant 1 - \epsilon + b$. To do this, we fix $\epsilon > b$, $\eta > 0$ and build ρ and u such that the algorithm of Lemma 119 outputs $\mathrm{Alg}(\mathcal{D}_N, \rho, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfying

$$(1+\eta)(1-\epsilon+b) \leqslant \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant (1+\eta)g(\rho) \leqslant 1 \ .$$

As $\rho \to \infty$, by Lemma 121, $g(\rho) \to b$. If $\alpha < 1/4$, $b < 1$, so $g(\rho) < 1$ for any ρ large enough. Fix

$$\eta \in \left(0, \frac{1-4\alpha}{2\alpha} \wedge \frac{\alpha}{1-\alpha} \wedge \frac{1}{6}\right) = \left(0, \left(\frac{1}{b} - 1\right) \wedge \frac{\alpha}{1-\alpha} \wedge \frac{1}{6}\right) \ .$$

In particular, $1/(1+\eta) > b$. Then, there exists ρ_0 such that $g(\rho_0) < 1/(1+\eta)$. For this value of ρ_0, by Lemma 119, the algorithm defined in this lemma outputs $\mathrm{Alg}(\mathcal{D}_N, \rho_0, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfying

$$\mathbb{P}(\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant 1 | \mathcal{D}_N) \geqslant 1 - \frac{e^{-u}}{R} \ .$$

On the other hand, when $\rho = 0$, the minimal value of $\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1$ is achieved by the couple $(\mathbf{M}, \mathbf{y}) \in \mathcal{C}_0$ such that $\mathbf{M} = 0$ and $\mathbf{y}$ is the vector with all coordinates equal to $1/[(1-\alpha)K]$. It follows that $g(0) = 1/(1-\alpha) > 1$.

Lemma 123. *For any fixed constant $\nu \in (6\eta, 1)$, it is possible to build in $O(T(u+\log R)Kd)$ operations, where $T = O(\log(\rho_0))$, a couple $(\boldsymbol{M}, \boldsymbol{y})$ satisfying*

$$\mathbb{P}(1 - \nu \leqslant \mathit{Tr}(\boldsymbol{M}) + \|\boldsymbol{y}\|_1 \leqslant 1 \wedge (1+\eta)g(\eta) | \mathcal{D}_N) \geqslant 1 - \frac{Te^{-u}}{R} \ .$$

If $R = \tau T$, this gives that, for any fixed constant $\nu \in (5\eta, 1)$, it is possible to build in $O(\log(\rho_0)(u + \log(\tau) + \log\log(\rho_0))K)$ operations, a couple $(\boldsymbol{M}, \boldsymbol{y})$ satisfying

$$\mathbb{P}(1 - \nu \leqslant \mathit{Tr}(\boldsymbol{M}) + \|\boldsymbol{y}\|_1 \leqslant 1 \wedge (1+\eta)g(\eta) | \mathcal{D}_N) \geqslant 1 - \frac{e^{-u}}{\tau} \ .$$

Proof. Fix ν_0 and ν_1 such that $(1-\nu_0) > (1+\eta)(1-\nu_1)$ and

$$\frac{2\eta}{1+\eta} < \nu_0 < \nu_1 < \frac{\nu - \eta + \nu\eta}{2} \ .$$

This is possible since $\nu > 5\eta$. $\nu_0 < \nu_1$ are chosen such that

$$\frac{1-2\nu_1}{1+\eta} \geqslant 1 - \nu \quad \text{and} \quad (1+\eta)\left(1 - \frac{\nu_0}{2}\right) \leqslant 1 \ .$$

Moreover, as $(1-\nu_0) > (1+\eta)(1-\nu_1)$, we have, for ρ satisfying $g(\rho) = 1 - \nu_1$ (which exists by continuity of g, $1 - \nu_1 = g(\rho) \leqslant (1+\eta)g(\rho) < 1 - \nu_0$. Hence, by Lemma 119, the algorithm defined in this lemma outputs $\mathrm{Alg}(\mathcal{D}_N, \rho, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfying

$$\mathbb{P}(1 - \nu_1 \leqslant \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant 1 - \nu_0 | \mathcal{D}_N) \geqslant 1 - e^{-u}/R \ .$$

Consider the following recursive algorithm:

1. Initialize $\rho_- = 0$, $\rho_+ = \rho_0$, $\mathrm{Alg}(\mathcal{D}_N, (\rho_+ + \rho_-)/2, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$, $V = \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1$.
2. if $V \in [1-\nu_1, 1-\nu_0]$, stop,
3. if $V < 1-\nu_1$, update $\rho_+ = (\rho_+ + \rho_-)/2$,
4. if $V > 1-\nu_0$, update $\rho_- = (\rho_+ + \rho_-)/2$,
5. update $\mathrm{Alg}(\mathcal{D}_N, (\rho_+ + \rho_-)/2, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$, $V = \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1$.
6. Return $\rho_* = (\rho_+ + \rho_-)/2$.

Clearly, on the event where $\mathrm{Alg}(\mathcal{D}_N, \rho_0, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfies

$$\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant 1 \ ,$$

this algorithm outputs ρ_* such that $\mathrm{Alg}(\mathcal{D}_N, \rho_*, \eta, u, \mathbf{x}) = (\mathbf{M}_*, \mathbf{y}_*)$ satisfies $\mathrm{Tr}(\mathbf{M}_*) + \|\mathbf{y}_*\|_1 \in [1-\nu_1, 1-\nu_0]$. Such a value exists, at least on an event with large probability, by the discussion preceding the algorithm. Let T denote a number of steps to be defined later. Using a union bound in Lemma 119, with probability larger than $1 - Te^{-u}/R$, for all ρ in the set of all $(\rho_+ + \rho_-)/2$ along the first T steps of the algorithm, $\mathrm{Alg}(\mathcal{D}_N, \rho, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfies

$$g(\rho) \leqslant \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant (1+\eta) g(\rho) \ .$$

Moreover, after k steps of the algorithm,

$$\rho_+ - \rho_- = \frac{\rho_0}{2^k} \ .$$

On this event ρ_* satisfies $g(\rho_*) \in [(1-\nu_1)/(1+\eta), 1-\nu_0]$. Now, by continuity of g (see Lemma 120), there exists a constant $\delta > 0$ such that, for any $\rho \in [\rho_* - \delta, \rho_* + \delta]$, $g(\rho) \in [(1-2\nu_1)/(1+\eta), 1-\nu_0/2]$, so $\mathrm{Alg}(\mathcal{D}_N, \rho, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfies

$$\frac{1-2\nu_1}{1+\eta} \leqslant g(\rho) \leqslant \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant (1+\eta) g(\rho) \leqslant (1+\eta)\left(1 - \frac{\nu_0}{2}\right) \ .$$

In particular, after $\log(\rho_0/\delta)/\log(2)$ steps, $\mathrm{Alg}(\mathcal{D}_N, \rho, \eta, u, \mathbf{x}) = (\mathbf{M}, \mathbf{y})$ satisfies

$$1 - \nu \leqslant \frac{1-2\nu_0}{1+\eta} \leqslant \mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \leqslant (1+\eta)\left(1 - \frac{\nu_0}{2}\right) \leqslant 1 \ .$$

Hence, the algorithm outputs in $T = \log(\rho_0/\delta)/\log(2)$ steps a solution satisfying the conclusions of Lemma 123. □

Let now $\mathbf{x}_0 = \widehat{\mu}^{(0)}$ defined in Lemma 108. This lemma grants that, with probability larger than $1 - e^{-K/128}$, Ω_0 holds, where $\Omega_0 = \{\|\widehat{\mu}^{(0)} - \mu_P\| \leqslant 12\sqrt{\mathrm{Tr}(\Sigma)K/N}\}$. Consider the event $\Omega = \Omega_0 \cap \Omega_\alpha$. We consider two cases. On Ω, either

$$\|\widehat{\mu}^{(0)} - \mu_P\| \leqslant 2r \quad \text{or} \quad \|\widehat{\mu}^{(0)} - \mu_P\| > 2r \ .$$

By Lemma 112, on Ω,

$$\mathrm{OPT}(\widehat{\mu}^{(0)}) \geqslant \frac{1-2\alpha}{1-\alpha}(\|\widehat{\mu}^{(0)} - \mu_P\| - r)^2 \ .$$

If $\|\widehat{\mu}^{(0)} - \mu_P\| > 2r$, it follows that

$$\mathrm{OPT}(\widehat{\mu}^{(0)}) \geqslant \frac{1-2\alpha}{1-\alpha} r^2 \ .$$

By (9.9), if $\|\widehat{\mu}^{(0)} - \mu_P\| > 2r$, therefore

$$g(\rho) \leqslant \frac{(1-\alpha)}{(1-2\alpha)\rho r^2} + b \ .$$

Hence, $g(\rho) < 1/(1+\eta)$ if $\rho_0 = C_\alpha / r^2 \leqslant C_\alpha N$.

9.5.6 Final algorithm

1. Compute $\widehat{\mu}^{(0)}$, fix $R = \log N$, $\tau = \log[144K]/\log[3/4]$, $\epsilon \in (2\alpha/(1-\alpha), 0.4(1-\alpha)/(1-2\alpha))$, $\nu = \epsilon - b$, $B = 0$.
2. While $t \leqslant \tau$ and $B = 0$,
   (a) Run the algorithm of Lemma 123 with $\rho_0 = N$ and output $(\mathbf{M}, \mathbf{y})$ and ρ^*.
   (b) If $\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \notin [1-\nu, 1]$, then $\widehat{\mu}^{(\kappa+1)} = \widehat{\mu}^{(t)}$, $B = 1$.
   (c) If $\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \in [1-\nu, 1]$, then apply the algorithm of Lemma 117 to output $(x, \mathbf{y}', \mathbf{M}')$ satisfying, for any $\mathbf{x}$ such that $\|\mathbf{x} - \mu_P\| > A_\alpha r$,
   $$h_{\mathbf{x}}(\mathbf{M}') \geqslant (\beta\|\mathbf{x} - \mu_P\| + r)^2 \ .$$
   (d) Update $\widehat{\mu}^{(t+1)}$ according to (9.5), with $\mathbf{x} = \widehat{\mu}^{(t)}$ and $\widehat{\mathbf{M}}_{\mathbf{x}} = \mathbf{M}'$.
3. Output $\widehat{\mu} = \widehat{\mu}^{(\tau+1)}$.

The algorithm terminates either after τ operations or when $\widehat{\mu}^{(t)}$ satisfies $\mathrm{Tr}(\mathbf{M}) + \|\mathbf{y}\|_1 \notin [1-\nu, 1]$. Using a union bound in $t \in \{1, \dots, \tau\}$, this last situation either happen if $\|\widehat{\mu}^{(t)} - \mu_P\| \leqslant 2r$ or on an event of probability at least $1 - e^{-u}$. If the algorithm runs τ steps without stopping, on $\Omega = \Omega_0 \cap \Omega_\alpha$, the output satisfies, by Proposition 115

$$\|\widehat{\mu} - \mu_P\|^2 \leqslant (3/4)^\tau \left(144 \frac{\mathrm{Tr}(\Sigma) K}{N}\right) + (A_\alpha^2 + 1) r^2 \sum_{i=0}^{+\infty} \left(\frac{3}{4}\right)^i \leqslant C_\alpha r^2 \ .$$

Overall, choosing for example $\alpha = 1/10$, we have obtained the following result.

Theorem 124. *There exists a numerical constant C and an algorithm that runs in $\tilde{O}(uK + Kd)$ operations and outputs an estimator $\widehat{\mu}$ of μ_P such that*

$$\mathbb{P}\left(\|\widehat{\mu} - \mu_P\| \leqslant C \left(\sqrt{\frac{Tr(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} K}{N}} \right) \right) \geqslant 1 - e^{-u \wedge (K/C)} \ .$$

Bibliography

[1] Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Ann. Statist.*, 47(4):2117–2144, 2019.

[2] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[3] Y. Baraud and L. Birgé. Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.*, 126(12):3888–3912, 2016.

[4] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017.

[5] Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[6] Yannick Baraud and Lucien Birgé. Rho-estimators revisited: general theory and applications. *Ann. Statist.*, 46(6B):3767–3804, 2018.

[7] Lucien Birgé. Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.*, 20(3):201–223, 1984.

[8] Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[9] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[11] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[12] O. Catoni and I. Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. Technical report, 2017. https://arxiv.org/pdf/1712.02747.pdf.

[13] Olivier Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[14] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[15] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 2018.

[16] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '19, pages 2755–2771, Philadelphia, PA, USA, 2019. Society for Industrial and Applied Mathematics.

[17] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806, Phoenix, USA, 25–28 Jun 2019. PMLR.

[18] G. Chinot, G. Lecué, and M. Lerasle. Statistical learning with lipschitz and convex loss functions. *to appear in Probab. Theory Related Fields, arXiv:1810.01090*, 2019.

[19] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *1906.03058*, 2019.

[20] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016.

[21] Magalie Fromont, Matthieu Lerasle, and Patricia Reynaud-Bouret. Family-wise separation rates for multiple testing. *Ann. Statist.*, 44(6):2533–2563, 2016.

[22] Frank R Hampel. Contribution to the theory of robust estimation. *Ph. D. Thesis, University of California, Berkeley*, 1968.

[23] Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

[24] Frank R Hampel. Robust estimation: A condensed partial survey. *Probability Theory and Related Fields*, 27(2):87–104, 1973.

[25] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[26] Frank R Hampel. Beyond location parameters: Robust concepts and methods. *Bulletin of the International statistical Institute*, 46(1):375–382, 1975.

[27] Samuel B. Hopkins. Fast mean estimation with sub-gaussian rates. *to appear in Ann. Statist., arXiv:1809.07425*, 2018.

[28] Peter J Huber. The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA, 1967.

[29] Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[30] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[31] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[32] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008, 2015.

[33] G. Lecué and M. Lerasle. Learning from mom's principle : Le cam's approach. Technical report, CNRS, ENSAE, Paris-sud. To appear in Stoch. Proc. App.

[34] G. Lecué and M. Lerasle. Robust machine learning by median-of-means : theory and practice. *to appear in Ann. Statist., arXiv:1711.10306*, 2019.

[35] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

[36] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[37] Z. Lei, K Luh, P Venkat, and F. Zhang. A fast spectral algorithm for mean estimation withsub-gaussian rates. *Arxive:1908.04468*, 2019.

[38] G. Lugosi and S. Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *1907.11391*, 2019.

[39] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in JEMS*.

[40] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, Aug 2019.

[41] Gábor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106, 2019.

[42] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 2019.

[43] Pascal Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. Probability theory.

[44] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[45] Shahar Mendelson. Learning without concentration. In *Proceedings of the 27th annual conference on Learning Theory COLT14*, pages pp 25–39. 2014.

[46] S. Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. *Preprint available on arXiv:1704.02658*.

[47] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[48] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[49] Richard Peng and Kanat Tangwongsan. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. In *Proceedings of the Twenty-fourth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '12, pages 101–108, New York, NY, USA, 2012. ACM.

[50] A. Prasad, A.S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *Arxive:1802.06485*, 2018.

[51] Adrien Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018.

[52] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[53] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[54] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[55] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[56] J.W. Tukey. Adress to international congress of mathematicians. Technical report, Vancouver, 1974.

[57] J.W. Tukey. T6: Order statistics. Technical report, In mimeographed notes for Statistics 411, Princeton Univ., 1974.

[58] Vladimir N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.