Statistical and Computational Complexities of Robust and High-Dimensional Estimation Problems

Jules Depersin

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAG009

Thèse de doctorat

ENSAE

IP PARIS

Statistical and Computational Complexities of Robust and High-Dimensional Estimation Problems

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques fondamentales

Thèse présentée et soutenue à Palaiseau, le 20 décembre 2021, par

JULES DEPERSIN

Composition du Jury :

Arnak Dalalyan Professeur, Centre de recherche en économie et statistique	Président
Peter Bartlett Professeur, University of California, Berkeley	Rapporteur
Gabor Lugosi Professeur, Universitat Pompeu Fabra	Rapporteur
Claire Brecheteau Maître de Conférences, Université Rennes 2	Examinateur
Stanislav Minsker Professeur, University of Southern California	Examinateur
Guillaume Lecué Professeur, Centre de recherche en économie et statistique	Directeur de thèse

Remerciements

Il y a tellement de personnes qui m'ont aidé et fait grandir pendant cette thèse que c'est impossible de les remercier assez avec quelques mots. Cette longue liste de merci n'est donc ni suffisante, ni exhaustive.

Je tiens à remercier Guillaume, qui m'a pris sous son aile depuis le début et m'a permis de découvrir le monde de la recherche avec sa sagesse, son humour, son humilité et son optimisme inébranlable. J'ai passé trois très belles années à apprendre avec toi. Je garderai précieusement avec moi les fruits de nos échanges.

Je remercie tous les membres du jury qui ont bien voulu me faire l'honneur de lire et de commenter ce travail. Un grand merci à Gabor Lugosi et à Peter Bartlett, qui ont été deux immenses sources d'inspirations pour ce travail et qui ont tous les deux acceptés d'être rapporteurs, ce qui me rend très fier. Cette thèse leur doit énormément, ainsi qu'au travail de Stanislav Minsker. Arnak, ton calme, ta patience, et la façon dont tu parles d'apprentissage statistique et dont tu l'enseignes ont déjà et continueront à marquer mon parcours scientifique. Merci d'avoir accepté de lire cette thèse. Enfin merci beaucoup à Claire Brecheteau d'avoir bien voulu prendre le temps de lire ce travail et de le commenter à la lumière de son expertise.

Je remercie tous les chercheurs du CREST qui sont autant d'exemples de professionnalisme et de rigueur. En particulier j'ai eu la chance d'aider Vianney et Pierre à enseigner, et donc de profiter de la profondeur de leur compréhension et de leurs intuitions. Matthieu a énormément contribué à cette thèse et m'a aidé dans tout les aspects du travail scientifique. Je ne pourrais pas te remercier assez pour tous tes conseils, toute ton aide, et toutes nos discussions. Tous ceux que j'ai moins côtoyés au quotidien m'ont pourtant souvent apporté, par une remarque ou une question, une aide précieuse : Jaouad Mourtada, Nicolas Chopin, Alexandre Tsybakov, Francis Kramarz, Victor-Emmanuel, entre autres.

Je remercie mes collègues de la direction des études, qui m'ont permis de relativiser les galères partagées et de se sortir tant bien que mal de ce Covid : Laurent et Frédéric, merci pour votre gentillesse, votre disponibilité, et vos conseils différents mais utiles. Rosalinda, Karine, Christine travaillent beaucoup, parfois dans l'ombre, pour que l'école fonctionne, je profite de l'occasion pour les remercier au nom de tous. Merci aussi à la famille des coordinateurs; Morgane, Jérome, Arthur, Lucie, Wissal.

Je remercie tous les doctorants que j'ai pu croiser pour nos échanges : Solène, Flore, Julien, Badr dont le sourire a été un rayon de soleil, Geoffrey qui m'a beaucoup appris, Amir, Hai-Dang, Avo, Suzanne, Lionel, Gabriel, François-Pierre, Meyer, Guido, Léa, Héloise, Inès, Cécile, Tang, Germain, Jeremy et Christophe entre autres.

Je profite de l'occasion pour redire à ma famille la source de bonheur et d'amour qu'ils sont

pour moi. Alice, Lucas, Papa, Maman, je vous aime. Mes grands-parents, mes modèles, cette thèse vous est dédiée. Adèle, merci pour ta patience, ta douceur et tout le reste pendant toutes ces années.

Je finis par mentionner ces quelques personnes rencontrées sur le chemin et qu'à la fin de ma thèse je suis heureux de pouvoir appeler mes amis : c'était tellement facile de commencer la journée avec la perspective de vous retrouver et de partager un potage au Magnan. Merci Guillaume Hollard de m'avoir fait rire et d'avoir partagé avec nous ta grande experience de la vie. Yannick et son petit sourire coquin, Gwen notre roc de certitude, Elia toujours là pour écouter ce qu'on a sur le coeur, Etienne et Remi mes frères avec qui j'ai l'impression d'avoir tout vécu, Remi qui a su me redonner le sourire dans les moments plus durs, qui sait modéliser économiquement les difficultés de la vie, Nicolas Schreuder, avec lequel j'ai tant partagé sur la recherche et sur la vie, et Fabien que j'ai plus vu que mon reflet pendant 3 ans, mon partenaire de musculation et de StatApp, son sourire, ses bonnes questions et ses passions vont beaucoup me manquer. Enfin merci à Bérengère pour toutes les aventures qu'on a vécues ensemble, et qui j'espère ne font que commencer.

A mes grands-parents,

Contents

CHAPTER 1

Introduction: Robustness and Complexity

This thesis tries to assess the complexity of some robust statistical tasks. In the introduction, the meaning and the use of this overall goal will be analysed and hopefully precised. First the two main terms at stake, robustness and complexity, will be successively discussed. Then the five works that make up this thesis will be presented, situated with respect to the general context, and their respective contributions and limitations will be exposed.

Contents

1.1 What is Robust Statistics?

The main goal of a statistician is to extract useful information from observational data, for instance inferring patterns, identifying causal effects, in a word learning about a given phenomenon from the data. Most of the time the statistician has to make some assumptions about the given data in order to get some guarantees that her procedures will indeed lead to trustworthy information. The most common assumptions are :

- that the data are independant and identically distributed, meaning that they are produced from the same process, and independently from one another, they are different realisations of the same *random variable*,
- this *random variable* is assumed by the statistician to have some given property (for instance a finite expected value),
- that the distribution of the random variable from which the observations are supposed to be drawn belongs to a specific family of distributions determined beforehand by the statistician (for instance, the family of normal distributions) and called the statistical model. When making this kind of assumption, the statistics at stake are called *parametric* because in this framework, the different distributions of the given family are often labeled with a parameter (and can by entirely described with this parameter), so searching for the right distribution among the family boils down to searching for the right parameter.

We note that these assumptions are increasingly strong. One could say that the overall goal of Robust Statistics as a field is to question these assumptions, to study if it is possible to give some guarantees under weaker and more realistic assumptions. To be more precise, the field of robust statistics tries to study

- what happens to classical non-robust procedures when the assumptions under which they were created are loosened,
- what are the minimal assumptions one has to make on the data so that it is theoretically possible to retrieve some information from those,
- and how to find procedures that still hold when making minimal assumptions.

Hampel et al. (1986) gives the following summary: “robust statistics is a body of knowledge relating to deviations from idealized assumptions in statistics.”

1.1.1 Why Robust Statistics?

One might ask what is the use of such a theory. It has been largely justified by a rich literature pioneered in the sixties by Tukey (1960), Huber (1964) and Hampel (1973), and numerous grounds for such a theory are exposed with great details in a number of books (see for instance Huber (1981), Hampel et al. (1986), Huber and Ronchetti (2009), or Maronna et al. (2006a)). We recall and illustrate two of the main arguments they develop, which can serve as starting points to understand the contributions of this thesis.

Gross error. That is a point made by both Huber (1981) and Hampel et al. (1986): the samples collected in physical, natural or social science contain "good data", that are well described by the model of the statistician, but they are mixed with a fraction of "bad data", also called gross error or blunders - typically in the range between 1% and 10% according to Huber (1981) for dataset of that time (that might have changed in magnitude with internet data and declarative data). These errors can come from mistakes in copying, in computation, inattention of the experimenter, and so on, they could also come from an adversary trying to lure the statistican. These bad data or "outliers" are not well described by the statistician model, and can be orders of magnitude away from the phenomenon which she wants to quantify.

For instance, a part of the data can be expressed in the wrong unit: when asked in on-line surveys about their monthly salary, a fraction of the survey respondents rather give their annual salary. Such mistakes can be very costly: one can think about the failed launch of the Mars Climate Orbiter in 1999 by NASA that was caused by some key data being expressed in non-metric units. Classical estimators are very sensitive to this kind of gross error: for instance, consider a series of observations of the height of people, with 99 observations expressed in meters and 1 observation mistakenly expressed in centimeter. The empirical mean of this series taken carelessly could lead to conclude that the average human size is around 3.3 meters high, showing that it only takes a small fraction of bad data to get a very wrong idea about the phenomenon at stake when using non-robust estimators.

Gross error are a violation of the first classical assumption that all data are produced from the same process. When there are gross errors, *all observations do not have the same informational value*. Most of classical statistical procedures only present guarantees when the data are i.i.d. (independant and identically distributed), thus the presence of blunders is a first argument to justify the need to develop new robust procedures.

Heavy-tail data. Another argument can be found to give grounds for the need of a robust theory: the presence of heavy-tail phenomena. Heavy tails are characteristic of phenomena where there is a significant probability of observing a single huge value, that is order of magnitudes away from other. In contrast to gross error, this huge value is not a mistake and contains information about the phenomenon at stake. Insurance losses, financial returns, social contagion, the retweet activity of a tweet are all examples of heavy-tailed phenomena, see Resnick (2007) for more examples and detailed explanations. The following example illustrates the problem raised by heavy-tailed data. Take a random variable X that is equal to 0 with probability 999/1000, and 1000 with probability 1/1000, so that its mean value is 1 (and its standard deviation around 32). When given 10 observations of this random variable, one has about a 1% chance of seeing the value 1000 which is an extra-ordinary event. When observing such an event among the 10 data, the statistician desiring to estimate the mean value of the phenomenon faces a dilemma: should she take into account this observation? Including this observation, the empirical mean of the observations is 100, which is a few standard deviation away from the true mean. Excluding this observation before taking the empirical mean allows to get a better estimate of the mean in this case, but on what ground should the statistician discard some values? Robust theory tries to answer such questions.

Even if this heavy-tail argument has mainly been explored in the robust literature for the last decade following the pioneer work of Catoni (2012), it is already mentioned in Huber (1981). The presence of heavy-tail data breaks the assumption, very common in statistics, that the observations are drawn from an underlying Gaussian (or sub-Gaussian) distribution; that is a distribution with good concentration properties.

What about outlier rejection? From the examples presented above, one could be under the impression that it is enough to pre-process the data to remove outliers before applying standard procedures, rather than finding new robust procedures. The example of heavy-tail data shows that it is not always simple to tell when an observation is an outlier, and it becomes even more complicated when the observations are multi-dimensional, as we will see throughout this thesis. While in some cases outlier rejection might be a good idea, it is often complicated to implement (see Diakonikolas et al. (2019b) for instance) and providing guarantees for such procedures requires the theoretical tools developed by robust statistics.

1.1.2 Models for gross errors

In order to give guarantees about their procedures, statisticians working on robust statistics have to make some assumptions about the data, even if these are weaker and more realistic than the classical ones. In this part we present the main models adopted by robust statistics and precise the framework that will be used throughout this thesis.

Huber's contamination model. The first generation of statisticians working on robust statistics proposed to take into account gross errors with the following model, which is sometimes called *Huber's contamination model* (presented and studied for instance in Huber (1981)): the observations are supposed to be i.i.d. realizations drawn from a random variable with cumulative distribution function

$$F = (1-\epsilon)F_\theta + \epsilon H,$$

where F_θ belongs to a parametric family known in advance by the statistician (for instance the Gaussian family) but where H can be *any* cumulative distribution function, and where ϵ represents the contamination rate, the assumed fraction of gross error (for which an upper bound is often assumed to be known). Most of the work of early robust statistics was concerned with finding estimators that were efficient in this framework, that is somehow close to a parametric framework. This model is still an active research area, see for instance Chen et al. (2017). We note that a lot of work dealing with this first model focus on providing asymptotic results, together with non-asymptotic properties such as the breakdown point. In contrast, the point of view adopted in this thesis is solely *non-asymptotic*, inspired by recent trends in robust statistics initiated by Catoni (2012) and by works from the computer science community such as Diakonikolas et al. (2016).

Adversarial contamination model. The model that we deal with in this work is more general than Huber's contamination model and is sometimes referred to as "adversarial contamination". It is difficult to trace back, but it seems to have been described and popularised by the computer science community, for instance in Diakonikolas et al. (2016). The samples are generated from the following process: First, N samples are drawn independently from some unknown distribution. Then, an adversary is allowed to look at the samples and arbitrarily corrupt an ϵ-fraction of them before turning the corrupted data to the statistician. The setting can be described more formally as follows :

Setting 1.1. *There exists N i.i.d random variables distributed like X denoted $(\tilde{X}_i)_{i=1}^N$ in $\mathbb{R}$ which are independent. These variables are not directly observed by the statistician, they are first given to an "adversary" who is allowed to modify up to $\lfloor \epsilon N \rfloor$ of these variables before returning a modified dataset $(X_i)_{i=1}^N$ to the statistician.*

The only information that the statistician have is that there exists a (possibly random, possibly data-dependant) set $\mathcal{O}$ such that, for any $i \in \mathcal{O}^c$, $X_i = \tilde{X}_i$. The only assumption on the set $\mathcal{O}$

concerns its size: the statistician knows that $|\mathcal{O}| \leq \lfloor \varepsilon N \rfloor$ *(the statistician may in some cases not even know* ε *and adapt to it). The statistician does not know which data has been modified, so the set* $\mathcal{O}^c$ *is unknown. The statistician tries to acquire some knowledge (such as the mean or the variance) about the random variable* X *from the corrupted dataset* $(X_i)_{i=1}^N$.

In this setup, not only can outliers be correlated to each other and to inliers, but inliers can also be correlated to one another (because the adversary can choose which original samples to keep and in doing so correlating the samples that he keeps, for instance only keeping the largest samples when they are real-valued), which can not be the case in the Huber's contamination model.

This model generalizes Huber's one, and it has been used, as Huber's one, to deals with deviations from a *parametric* model, for instance in Du et al. (2017), Diakonikolas et al. (2019c) or Cheng et al. (2019b), where the inliers are supposed to follow a Gaussian distribution. In contrast this thesis, along with a modern line of work in robust learning opened by Catoni (2012), deals with *non-parametric statistics* (we often try to estimate the mean of a sample distribution), with a particular emphasis put on heavy-tail data, as will be explained below. The adversarial contamination model is used here in a non-parametric way, the general form of the distribution that the inliers follow is *not* assumed to be known in advance.

1.1.3 Dealing with heavy-tail data

The two models described above (Huber and adversarial contaminations) allow to deal with gross errors, but they do not directly address the problems raised by heavy-tail phenomena. Let us get back to our example, where a statistician wants to estimate the mean value of the random variable X that equals either 0 (with high probability) or 1000 (with small probability), when given $N = 10$ observations. If she uses the usual empirical mean, which is the standard tool in classical statistics to estimate a mean, she will be "close" to the true value with probability 99% and drastically wrong ($\sim 3\sigma$) with probability 1%, a small but not negligible probability. Catoni (2012) raises the following questions: when given a "precision radius" r, what is the smallest *failure probability* $\delta(r)$ such that it is possible to find an estimator that lies with probability $1 - \delta(r)$ within a radius r of the true mean value of the random variable? In our example, for a radius $\sim \sigma$ for instance, is it possible to get a better failure probability than 1%? What estimation procedures can lead to such an estimator? And what are the minimal assumptions one has to make on the distribution of the random variable?

What assumptions on the underlying distribution? To capture the heavy-tail phenomenon, we want to obtain statistical properties *without making either boundedness or gaussian assumptions on the data* (or any other strong concentration assumptions), and it is in this sense that we will call our estimators robust to heavy-tails. What weaker assumption should we make on the underlying distribution of the data, that would be weak enough not to limit severely the applicability of the results, and strong enough to lead to interesting and significant results? This question is investigated in details in Devroye et al. (2016): authors show that, in order to reach a rate that resembles the one asymptotically reached with the central limit theorem, rate which can be proven to be "essentially optimal" and that will be discussed in greater details below, the minimal assumption to make on the data is the existence of a finite second moment: $\mathbb{E}(X^2) < \infty$. Like a large majority of the robust literature that deals with heavy-tail data, the better part of this thesis complies with this analysis (Chapter 3, 4 and 5)

Setting 1.2. *The random variable* X *from which are sampled the* N *variables* $(\tilde{X}_i)_{i=1}^N$ *in* $\mathbb{R}$ *has a mean* μ*, which we try to estimate, and a (possibly unknown) second order moment* $\sigma^2 = \mathbb{E}[(X - \mu)^2]$.

Most of the work presented deals with the case where the underlying distribution has a -sometimes known, sometimes unknown- second-order moment. We however note that some recent papers investigate cases where the second moment does not even exists and is replaced by moments of order $1+\alpha$ with $\alpha < 1$ for instance Cherapanamjeri et al. (2020b). The end of this thesis also presents new settings that does not require the existence of a finite second moment (in parts of Chapter 6 and 7), using as inspiration other asymptotic results than the Central Limit Theorem.

The empirical mean. Let us come back to our example, and to the insufficiency of the empirical mean. If we are given N independent realizations of a random variable with mean μ and variance σ^2, Chebyshev's inequality tells that the empirical mean $\hat{\mu}$ satisfies, with probability greater than $1-\delta$, $|\hat{\mu}-\mu| \leq \sigma/\sqrt{N\delta}$. Catoni (2012) states that this rate is sharp for the empirical mean in general, so one can not give a better inequality for the empirical mean that holds for all distributions with a variance. So for the empirical mean $r(\delta) = \sigma/\sqrt{N\delta}$, or equivalently $\delta(r) = \sigma^2/(Nr^2)$. This bound rapidly deteriorates for small values of δ: if one wants to get high-probability results, for instance with $\delta \sim 10^{-5}$, with about a thousand data points, one gets a radius of $\simeq 10\ \sigma$. Is it possible to do better, can one find procedures that gives smaller radius? This question was raised and answered in Catoni (2012).

Sub-Gaussian estimation. Catoni (2012) states that the rate attained by the empirical mean can indeed be drastically improved. The rate he proposes is inspired by asymptotic theory. Indeed, when the data have a finite second moment σ, the central limit theorem guarantees the empirical mean has Gaussian tails, asymptotically, without making further assumptions on the data, so, when $N \to \infty$,

$$\mathbb{P}\left(|\hat{\mu}-\mu| < \frac{\sigma\phi^{-1}(1-\delta/2)}{\sqrt{N}}\right) \to \delta, \tag{1.1}$$

where ϕ is the cumulative distribution function of the standard normal distribution. Catoni (2012) also proves that this asymptotic rate cannot be improved: no non-asymptotic estimator can achieves better-than-Gaussian tails for all distributions in a class that is "large enough" (that contains at least all Gaussian distributions with a given variance σ^2). The main finding from Catoni (2012) is that it is possible to find non-asymptotic estimators reaching the rate (1.1), up to universal multiplicative constants.

Note that the rate $\sigma\phi^{-1}(1-\delta/2)\sqrt{N}$ obtained in (1.1) is the rate achieved by the empirical mean when the data is a N-sample of i.i.d. Gaussian variables. We thus say that an estimator *achieves a subgaussian rate* if it achieves the rate (1.1) up to multiplicative constants. In a sense, we want our estimators to be as good as if the data were Gaussian, *even when the real sample is heavy tailed* (it is only assumed to have a second moment).

Remark 1.1. *We know that* $\phi^{-1}(1-\delta/2) \leq \sqrt{2\ln(2/\delta)}$*, and for small* δ*,* $\phi^{-1}(1-\delta/2) \sim \sqrt{2\ln(2/\delta)}$*. As our results are all formulated up to multiplicative constants*[1]*, we use the explicit formulation* $C\sqrt{\ln(1/\delta)}$ *in most of our results.*

Coming back to the example, the rate (1.1) with $\delta \sim 10^{-5}$ and $N \sim 1000$ gives a radius of around 0.15σ, and the crucial logarithmic dependence allows it to deteriorate drastically slower than the rate obtained by Chebychev's inequality when δ goes to 0.

[1] In this thesis, we have not tried to optimize the constants, even though it is an important and interesting problem, see Section 1.5.

Catoni (2012) builds a subgaussian estimator $\hat{\theta}$ implicitly, as a solution of the following equation:

$$\sum_i \psi\left[\alpha(X_i - \hat{\theta})\right] = 0,$$

where α is a carefully chosen positive real, and ψ is a bounded influence function such that $\psi(x) \sim x$ when x is close to 0. In contrast, this thesis mainly studies an other way to build such subgaussian estimator, the Median-Of-Mean (MOM) heuristic.

The Median-Of-Mean heuristic. This approach which was first described in Nemirovsky and Yudin (1983) and Jerrum et al. (1986), and has received a lot of attention in the statistical and machine learning communities in the last decade, for instance in Bubeck et al. (2013), Lerasle and Oliveira (2011), Devroye et al. (2016), Minsker and Strawn (2017) or Minsker (2015). This approach was originally designed to build estimators that are robust to heavy-tail data in various settings (see Alon et al. (1999), Jerrum et al. (1986) and Birgé (1984) for instance). It can be defined as follows: we first randomly split the data into K blocks $B_1, \dots, B_K$ of equal-size $m = \lfloor N/K \rfloor$ (if K does not divide N, we just remove some data), K being chosen of order $\log(1/\delta)$ with δ the failure probability desired by the statistician. We then compute the empirical mean within each block: for $k = 1, \dots, K$,

$$\bar{X}_k = \frac{1}{m} \sum_{i \in B_k} X_i.$$

The final estimator $\hat{\mu}_K$ is the median of the latter K empirical means. The first paper that formally proves that this estimator has sub-Gaussian tails is Devroye et al. (2016).

Theorem 1.1 (Devroye et al. (2016))**.** *The estimator $\hat{\mu}_K$ has subgaussian deviations: with probability $\geq 1 - e^{-K/32}$,*

$$|\hat{\mu}_K - \mu| \leq \sqrt{8\frac{\sigma^2 K}{N}}$$

The main idea lies in the switch from an unbounded variable to a bounded one thanks to the median operator. Indeed, to know whether the median is within an interval I, we compute $Z := \sum_{k=1}^{K} \mathbf{1}_{\bar{X}_k \in I}$: when this quantity is greater than $K/2$, the median lies in I. The quantity Z, unlike the empirical mean, is a sum of bounded variable, thus the Hoeffding inequality (see Hoeffding (1963)) states that it is close to its mean with exponentially low failure probability, leading to the subgaussian rate in Theorem 1.1.

Note that this procedure does not hold simultaneously for all δ: one has to compute a different estimator for each value of δ. This issue can be overcome using Lepskii adaptation method presented in Lepskiĭ (1990), as explained in greater details in Chapter 3.

This technique allows to fully handle the one-dimensional case, and to deal with both heavy-tail data and adversarial contamination. However, the rapid development of machine learning and the growing amount of available high-dimensional data has led many statisticians to focus on high-dimensional tasks. In these settings, the given data are not observations of a one-dimensional random variable, but rather observations of a d-dimensional random vector, with $d \gg 1$. The field of robust statistics has taken up this issue, raising new questions: what rate can we reach in this case? What role does the dimension play? How to change and adapt the procedures for this setting? We note that the extension of the one dimensional result is not trivial since there exist several possible generalizations of the median in multi-dimensional set-ups (for instance the geometric median, see Minsker (2015) for a definition, or the coordinate-wise median).

1.2 High-dimensional Robustness

1.2.1 The subgaussian rate in high dimension: decoupling complexity and deviation.

What rates is it possible to reach in high-dimensional setups? One can still try to reach a rate inspired by the asymptotic normality of the empirical mean of observations with finite second moment, which is the rate achieved by the empirical mean when the data is a N-sample of i.i.d. Gaussian variable, that we will still call *subgaussian rate*. We compute this rate using Borell-TIS's inequality (see (Ledoux, 2001, Theorem 7.1)): if $Z_1, Z_2, ..., Z_N$ are independent identically distributed Gaussian variables $\mathcal{N}(\mu, \mathrm{Id})$, it follows from this inequality that with probability at least $1-\delta$,

$$\left\|\bar{Z}_N - \mu\right\|_2 = \sup_{\|v\|_2\leq 1} \langle \bar{Z}_N - \mu, v\rangle \leq \mathbb{E}\sup_{\|v\|_2\leq 1} \langle \bar{Z}_N - \mu, v\rangle + \gamma\sqrt{2\log(1/\delta)},$$

where $\gamma = \sup_{\|v\|_2\leq 1}\sqrt{\mathbb{E}\langle \bar{Z}_N - \mu, v\rangle^2}$. It is elementary knowledge about multivariate Gaussian distributions that $\mathbb{E}\sup_{\|v\|_2\leq 1}\langle \bar{Z}_N - \mu, v\rangle \leq \sqrt{d/N}$ and $\gamma = \sqrt{1/N}$, which leads to the following subgaussian rate (where C is an absolute constant),

$$\left\|\bar{Z}_N - \mu\right\|_2 \leq \left(\sqrt{\frac{d}{N}} + \sqrt{\frac{2\log(1/\delta)}{N}}\right) := Cr_\delta. \tag{1.2}$$

Whether it was possible to reach such a rate with heavy-tail and corrupted data was an open problem during a few years. The first attempts to adapt the Median-Of-Mean heuristic, using for instance the geometric median (also called Fermat point) instead of the one-dimensional median, and presented in Minsker (2015) or in Hsu and Sabato (2016), led to estimators $\hat{\mu}$ achieving with probability larger than $1-\delta$,

$$\|\hat{\mu} - \mu\|^2 \leq \frac{d\log(1/\delta)}{N}, \tag{1.3}$$

where $\|\cdot\|$ is the canonical Euclidean norm on $\mathbb{R}^d$. This bound is proportional to $\log(1/\delta)$, thus the estimators are somehow robust to heavy tails, but they do not quite achieve the subgaussian rate. Indeed, in the subgaussian rate, the complexity term d, which captures how involved is the ambient space, and the failure-dependant factor $\log(1/\delta)$ are not multiplied, but added instead, and are thus in a way *decoupled* from each other.

After a few years, the seminal paper of Lugosi and Mendelson (2019c) described the first estimator to reach the subgaussian rate only assuming finite second moment, using the Median-Of-Mean heuristic coupled with a tournament procedure. An idea is to make a clever use of a generalization of the Hoeffding inequality used in Theorem 1.1, called the *Bounded Difference inequality* (also called McDiarmid or Hoeffding/Azuma inequality), that we recall here (see for instance Theorem 6.2 in Boucheron et al. (2013)):

Theorem 1.2 (McDiarmid's inequality)**.** *Consider independent random variables $X_1, \cdots, X_n \in E$ and a mapping $\psi : E^n \mapsto \mathbb{R}$. If for all $i \in [\![1, n]\!]$ and for all $x_1, \cdots, x_n, x_i'$*

$$|\psi(x_1, \cdots, x_i, \cdots, x_n) - \psi(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i \ ,$$

then for every $t > 0$,

$$\mathbb{P}(|\psi(X_1, \cdots, X_n) - \mathbb{E}[\psi(X_1, \cdots, X_n)]| \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

This inequality can be found in McDiarmid (1997) and it allows to deal with functions ϕ more general than the simple sum of random variables. In the problem at stake, we use this inequality to bound the high-dimensional equivalent of the quantity Z defined above,

$$\tilde{Z} := \sup_{v \in \mathcal{S}_2^{d-1}} \left(\sum_{k \in [K]} \mathbf{1}(|\langle \overline{X}_k - \mu, v \rangle| > r) \right), \tag{1.4}$$

which is this time a supremum over all possible d-dimensional direction, but remains a supremum of bounded quantity. Theorem 1.2 states that $\tilde{Z}$ is exponentially likely to be close to its expectation. The remaining steps of the proof, that are detailed in Chapter 3, are about bounding this expectation $\mathbb{E}(\tilde{Z})$, using tools from empirical processes theory.

This pioneer paper, while answering an open question, opens two main research directions that are at the heart of this thesis, and that both deals with notions of complexity. The first one is about *computational complexity*: to compute the estimator described in Lugosi and Mendelson (2019c), one needs a number of steps that grows exponentially with the dimension, so this estimator cannot be computed in practice, even for moderate values of d. Thus our first question is the following one: *can we find tractable subgaussian estimators in high-dimensions?* Chapter 3 and 4 mainly deal with this question. The second one is about statistical complexity in broad sense and asks *whether it is always possible to find procedures that reach a Gaussian rate* for other estimation tasks, such as estimation with respect to non-euclidean norms for instance. Chapter 5, 6 and 7 follow that second direction.

1.2.2 Computational complexity

The first kind of complexity this thesis explores is a computational one. Diakonikolas et al. (2016) is one of the first papers to raise the question of tractability for robust mean estimation in high dimension, and to show that high-dimensional robust learning is *algorithmically* possible. Before this pioneer paper, the computational considerations were mainly hardness results, showing that traditional robust estimators such as the Tukey median, although provably robust, are provably hard to compute (see for instance Johnson and Preparata (1978) or Bernholt (2006)). Using new techniques, Diakonikolas et al. (2016) opens the way to computational, tractable robust statistics and leaves many exciting questions to explore. Indeed, even though estimators proposed in Diakonikolas et al. (2016) are robust to adversarial contamination, they fail with constant probability (for instance 1/100), and do not achieve the aforementioned *subgaussian rate*. A similar observation can be made about the works of Minsker (2015) or Hsu and Sabato (2016). In these two papers, the estimators are very fast to compute: the geometric median of mean can be computed as fast as the empirical mean up to multiplicative factors logarithmic in the dimension and the number of point. However as pointed out earlier, the estimators proposed fail to achieve the subgaussian rate. A large part of this thesis focus on developing procedures that are computationally efficient and reach the subgaussian rate.

1.3 Statistical complexities

A second question that this work tackles is to know whether it is still possible to find estimators that behaves as if the data were Gaussian for other estimation tasks, or if it is only possible in a few special sub-cases as estimating the mean with respect to the usual euclidean norm.

Let us illustrate that question with an example. We now want to estimate the mean with respect to a sparse norm: we denote the set of s-sparse vectors $\mathcal{U}_s$, and consider the following

norm:

$$\|a\|_s = \sup_{u\in\mathcal{U}_s, \|u\|_2=1} \langle a, u\rangle .$$

In a sense we aim to replace the supremum over the euclidean ball $\mathcal{B}_2^d$ by a supremum over some other set, here $\mathcal{U}_s \cap \mathcal{B}_2^d$. We see, taking again the Gaussian variables Z_i, that with probability at least $1-\delta$, (still from (Ledoux, 2001, Theorem 7.1))

$$\left\|\bar{Z}_N - \mu\right\|_s = \sup_{\|v\|_2\leq 1, v\in\mathcal{U}_s} \langle \bar{Z}_N - \mu, v\rangle \leq \mathbb{E} \sup_{\|v\|_2\leq 1, v\in\mathcal{U}_s} \langle \bar{Z}_N - \mu, v\rangle + \gamma\sqrt{2\log(1/\delta)}$$
$$\leq C\left(\frac{s\log(ed/s)}{N} + \frac{\log(1/\delta)}{N}\right)^{1/2}. \tag{1.5}$$

The subgaussian rate for the sparse mean estimation problem (1.5) is different from (1.2): the "deviation term" containing the failure probability δ remains unchanged, but the "complexity term" (the one that does not depend on δ) goes from d to $s\log(ed/s)$. Can this rate be reached only assuming second order moment on the random variables? While this question is answered in this thesis for the special case of the supremum over $\mathcal{U}_s \cap \mathcal{B}_2^d$, its generalisation to other set V is still an open question, that is only partly answered in Chapter 5, 6 and 7. We can rephrase this question in another way: is the complexity term coming from the Gaussian perspective reachable for non-Gaussian variables, or is there some other complexity measure intrinsically linked to the non-Gaussian situation? We now give a first overview of the tools used to think about this question and to measure the complexity of a set V, that mainly come from empirical processes theory (see Ledoux and Talagrand (2011), Koltchinskii (2011) and van der Vaart and Wellner (1996)).

1.3.1 Gaussian complexity

The first notion of complexity that naturally arise from the comparison with the Gaussian set up used as a benchmark (always because of the central limit theorem, that is wished to be made non-asymptotic in a sense) is the *Gaussian complexity*, also called *Gaussian Mean-Width* when talking about sets $V \subset \mathbb{R}^d$. It is the complexity that appears when using the Borell-TIS inequality:

$$w^*(V) = \mathbb{E}\sup_{v\in V}\langle G, v\rangle, \tag{1.6}$$

where $G \sim \mathcal{N}(0, I_p)$. This quantity is often used in Banach space theory, see Vershynin (2018), Ledoux and Talagrand (2011), Pisier (1989) or Holmes (2012). One can think of the Gaussian mean-width as one of the basic intrinsic geometric quantities associated with sets $V \subset \mathbb{R}^p$, such as volume, surface area,... The Gaussian mean-width of various sets V is known, see for example Vershynin (2018). It is easily computable in some cases: for any subspace F of dimension k, $w^*(\mathcal{B}_2^d \cap F) = k$, so the Gaussian Mean-Width captures well the usual dimension of a subspace. We have already mentioned that for the set of sparse vector with unit euclidean norm $w^*(\mathcal{U}_s \cap \mathcal{B}_2^d) = s\log(d/s)$. It can be more involved and hard to compute in some other cases.

Even if this notion of complexity naturally appears using a Gaussian benchmark, it is for now unclear whether such a Gaussian rate can be attained only assuming second order moment on the data, aside from the case $V = \mathcal{B}_2^d$. Recent research, including works presented in this thesis, show that one can reach different but closely related rates, where the deviation term remains the same but where the complexity of the set V comes into play in another form.

1.3.2 Rademacher complexity

In the early 2000 several papers including for instance Koltchinskii (2001) and Bartlett and Mendelson (2003) proposed a new way to measure the complexity of a class of functions called the *Rademacher complexity*. This measure somehow arises from the symmetrization lemma, which is much earlier and can be traced back to Vapnik and Chervonenkis (1971).

Definition 1.1 (Rademacher complexity)**.** *Let $X_1, \cdots, X_n$ be independent random variables taking values in a measurable space $(E, \mathcal{E})$. Let $\mathcal{F}$ be a class of functions from E to $\mathbb{R}$. The Rademacher complexity of the class $\mathcal{F}$ is defined as*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right)\right] ,$$

where the variables $\sigma_1, \cdots, \sigma_n$ are i.i.d Rademacher random variables ($\mathbb{P}(\sigma_1 = 1) = \mathbb{P}(\sigma_1 = -1) = 1/2$) independent of $X_1, \cdots, X_n$. The expectation is taken with respect to both the Rademacher random variables and the data $X_1, \cdots, X_n$.

The Rademacher complexity of a class $\mathcal{F}$ quantifies the extent to which one can find, for any given Bernoulli noise sequence, a function in $\mathcal{F}$ that correlates with this particular sequence. The richer the class of functions, the more likely to find for each given noise sequence one function that correlates well with it (see Koltchinskii (2011)). In the usual learning setting, choosing classes that are "too big" and that can mimic any noise usually leads to over-fitting.

In order to come back to the setting presented in this introduction, one can identify a set $V \subset \mathbb{R}^d$ with the class of linear functions $\mathcal{F}_V = \{\langle v, \cdot\rangle : v \in V \subset \mathbb{R}^d\}$, so that

$$\mathcal{R}_n(V) = \mathbb{E}\left[\sup_{v\in V}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i \langle v, X_i\rangle\right)\right] . \tag{1.7}$$

As stated above, the symmetrisation lemma is one of the tools that make this complexity measure popular. It is used throughout this work (in Chapter 3 and 5 for instance), for instance to bound expectations of suprema such as $\tilde{Z}$ introduced earlier in equation (1.4).

Lemma 1.1 (Symmetrization)**.** *Let $\mathcal{F}$ be a class of functions from E to $\mathbb{R}$. Then,*

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

We note that, contrary to the Gaussian complexity, the Rademacher complexity depends not only on the set V, but also on the distribution of the random variables X_i.

1.3.3 VC-dimension.

The Vapnik–Chervonenkis (VC) dimension is one of the first way ever proposed to measure the richness and the flexibility of a class of functions. It was first introduced in Vapnik and Chervonenkis (1971). Unlike the Rademacher complexity, it is restricted to class $\mathcal{F}$ of classifiers (or boolean functions) that take values in the set $\{0, 1\}$. It is a fundamentally combinatorial measure of the complexity, and its combinatorial nature often makes it loose when compared with finer measure such as the Gaussian complexity. We will however see throughout Chapter 5 that it can at time give lead to state-of-the-art results. The definitions and facts that we state here can be found in a lot of textbooks, see for instance Vapnik (2013) or Ahsen and Vidyasagar (2019) and references therein.

Definition 1.2. *Let $\mathcal{F}$ be a set of Boolean functions on a space $\mathcal{X}$. We say that a finite set $S \subset \mathcal{X}$ is shattered by $\mathcal{F}$ if, for every subset $B \subset S$, there exists $f \in \mathcal{F}$ such that $S \cap f^{-1}(\{1\}) = B$. We call VC-dimension of $\mathcal{F}$ (and note $\mathrm{VC}(\mathcal{F})$) the largest integer n such that there exists a set S of cardinality n that is shattered by $\mathcal{F}$.*

Once again, in order to deal with the setting presented in this introduction we abusively call VC-dimension of a set $V \subset \mathbb{R}^d$ and note $\mathrm{VC}(V)$ the VC-dimension of the set of half-spaces generated by the vectors in V:

$$\mathrm{VC}(V) = \mathrm{VC}(\{x \in E \to \mathbf{1}_{\langle x,v\rangle \geq 0},\ v \in V\}).$$

We give a few examples to illustrate what this notion does and does not capture.

1. $\mathrm{VC}(\mathbb{R}^d) = d+1$. More generally, if F is a set of real-valued functions in a k-dimensional linear space, then $\mathrm{Pos}(F) = \{x \to \mathbf{1}_{f(x)\geq 0}, f \in F\}$ has VC-dimension $k+1$ (see for instance Dudley (1978), Theorem 7.2), so when it comes to linear sub-spaces, the VC-dimension captures the dimension.

2. $\mathrm{VC}(\mathcal{B}_2^d) = d+1$: when it comes to VC-dimension, the unit euclidean ball and the whole space $\mathbb{R}^d$ have the same complexity. The diameter of the set does not matter ; the combinatorial complexity of the structure only is measured. For instance the set V and the set αV for $\alpha \in \mathbb{R}_+^*$ always have the same VC-dimension, while it is not the case for Gaussian or Rademacher complexity: $\mathcal{R}_n(\alpha V) = \alpha \mathcal{R}_n(V)$, and $w^*(\alpha V) = \alpha w^*(V)$. Even more, if A is any invertible matrix, then $\mathrm{VC}(AV) = \mathrm{VC}(V)$. In consequence the VC-dimension can not measure any fine details of problems where there are no scale/rotational invariance.

3. Sparse vectors: Let $e_1, ..., e_d \in \mathbb{R}^d$ be the canonical basis of $\mathbb{R}^d$ note $\mathcal{U}_s = \{\sum_i \lambda_i v_i \mid \lambda_i \in \mathbb{R}\ \&\ \sum_i \mathbf{1}_{\lambda_i \neq 0} \leq s\}$ the set of s-sparse vectors, then

$$C_1 s \log_2(ed/s) \leq \mathrm{VC}(\mathcal{U}_s) \leq C_2 s \log_2(ed/s),$$

where C_1 and C_2 are universal constants. This can be found for instance in Theorem 3 in Ahsen and Vidyasagar (2019). In this case, the VC-dimension captures well the Gaussian complexity of the set.

We note that, unlike the Rademacher complexity, the VC-dimension does not depend on the distribution of the random variable X_i.

1.3.4 Entropy-based complexity.

Entropy-based complexity has been used since the fifties to measure how complex a set is, see for instance Kolmogorov (1959). The principle is the following one: we denote $N(V, \eta\mathcal{B}_2^d)$ or $N(V, \mathcal{B}_2^d, \eta)$ the minimal number of translated $\eta\mathcal{B}_2^d$ balls needed to cover the set V. This quantity measures how many points it takes to "discretize" the set V within radius η. This number is usually called the η-covering numbers of V, and are also frequently used in geometric functional analysis and in empirical processes theory (see for instance Vershynin (2018), Ledoux and Talagrand (2011) or van der Vaart and Wellner (1996)).

How to use such a family of numbers to describe the complexity of a set? The behaviour of $N(V, \mathcal{B}_2^d, \eta)$ when η goes to 0 often gives an idea about the complexity of V. For instance, for a linear sub-space F of dimension k, $(1/\eta)^k \leq N(F \cap \mathcal{B}_2^d, \mathcal{B}_2^d, \eta) \leq (3/\eta)^k$ (this can be found in lot of textbooks, for instance Vershynin (2018) p.85). As the dimension appears in the exponent, we are prone to turn to the logarithm of covering numbers, which is sometimes called the entropy:

$$H(V, \mathcal{B}_2^d, \eta) = \log N(V, \mathcal{B}_2^d, \eta)\ .$$

For the case of linear subsets, we see that $H(F, \mathcal{B}_2^d, \eta) \sim k \log(1/\eta)$ when η goes to 0. One way to measure the complexity of V is thus to study the behavior of $H(V, \mathcal{B}_2^d, \eta)$ when η goes to 0. Two quantities related to entropy appears naturally in empirical process theory: Sudakov's bound and Dudley's entropy integral. This two quantities are often used to bound below and above the Gaussian complexity, thanks to the two following theorems (see Ledoux and Talagrand (2011)):

Theorem 1.3 (Sudakov's minoration inequality)**.** *Let $V \subset \mathbb{R}^d$. Then, for any η, we have*

$$\mathbb{E} \sup_{t \in V} \langle G, t \rangle \geq c\eta \sqrt{H(V, \mathcal{B}_2^d, \eta)},$$

where $G \sim \mathcal{N}(0, I_p)$ and c is a universal constant.

We thus define the Sudakov's bound as $\varsigma(V) = \sup_\eta \eta \sqrt{H(V, \mathcal{B}_2^d, \eta)}$, so that $w^*(V) \geq \varsigma(V)$. We note that for the case of linear subset, we will capture the right behaviour up to universal constants: $\varsigma(H(V, \mathcal{B}_2^d, \eta)) \propto k$. The second theorem bounds by above the Gaussian complexity and it features the whole range of entropy numbers within an integral instead of using a supremum.

Theorem 1.4 (Dudley integral)**.** *Let $V \subset \mathbb{R}^d$. Then there exist an absolute constant c such that*

$$w^*(V) \leq c \int_0^\infty \sqrt{H(V, \mathcal{B}_2^d, \eta)} \mathrm{d}\eta.$$

Theorem 1.4 derives from chaining techniques (detailed for instance in Ledoux and Talagrand (2011) or Vershynin (2018)), that can be further refined using "generic chaining" techniques developed by Talagrand (1996).

1.4 Main contributions

With the main concepts, tools and questions introduced we present the various contributions of the five works that make up this thesis.

1.4.1 Robust Subgaussian Estimation of a Mean Vector in Nearly Linear Time

The first contribution of this thesis, detailed in Chapter 3, deals with the computational complexity of subgaussian mean estimation, with respect to the traditional euclidean norm. As this thesis began, two very important papers were released, that were the firsts to propose procedures achieving the subgaussian rate (1.2) while running in polynomial time in both variables N and d: Hopkins (2018) and Cherapanamjeri et al. (2019). They both run in polynomial time: $\mathcal{O}(N^{24} + Nd)$ for Hopkins (2018) and $\mathcal{O}(N^4 + N^2 d)$ for Cherapanamjeri et al. (2019) (see Cherapanamjeri et al. (2019) for more details on these running times). They do not consider an adversarial contamination of the dataset even though their results easily extend to this setup. The first chapter of this thesis thus proposes the third polynomial algorithm reaching the subgaussian rate for mean estimation, and the first one to explicitly deal with adversarial contamination. Moreover, it improves the run time of the first procedures proposed: we construct an algorithm running in time $\tilde{\mathcal{O}}(Nd + u \log(1/\delta)d)$ which outputs an estimator of the true mean achieving the subgaussian rate (1.2) with confidence $1 - \delta - (1/10)^u$ (for $\exp(-c_0 N) \leq \delta \leq \exp(-c_1 |\mathcal{O}|)$) on a corrupted database and under a second moment assumption only. In the worst case, the run time is thus of $\tilde{\mathcal{O}}(N^2 d)$.

In order to do so, this paper uses the Median of Mean heuristic, like the two previous polynomial-time procedures. Our approach in fact takes ideas from two communities: the

median-of-means principle from the statistics community and a SemiDefinite Programming (SDP) relaxation used in the Computer Science community by Cheng et al. (2019a) which can be theoretically computed fast. The computational time improvement upon the procedure in Cherapanamjeri et al. (2019) is due to the use of this covering Semidefinite program (SDP) studied in Allen-Zhu et al. (2014), Peng et al. (2012), and Cheng et al. (2019a), and popularised in the robust statistic field by Cheng et al. (2019a), at each iteration of the robust gradient descent algorithm. While in Cheng et al. (2019a) this SDP leads to procedures failing with constant probability and thus failing to reach the subgaussian rate, we show that using this kind of SDP on block means rather than on the data themselves leads not only to reach the subgaussian rate, but also to improvement in the computational cost of the algorithm. So the median of mean heuristic presents in this case a stochastic advantage and a computational one.

To prove that the proposed procedure indeed reaches the subgaussian rate, we had to come up with a new stochastic lemma interesting in its own, generalizing the one from Lugosi and Mendelson (2019c) and Lerasle et al. (2019). This lemma has been used since in a variety of other works and contexts, for instance in Regression in Cherapanamjeri et al. (2020b), or to define the notion of Stability in Diakonikolas et al. (2020). The proof of this Lemma relies on a Gaussian rounding technique similar to the one used in Grothendieck's inequality.

Very recent works Lei et al. (2020); Hopkins et al. (2020); Depersin (2020a) obtain similar results to the one from this work. They were also able to replace SDPs by spectral methods for the computations of a robust descent direction at each step. Indeed, even though cover SDPs are from a theoretical point of view computationally efficient (see Allen-Zhu et al. (2014),Peng et al. (2012)) they are notoriously difficult to implement in practice whereas the power methods used in Lei et al. (2020); Hopkins et al. (2020); Depersin (2020a) open the door to implementable algorithms. It is interesting to note that the computational time proposed in this work is still to this date the best run time known, and it is a conjecture that it might even be the fastest possible way to reach the subgaussian rate.

1.4.2 A Spectral Algorithm for Robust Regression with Subgaussian Rates

The second contribution we make, detailed in Chapter 4 deals with the question of reaching sub-Gaussian bounds in polynomial time, but for regression instead of mean estimation. We recall quickly the standard linear regression setting where data are couples $(X_i, Y_i)_i \in \mathbb{R}^d \times \mathbb{R}$ and where one looks for the best linear combination of the coordinates of an input vector X to predict the output Y, that is we look for β^* defined as follows.

$$\beta^* = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \ell(\beta) = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \mathbb{E}(Y_1 - \langle \beta, X_1 \rangle)^2.$$

Whether reaching sub-Gaussian rates in that framework, under weak moment assumptions, was even possible was an open question for a long time. Indeed for a time the best known polynomial algorithms were the one from Prasad et al. (2018) or from Hsu and Sabato (2016). The guarantee for those two algorithms is the following: when the covariance of X is the identity and when the noise $\xi = Y - \langle \beta^*, X \rangle$ has bounded variance, $\ell(\hat{f}) - \ell(f^*) \leq \mathcal{O}(\frac{\log(1/\delta)d}{N})$ with probability $1-\delta$. This rate does not present this decoupling between complexity and deviation that we called sub-Gaussian. The article from Cherapanamjeri et al. (2020a) has been the first to construct a polynomial-time method achieving the sub-Gaussian rate of the OLS in the Gaussian setting $\ell(\hat{f}) - \ell(f^*) \leq \mathcal{O}(\frac{\log(1/\delta) \vee d}{N})$. When Chapter 4 was first published, Cherapanamjeri et al. (2020a) was the only procedure running in polynomial algorithm achieving the optimal subgaussian rate. However, Cherapanamjeri et al. (2020a) uses the Sum of Square (SoS) programming hierarchy to design their algorithm. Even if SoS hierarchy runs in polynomial time, its reliance on solving

large semi-definite programs makes it impractical and remains a theoretical result leaving still open the question on the existence of a practical efficient algorithm achieving optimal subgaussian rates.

In Chapter 4, we tackle this issue, showing that techniques from Lei et al. (2020) combined with lemmas from Depersin (2020a) can be used to give the first practical, nearly quadratic (and in fact in most cases nearly-linear) algorithm that reaches the subgaussian rate. We also conduct numerical experiments on simulated data with our proposed procedure to show that it is indeed practical and fast. Moreover, as predicted by our theoretical findings, our simulation analysis shows that it is robust both to heavy-tailed data and to outliers. To the best of our knowledge, this is the first time that numerical experiments implementing the exact formulation of a sub-gaussian estimator are conducted for a regression algorithm with sub-gaussian rates and polynomial time guarantees.

1.4.3 Robust subgaussian estimation with VC-dimension

The third contribution we make is detailed in Chapter 5, and it deals with statistical complexity rather than computational complexity. We tried to show how one can use VC-dimension to get state-of-the art bounds in non-euclidean estimation with heavy-tail data.

In this work, we show that the analysis presented in Lugosi and Mendelson (2019c), in Lecué and Lerasle (2019), Lerasle (2019) or in Lecué and Lerasle (2020), and generalized in Lugosi and Mendelson (2019b), all based on the Median-of-mean principle and the use of Rademacher complexities, can be modified in order to achieve sub-gaussian rates for sparse or structured problems assuming only bounded two-order moments. The method developed in Lerasle (2019) or in Lecué and Lerasle (2020) requires data to have at least $\log(d)$ finite moments (where d is the dimension of the space) in order to exploit the sparsity of the problem and offers no guarantees without that requirement, and to the date is the best known. We show that we can drop this condition by judiciously introducing VC-dimension in the different proofs, and exploit the sparsity of the problem *with only two moments*. We show in Chapter 5 that classical approaches using local Rademacher complexities cannot achieve this type of subgaussian bounds under only a second moment assumption. Somehow the classical approach used so far does not capture the right statistical complexity of high-dimensional problems under low-dimensional structural assumptions and under only a second moment assumption: **it seems that the Rademacher complexity is not the right way to measure the complexity of the problem of structured mean estimation in any norm**. Our VC-dimension based approach allows to overcome this issue and to go beyond this $\log d$ subgaussian moments assumption that has appeared in all works on robust and subgaussian estimation in the high-dimensional framework Lerasle (2019). We also show that this general technique can be easily replicated and give new robust estimators that achieve state-of-the-art bounds for different estimation tasks such as Regression, Mean estimation with non-Euclidean norms, Robust low-rank matrix estimation and Covariance estimation.

This Chapter is not the first to introduce VC-dimension in robust estimation problems: it has been inspired by Chen et al. (2018) and Gao (2017) for instance. In those two papers, estimation and regression with possible sparsity structure and outliers are also achieved with optimal rates, using VC-dimension techniques, but their assumption and their framework is somehow different from the one we studied in this thesis. For instance, Chen et al. (2018) estimates the center of *symmetric distributions* without moment assumption. In comparison, our estimators are for mean and covariance, thus moment assumption is needed, but we do not need the distributions at stake to be symmetric.

We note that VC-dimension has some advantages over Rademacher complexity in some cases, but this forth Chapter shows that is does not leads to optimal rates in all cases, and thus that it is not always the right way to measure the complexity of a robust estimation task. Indeed, using VC-dimension in mean estimation, we lose a nice dependence of the risk bounds in the covariance structure: our rates for (non-sparse) mean estimation depend on the ambient dimension d instead of the effective rank $\mathrm{Tr}(\Sigma)/||\Sigma||_{op}$ (that is captured by Rademacher Complexity). In particular, the general VC-dimension approach does not generalize directly to infinite dimensional spaces. In the last section of Chapter 5, we show that this issue can be overcome if we have some knowledge on the covariance matrix, proposing a new procedure for that case.

1.4.4 On the robustness to adversarial corruption and to heavy-tailed data of the Stahel-Donoho median of means

In Chapter 6, we deal with estimation with respect to the norm $x \in \mathbb{R}^d \to \left\|\Sigma^{-1/2}x\right\|_2$, where Σ is the covariance matrix of the data, and we assume that this matrix is not known beforehand by the statistician. So in a sense, we try to estimate with respect to an unknown norm, while techniques derived in Chapter 5 mainly apply to known norms. We study this particular norm, whose unit ball is the ellipsoid $\Sigma^{1/2}B_2^d$, because it is the best metric – that is the one leading to minimal volume confidence sets for a given confidence – in the benchmark i.i.d. Gaussian case.

While the subgaussian rate could be obtained with estimators from Lugosi and Mendelson (2019b) or Lerasle et al. (2019), these estimators would require knowledge about Σ in their construction. One therefore has to consider other techniques. In this Chapter, we show that it can be done thanks to a notion of depth/outlyingness introduced at the beginning of the 80's which uses a normalization by a robust estimation of the scale called *Stahel-Donoho Outlyingness* (SDO), which has been first introduced in Donoho and Gasko (1992). We couple this notion of outlyingness with the Median-Of-Mean heuristic to get subgaussian estimators with respect to the metric $\left\|\Sigma^{-1/2}.\right\|_2$. On our way to our goal, we complement the results on the $\sqrt{n}$-consistency and the asymptotic normality of Stahel-Donoho estimators that can be found in Maronna and Yohai (1995) and Tyler (1994) by deriving the first non-asymptotic convergence rate for the SDO median (as well as its median of means version). We also show that the robustness properties of the original SDO median and its MOM version goes beyond the Huber's contamination model and that they still persist in the adversarial corruption model from Setting 1.1. We also use the robust scaling from the Stahel-Donoho Outlyingness to build estimators of the covariance matrix under some regularity assumption.

1.4.5 Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms

Our last contribution, detailed in Chapter 7, is to give a new lower bound for mean estimation in any norm. Lugosi and Mendelson (2019b) gives the following lower bound on mean estimation :

Theorem 1.5. *[Theorem 3 from Lugosi and Mendelson (2019b)] There exists an absolute constant $c > 0$ such that the following holds. If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4)$,*

$$\mathbb{P}_{\mu^*}^N \left[\|\hat{\mu} - \mu^*\| \leq r^*\right] \geq 1 - \delta$$

where $\mathbb{P}_{\mu^*}^N$ is the probability distribution of $(X_i)_{i\in[N]}$ when the X_i are i.i.d. $\mathcal{N}(\mu^*, \Sigma)$ then

$$r^* \geq \frac{c}{\sqrt{N}} \left(\sup_{\eta>0} \eta \sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)} + \sup_{v\in B^\circ} \left\|\Sigma^{1/2}v\right\|_2 \sqrt{\log(1/\delta)} \right)$$

where $N(\Sigma^{1/2}B^\circ, \eta B_2^d)$ *is the minimal number of translated of* ηB_2^d *needed to cover* $\Sigma^{1/2}B^\circ$.

The complexity term in this lower bound is thus measured using the Sudakov's bound that we defined in Theorem 1.3. However, there is a gap between this lower bound and the upper bounds depending on the Gaussian mean width in the Gaussian case. This gap that comes from the looseness of Sudakov's inequality presented in Theorem 1.3. For ellipsoids for instance, Sudakov's bound is not sharp in general and therefore the lower bound from Theorem 1.5 fails to recover the classical subgaussian rate for the standard Euclidean norm case (that is for $S = B_2^d$) which is given in Lugosi and Mendelson (2019c) by

$$\sqrt{\frac{\mathrm{Tr}\,(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op}\log(1/\delta)}{N}}. \tag{1.8}$$

Indeed, when $\|\cdot\|$ is the ℓ_2^d Euclidean norm then $\mathbb{E}\left\|\Sigma^{1/2}G\right\| = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \sim \sqrt{\mathrm{Tr}(\Sigma)}$. In contrast the entropy of $\Sigma^{1/2}B^\circ = \Sigma^{1/2}B_2^d$ w.r.t. ηB_2^d can be computed using equation (5.45) in Pisier (1989) that

$$\sup_{\eta>0}\eta\sqrt{\log_2 N(\Sigma^{1/2}B_2^d, \eta B_2^d)} \sim \sup_{n\geq 1, k\in[d]} \frac{\sqrt{n}}{2^{n/k}}\left|\prod_{j=1}^k \sqrt{\lambda_j}\right|^{1/k} \sim \sqrt{\sup_{k\in[d]} k\left|\prod_{j=1}^k \lambda_j\right|^{1/k}} \tag{1.9}$$

where $\lambda_1 \geq \ldots \geq \lambda_d$ are the singular values of Σ. In particular, when $\lambda_j = 1/j$, the entropy bound (1.9) is of the order of a constant whereas the Gaussian mean width is of the order of $\sqrt{\log d}$. We fill this gap in Chapter 7 by showing a lower bound where the entropy is replaced by the (larger) Gaussian mean width. In order to do so, we use Anderson's Lemma, and analytic arguments, intead of geometric and volumetric arguments used by Lugosi and Mendelson (2019b) to get the Sudakov's bound as a lower bound.

We also show that this rate can sometimes be achieved by a solution to a convex optimization problem in the adversarial and L_2 heavy-tailed setup by considering minimum of some Fenchel-Legendre transforms constructed using the Median-of-means principle.

1.5 Unanswered Questions and Future Research Direction.

After those different contributions, there is still a number of exciting unanswered questions that can be starting points for future research.

1.5.1 A new notion of complexity?

The different complexity measures mentioned in Section 1.3 gives standard tools to measure the complexity of a set and give some ideas about the rates that could be achieved with heavy-tailed data. However, it is plausible that none of these measures is the right one, and that the right way to measure the complexity of robust estimation tasks is yet to be found. The quantity that is crucial in all the works is

$$r \to \mathbb{E}\left(\sup_{v\in V}\sum_{k=0}^K \mathbf{1}_{\langle \bar{Y}_k - \mu, v\rangle \geq r} - K\mathbb{E}(\mathbf{1}_{\langle \bar{Y}_k-\mu, v\rangle \geq r})\right).$$

Bounding this quantity using the VC-dimension of V yields a bound independent of the covariance of Y. On the other hand, bounding that quantity by the Rademacher complexity of the Y_i (like in Lecué and Lerasle (2020), Lugosi and Mendelson (2019b)) does not exploit the boundedness of the indicator function and necessitates unnecessary stronger assumptions on data (see Chapter 5). The ideal would be to conciliate both ideas, and to find a nice in-between that would take into account both the boundedness and the dependency in the covariance structure.

1.5.2 Tightening the constants

In most of the work we present in this thesis, the bounds are all given up to universal constants: we have not focused on the constants but rather on how the rate depends on d, N, Σ, or other parameters of interest, such as the sparsity s. In consequence, the constants we give in most of this work are not optimized and analysed. However it seems that, when computed, they are huge and make most of the theory not usable as such for practical implementation. Simulations tends to show that the procedures proposed seem to reach in practice rates with much smaller and practical constants.

Trying to optimize the constants to find the "sharp rate", and search for procedures that could lead to better constants, even in the most simple case of estimation with respect to the euclidean norm, has not been done in the literature to the best of our knowledge. It seems to be both challenging from a theoretical perspective and interesting from a practical point of view.

1.5.3 Finding algorithms for estimation in any norms

This thesis is somehow separated in two parts: a more practical one and a more theoretical one. Chapter 3 and 4 mainly deal with tractable estimation in euclidean set-ups, and their contributions are (mainly) algorithms, while Chapter 5, 6 and 7 mainly deal with theoretical estimators in non-euclidean set-ups. It seems like it would be promising to build a bridge between those two parts and to look for tractable procedures in various non-euclidean set-ups.

For instance, only very little is known on the theoretical computational side for the Stahel-Donoho outlyingness. In Section 5 of Donoho and Gasko (1992), an algorithm running in time $\mathcal{O}(K^{d+1} \log K)$ is mentioned but its time complexity is making this approach impractical for dimensions larger than 5. There are to our knowledge no theoretical results of any kind on the convergence of some approximate algorithm for the computation of the SDO of a point in $\mathbb{R}^d$ that could be used in practice. As mentioned already in Donoho and Gasko (1992), "some sort of computational breakthrough is necessary to make the estimators, as defined here, really practical". This looks to be still the case.

In the same way, there is very little known about the computational side of robust sparse estimation. To the best of our knowledge, there is no tractable algorithm suitable for sparse mean estimation with subgaussian rates. Knowing if it is even possible to find such an algorithm is one of the main open question left unanswered by this thesis, and that seems like a very promising research direction.

CHAPTER 2

Introduction en français : Robustesse et complexité

Cette thèse tente d'évaluer la complexité de certaines tâches statistiques robustes. Dans l'introduction, la signification et l'utilité de cet objectif global seront analysées et précisées. D'abord, les deux principaux termes en jeu, robustesse et complexité, seront successivement discutés. Puis les cinq travaux qui composent cette thèse seront présentés, situés par rapport au contexte général, et leurs apports et limites respectifs seront exposés.

Contents

2.1 Qu'est-ce que la statistique robuste?

L'objectif principal d'un statisticien est d'extraire des informations utiles de données d'observation, par exemple en déduisant des modèles, en identifiant des effets causaux, en un mot en apprenant sur un phénomène donné à partir des données. La plupart du temps, le statisticien doit faire des

hypothèses sur les données afin d'obtenir des garanties que ses procédures conduiront effectivement à des informations fiables. Les hypothèses les plus courantes sont :

- que les données sont indépendantes et identiquement distribuées, ce qui signifie qu'elles sont produites à partir du même processus, et indépendamment les unes des autres, qu'elles sont différentes réalisations de la même *variable aléatoire*,
- que cette *variable aléatoire* est supposée par le statisticien avoir une certaine propriété donnée (par exemple une esperance finie),
- que la distribution de la variable aléatoire dont les observations sont censées être tirées appartient à une famille spécifique de distributions déterminée au préalable par le statisticien (par exemple, la famille des distributions normales) et appelée le modèle statistique. Lorsqu'on fait ce genre d'hypothèse, les statistiques en jeu sont dites *paramétriques* car dans ce cadre, les différentes distributions de la famille donnée sont souvent identifiées par un paramètre (et peuvent être entièrement décrites avec ce paramètre), de sorte que la recherche de la bonne distribution parmi la famille revient à rechercher le bon paramètre.

Nous remarquons que ces hypothèses sont de plus en plus fortes. On pourrait dire que l'objectif global de la statistique robuste en tant que domaine est de remettre en question ces hypothèses et d'étudier s'il est possible de donner certaines garanties sous des hypothèses plus faibles et plus réalistes. Pour être plus précis, le domaine des statistiques robustes tente d'étudier

- ce qui arrive aux procédures classiques (non robustes) lorsque les hypothèses sur lesquelles elles ont été créées sont relâchées,
- quelles sont les hypothèses minimales que l'on doit faire sur les données pour qu'il soit théoriquement possible d'en extraire des informations,
- et comment trouver des procédures qui fonctionnent toujours lorsqu'on fait des hypothèses minimales.

Hampel et al. (1986) donne le résumé suivant : "La statistique robuste est un ensemble de connaissances relatives aux déviations des hypothèses idéalisées en statistique."

2.1.1 Pourquoi des statistiques robustes?

On peut se demander quelle est l'utilité d'une telle théorie. C'est l'objet d'une littérature riche, initiée dans les années 60 par Tukey (1960), Huber (1964) et Hampel (1973), et les fondements théoriques de la robustesse sont exposés avec beaucoup de détails dans un certain nombre d'ouvrage (voir par exemple Huber (1981), Hampel et al. (1986), Huber and Ronchetti (2009), ou Maronna et al. (2006a)). Nous rappelons et illustrons deux des principaux arguments qu'ils développent, qui peuvent servir de points de départ pour comprendre les apports de cette thèse.

OUtliers. C'est un point soulevé à la fois par Huber (1981) et Hampel et al. (1986) : les échantillons collectés dans les sciences physiques, naturelles ou sociales contiennent de "bonnes données", qui sont bien décrites par le modèle du statisticien, mais elles sont mélangées à une fraction de "mauvaises données", également appelées erreurs grossières ou outliers - typiquement dans une fourchette comprise entre 1% et 10% selon Huber (1981) pour les ensembles de données de l'époque (dont l'ampleur a pu changer avec les données Internet et les données déclaratives). Ces erreurs peuvent provenir d'erreurs de copie, de calcul, d'inattention de l'expérimentateur, etc., mais aussi d'un adversaire qui tente de tromper le statisticien. Ces mauvaises données ou "valeurs aberrantes" ne sont pas bien décrites par le modèle du statisticien, et peuvent être éloignées de plusieurs ordres de grandeur du phénomène qu'il veut quantifier.

Par exemple, une partie des données peut être exprimée dans une unité erronée : lorsqu'on leur demande leur salaire mensuel dans les enquêtes en ligne, une fraction des répondants donnent plutôt leur salaire annuel. De telles erreurs peuvent être très coûteuses : on peut penser à l'échec du lancement de l'orbiteur climatique de Mars en 1999 par la NASA, causé par l'expression de certaines données clés dans des unités non métriques. Les estimateurs classiques sont très sensibles à ce type d'erreur grossière : par exemple, considérons une série d'observations de la taille des personnes, avec 99 observations exprimées en mètres et 1 observation exprimée par erreur en centimètres. La moyenne empirique de cette série pourrait conduire à conclure que la taille moyenne des personnes est d'environ $3,3$ mètres, ce qui montre qu'il suffit d'une petite fraction de mauvaises données pour se faire une idée très erronée du phénomène en jeu lorsqu'on utilise des estimateurs non robustes.

Les erreurs grossières sont une violation de la première hypothèse classique selon laquelle toutes les données sont produites à partir du même processus. En présence d'erreurs grossières, *toutes les observations n'ont pas la même valeur informative*. La plupart des procédures statistiques classiques ne présentent des garanties que lorsque les données sont i.i.d. (indépendantes et identiquement distribuées), ainsi la présence d'erreurs est un premier argument pour justifier le besoin de développer de nouvelles procédures robustes.

Données à queue lourde. Un autre argument peut être trouvé pour justifier le besoin d'une théorie robuste : la présence de phénomènes à queue lourde. Les queues lourdes sont caractéristiques des phénomènes où il existe une probabilité significative d'observer une seule valeur énorme, c'est-à-dire d'un ordre de grandeur différent des autres. Contrairement à l'erreur grossière, cette valeur énorme n'est pas une erreur et contient des informations sur le phénomène en jeu. Les pertes d'assurance, les rendements financiers, la contagion sociale, l'activité de retweet d'un tweet sont tous des exemples de phénomènes à queue lourde, voir Resnick (2007) pour plus d'exemples et des explications détaillées. L'exemple suivant illustre le problème soulevé par les données à queue lourde. Prenons une variable aléatoire X qui est égale à 0 avec une probabilité de 999/1000, et 1000 avec une probabilité de 1/1000, de sorte que sa valeur moyenne est de 1 (et son écart-type d'environ 32). Lorsque l'on dispose de 10 d'observations de cette variable aléatoire, on a environ 1% de chance de voir la valeur 1000, ce qui constitue un événement extraordinaire. En observant un tel événement parmi les 10 de données, le statisticien désireux d'estimer la valeur moyenne du phénomène est confronté à un dilemme : doit-il prendre en compte cette observation ? En incluant cette observation, la moyenne empirique des observations est de 100, ce qui se situe à quelques écarts types de la vraie moyenne. Exclure cette observation avant de prendre la moyenne empirique permet d'obtenir une meilleure estimation de la moyenne dans ce cas, mais sur quelle base le statisticien doit-il écarter certaines valeurs ? La théorie robuste tente de répondre à ces questions.

Même si cet argument de la queue lourde a surtout été exploré dans la littérature robuste au cours de la dernière décennie suite aux travaux de Catoni (2012), il est déjà mentionné dans Huber (1981). La présence de données à queue lourde brise l'hypothèse, très courante en statistique, selon laquelle les observations sont tirées d'une distribution gaussienne (ou sous-gaussienne) sous-jacente, ou plus généralement d'une distribution présentant de bonnes propriétés de concentration.

Rejet des outliers? D'après les exemples présentés ci-dessus, on pourrait avoir l'impression qu'il suffit de prétraiter les données pour éliminer les valeurs aberrantes avant d'appliquer les procédures standard, plutôt que de trouver de nouvelles procédures robustes. L'exemple des données à forte queue de distribution montre qu'il n'est pas toujours simple de savoir si une

observation est aberrante, et cela devient encore plus compliqué lorsque les observations sont multidimensionnelles, comme nous le verrons tout au long de cette thèse. Si dans certains cas le rejet des outliers peut être une bonne idée, il est souvent compliqué à mettre en œuvre (voir Diakonikolas et al. (2019b) par exemple) et fournir des garanties pour de telles procédures nécessite les outils théoriques développés par la statistique robuste.

2.1.2 Modèles pour les outliers

Afin de donner des garanties sur leurs procédures, les statisticiens travaillant sur la statistique robuste doivent faire certaines hypothèses sur les données, même si celles-ci sont plus faibles et plus réalistes que les hypothèses classiques. Dans cette partie, nous présentons les principaux modèles adoptés par la statistique robuste et précisons le cadre qui sera utilisé tout au long de cette thèse.

Modèle de contamination de Huber . La première génération de statisticiens travaillant sur les statistiques robustes a proposé de prendre en compte les outliers avec le modèle suivant, parfois appelé *Modèle de contamination de Huber* (présenté et étudié par exemple dans Huber (1981)) : les observations sont supposées être des réalisations i.i.d. tirées d'une variable aléatoire avec une fonction de distribution cumulative

$$F = (1 - \epsilon)F_\theta + \epsilon H,$$

où F_θ appartient à une famille paramétrique connue à l'avance par le statisticien (par exemple la famille gaussienne) mais où H peut être *n'importe quelle* fonction de distribution cumulative, et où ϵ représente le taux de contamination, la fraction supposée de l'erreur brute (pour laquelle une limite supérieure est souvent supposée connue). La plupart des premiers travaux en statistiques robustes visaient à trouver des estimateurs efficaces dans ce cadre proche d'un cadre paramétrique. Ce modèle est toujours un domaine de recherche actif, voir par exemple Chen et al. (2017). Nous remarquons que de nombreux travaux traitant de ce premier modèle se concentrent sur l'obtention de résultats asymptotiques, ainsi que de propriétés non asymptotiques telles que le point de rupture. En revanche, le point de vue adopté dans cette thèse est uniquement *non-asymptotique*, inspiré par les tendances récentes en statistique robuste initiées par Catoni (2012) et par des travaux de la communauté informatique tels que Diakonikolas et al. (2016).

Modèle de contamination adversarial. Le modèle que nous traitons dans ce travail est plus général que le modèle de contamination de Huber et est parfois appelé "contamination adversarial". Il est difficile de le retracer, mais il semble avoir été décrit et popularisé par la communauté informatique, par exemple dans Diakonikolas et al. (2016). Les échantillons sont générés à partir du processus suivant : D'abord, N échantillons sont tirés indépendamment d'une certaine distribution inconnue. Ensuite, un adversaire est autorisé à regarder les échantillons et à en corrompre arbitrairement une fraction de ϵ avant de remettre les données corrompues au statisticien. Le cadre peut être décrit plus formellement comme suit :

Setting 2.1. *Il existe N variables aléatoires i.i.d. distribuées comme X notées $(\tilde{X}_i)_{i=1}^N$ dans $\mathbb{R}$ qui sont indépendantes. Ces variables ne sont pas directement observées par le statisticien, elles sont d'abord données à un "adversaire" qui est autorisé à modifier jusqu'à $\lfloor \epsilon N \rfloor$ de ces variables avant de retourner un jeu de données modifié $(X_i)_{i=1}^N$ au statisticien.*

La seule information dont dispose le statisticien est qu'il existe un ensemble (éventuellement aléatoire, éventuellement dépendant des données) $\mathcal{O}$ tel que, pour tout $i \in \mathcal{O}^c$, $X_i = \tilde{X}_i$. La seule hypothèse sur l'ensemble $\mathcal{O}$ concerne sa taille : le statisticien sait que $|\mathcal{O}| \leq \lfloor \varepsilon N \rfloor$ (le

statisticien peut dans certains cas ne même pas connaître ε et s'y adapter). Le statisticien ne sait pas quelles données ont été modifiées, l'ensemble $\mathcal{O}^c$ est donc inconnu. Le statisticien tente d'acquérir certaines connaissances (telles que la moyenne ou la variance) sur la variable aléatoire X à partir de l'ensemble de données corrompues $(X_i)_{i=1}^N$.

Dans cette configuration, non seulement les valeurs aberrantes peuvent être corrélées entre elles et avec les valeurs aberrantes, mais les valeurs non aberrantes peuvent également être corrélées entre elles (car l'adversaire peut choisir les échantillons originaux à conserver et, ce faisant, corréler les échantillons qu'il conserve, par exemple en ne conservant que les échantillons les plus grands lorsqu'ils sont à valeur réelle), ce qui ne peut pas être le cas dans le modèle de contamination de Huber.

Ce modèle généralise celui de Huber, et il a été utilisé, comme celui de Huber, pour traiter les déviations d'un modèle *paramétrique*, par exemple dans Du et al. (2017), Diakonikolas et al. (2019c) ou Cheng et al. (2019b), où les inliers sont supposés suivre une distribution gaussienne. En revanche, cette thèse, ainsi qu'une ligne de travail moderne en apprentissage robuste ouverte par Catoni (2012), traite des *statistiques non-paramétriques*. (nous essayons souvent d'estimer la moyenne d'une distribution d'échantillon), avec un accent particulier mis sur les données à forte queue de distribution, comme nous l'expliquerons ci-dessous. Le modèle de contamination contradictoire est utilisé ici de manière non paramétrique, la forme générale de la distribution que suivent les valeurs aberrantes étant supposée connue à l'avance.

2.1.3 Traitement des données mal concentrées

Les deux modèles décrits ci-dessus (contaminations de Huber et adversaires) permettent de traiter les erreurs grossières, mais ils ne répondent pas directement aux problèmes soulevés par les phénomènes de queue lourde. Revenons à notre exemple, où un statisticien veut estimer la valeur moyenne de la variable aléatoire X qui est égale soit à 0 (avec une forte probabilité) soit à 1000 (avec une faible probabilité), lorsqu'on lui donne $N = 10$ observations. Si elle utilise la moyenne empirique habituelle, qui est l'outil standard en statistique classique pour estimer une moyenne, elle sera "proche" de la vraie valeur avec une probabilité de 99% et radicalement fausse ($\sim 3\sigma$) avec une probabilité de 1%, une probabilité faible mais non négligeable. Catoni (2012) soulève les questions suivantes : étant donné un "rayon de précision" r, quelle est la plus petite *probabilité d'échec* $\delta(r)$ telle qu'il est possible de trouver un estimateur qui se situe avec une probabilité $1 - \delta(r)$ dans un rayon r de la vraie valeur moyenne de la variable aléatoire ? Dans notre exemple, pour un rayon $\sim \sigma$ par exemple, est-il possible d'obtenir une meilleure probabilité de défaillance que 1% ? Quelles procédures d'estimation peuvent conduire à un tel estimateur ? Et quelles sont les hypothèses minimales que l'on doit faire sur la distribution de la variable aléatoire ?

Quelles hypothèses sur la distribution sous-jacente? Pour capturer le phénomène des queues lourdes, nous voulons obtenir des propriétés statistiques *sans faire l'hypothèses que les données soient gaussiennes ou bornées* (ou toute autre hypothèse de concentration forte), et c'est dans ce sens que nous appellerons nos estimateurs robustes aux queues lourdes. Quelle hypothèse plus faible devrions-nous faire sur la distribution sous-jacente des données, qui serait suffisamment faible pour ne pas limiter sévèrement l'applicabilité des résultats, et suffisamment forte pour conduire à des résultats intéressants et significatifs ? Cette question est étudiée en détail dans Devroye et al. (2016) : les auteurs montrent que, afin d'atteindre un taux qui ressemble à celui atteint asymptotiquement avec le théorème central limite, taux dont on peut prouver qu'il est "essentiellement optimal" et qui sera discuté plus en détail ci-dessous, l'hypothèse minimale à faire sur les données est l'existence d'un second moment fini : $\mathbb{E}(X^2) < \infty$. Comme une grande

majorité de la littérature robuste qui traite des données à queue lourde, la majeure partie de cette thèse se conforme à cette analyse (Chapitre 3, 4 et 5)

Setting 2.2. *La variable aléatoire X à partir de laquelle sont échantillonnées les N variables $(\tilde{X}_i)_{i=1}^N$ dans $\mathbb{R}$ a une moyenne μ, que nous essayons d'estimer, et un moment du second ordre (éventuellement inconnu) $\sigma^2 = \mathbb{E}[(X-\mu)^2]$.*

La plupart des travaux présentés traitent du cas où la distribution sous-jacente a un moment du second ordre, parfois connu, parfois inconnu. Nous notons cependant que certains papiers récents étudient des cas où le second moment n'existe même pas et est remplacé par des moments d'ordre $1+\alpha$ avec $\alpha < 1$ par exemple Cherapanamjeri et al. (2020b). La fin de cette thèse présente également de nouveaux paramètres qui ne nécessitent pas l'existence d'un second moment fini (dans certaines parties du chapitre 6 et 7), en utilisant comme inspiration d'autres résultats asymptotiques que le théorème central limite.

La moyenne empirique. Revenons à notre exemple, et à l'insuffisance de la moyenne empirique. Si l'on dispose de N réalisations indépendantes d'une variable aléatoire de moyenne μ et de variance σ^2, l'inégalité de Chebyshev nous dit que la moyenne empirique $\hat{\mu}$ vérifie, avec une probabilité supérieure à $1-\delta$, $|\hat{\mu}-\mu| \leq \sigma/\sqrt{N\delta}$. Catoni (2012) indique que ce taux est net pour la moyenne empirique en général, on ne peut donc pas donner une meilleure inégalité pour la moyenne empirique qui soit valable pour toutes les distributions avec une variance. Ainsi, pour la moyenne empirique, $r(\delta) = \sigma/\sqrt{N\delta}$, ou de manière équivalente $\delta(r) = \sigma^2/(Nr^2)$. Cette limite se détériore rapidement pour les petites valeurs de δ : si l'on veut obtenir des résultats à haute probabilité, par exemple avec $\delta \sim 10^{-5}$, avec environ un millier de points de données, on obtient un rayon de $\simeq 10\,\sigma$. Est-il possible de faire mieux, peut-on trouver des procédures qui donnent un rayon plus petit ? Cette question a été soulevée dans Catoni (2012), qui donne également une réponse.

Estimation sous-gaussienne. Catoni (2012) affirme que le taux atteint par la moyenne empirique peut effectivement être amélioré de manière drastique. Le taux qu'il propose s'inspire de la théorie asymptotique. En effet, lorsque les données ont un second moment fini σ, le théorème central limite garantit que la moyenne empirique a des queues gaussiennes, asymptotiquement, sans faire d'autres hypothèses sur les données, ainsi, lorsque $N \to \infty$,

$$\mathbb{P}\left(|\hat{\mu}-\mu| < \frac{\sigma\phi^{-1}(1-\delta/2)}{\sqrt{N}}\right) \to \delta, \tag{2.1}$$

où ϕ est la fonction de distribution cumulative de la distribution normale standard. Catoni (2012) prouve également que ce taux asymptotique ne peut pas être amélioré : aucun estimateur non-asymptotique ne peut atteindre des queues meilleures que gaussiennes pour toutes les distributions d'une classe "suffisamment grande" (qui contient au moins toutes les distributions gaussiennes avec une variance donnée σ^2). La principale conclusion de Catoni (2012) est qu'il est possible de trouver des estimateurs non-asymptotiques atteignant le taux (1.1), à constantes multiplicatives universelles près.

Notons que le taux $\sigma\phi^{-1}(1-\delta/2)\sqrt{N}$ obtenu dans (1.1) est le taux atteint par la moyenne empirique lorsque les données sont un échantillon de N de variables gaussiennes i.i.d.. Nous disons donc qu'un estimateur *réalise un taux subgaussien* s'il réalise le taux (1.1) à constantes multiplicatives près. En un sens, nous voulons que nos estimateurs soient aussi bons que si les données étaient gaussiennes, *même lorsque l'échantillon réel est mal concentré* (on suppose seulement qu'il a un second moment).

Remark 2.1. *Nous savons que* $\phi^{-1}(1-\delta/2) \leq \sqrt{2\ln(2/\delta)}$*, et pour de petits* δ*,* $\phi^{-1}(1-\delta/2) \sim \sqrt{2\ln(2/\delta)}$*. Comme nos résultats sont tous formulés à constantes multiplicatives près*[1]*, nous utilisons la formulation explicite* $C\sqrt{\ln(1/\delta)}$ *dans la plupart de nos résultats.*

Pour revenir à l'exemple, le taux (1.1) avec $\delta \sim 10^{-5}$ et $N \sim 1000$ donne un rayon d'environ $0,15\sigma$, et la dépendance logarithmique cruciale lui permet de se dégrader drastiquement moins vite que le taux obtenu par l'inégalité de Tchebychev lorsque δ tends vers 0.

Catoni (2012) construit un estimateur subgaussien $\hat{\theta}$ implicitement, comme solution de l'équation suivante :

$$\sum_i \psi\left[\alpha(X_i - \hat{\theta})\right] = 0,$$

où α est un réel positif soigneusement choisi, et ψ est une fonction d'influence bornée telle que $\psi(x) \sim x$ lorsque x est proche de 0. Cette thèse au contraire étudie principalement une autre manière de construire un tel estimateur subgaussien, l'heuristique Median-Of-Mean (MOM).

L'heuristique de la médiane des moyennes. Cette approche, décrite pour la première fois dans Nemirovsky and Yudin (1983) et Jerrum et al. (1986), a fait l'objet d'une attention particulière dans les communautés de la statistique et de l'apprentissage automatique au cours de la dernière décennie, par exemple dans Bubeck et al. (2013), Lerasle and Oliveira (2011), Devroye et al. (2016), Minsker and Strawn (2017) ou Minsker (2015). Cette approche a été conçue à l'origine pour construire des estimateurs robustes aux données à queue lourde dans divers contextes (voir Alon et al. (1999), Jerrum et al. (1986) et Birgé (1984) par exemple). Elle peut être définie comme suit : on commence par diviser aléatoirement les données en K blocs $B_1, \ldots, B_K$ de taille égale $m = \lfloor N/K \rfloor$ (si K ne divise pas N, on enlève juste quelques données), K étant choisi d'ordre $\log(1/\delta)$ avec δ la probabilité d'échec souhaitée par le statisticien. On calcule ensuite la moyenne empirique au sein de chaque bloc : pour $k = 1, \ldots, K$,

$$\bar{X}_k = \frac{1}{m} \sum_{i \in B_k} X_i.$$

L'estimateur final $\hat{\mu}_K$ est la médiane des dernières K moyennes empiriques. Le premier article qui prouve formellement que cet estimateur a des queues sub-gaussiennes est Devroye et al. (2016).

Theorem 2.1 (Devroye et al. (2016))**.** *L'estimateur* $\hat{\mu_K}$ *a des déviations subgaussiennes : avec la probabilité* $\geq 1 - e^{-K/32}$*,*

$$|\hat{\mu}_K - \mu| \leq \sqrt{8\frac{\sigma^2 K}{N}}$$

L'idée principale réside dans le passage d'une variable non bornée à une variable bornée grâce à l'opérateur médian. En effet, pour savoir si la médiane est dans un intervalle I, on calcule $Z := \sum_{k=1}^{K} \mathbf{1}_{\bar{X}_k \in I}$: lorsque cette quantité est supérieure à $K/2$, la médiane est dans I. La quantité Z, contrairement à la moyenne empirique, est une somme de variables bornées, ainsi l'inégalité de Hoeffding (voir Hoeffding (1963)) stipule qu'elle est proche de sa moyenne avec une probabilité d'échec exponentiellement faible, ce qui conduit au taux sous-gaussien du théorème 2.1.

Notez que cette procédure n'est pas valable simultanément pour tous les δ : il faut calculer un estimateur différent pour chaque valeur de δ. Ce problème peut être résolu en utilisant la

[1] Dans cette thèse, nous n'avons pas essayé d'optimiser les constantes, bien que ce soit un problème important et intéressant, voir la section 1.5.

méthode d'adaptation de Lepskii présentée dans Lepskiĭ (1990), comme expliqué plus en détail dans le chapitre 3.

Cette technique permet de traiter pleinement le cas unidimensionnel, ainsi que les données à queue lourde et la contamination adverse. Cependant, le développement rapide de l'apprentissage automatique et la quantité croissante de données hautement dimensionnelles disponibles ont conduit de nombreux statisticiens à se concentrer sur les tâches en grande dimension. Dans ces contextes, les données données ne sont pas des observations d'une variable aléatoire unidimensionnelle, mais plutôt des observations d'un vecteur aléatoire à d dimensions, avec $d \gg 1$. Le domaine de la statistique robuste s'est emparé de cette problématique, soulevant de nouvelles questions : quel taux peut-on atteindre dans ce cas ? Quel rôle joue la dimension ? Comment modifier et adapter les procédures pour ce cadre ? Nous notons que l'extension du résultat unidimensionnel n'est pas triviale car il existe plusieurs généralisations possibles de la médiane dans des configurations multidimensionnelles (par exemple la médiane géométrique, voir Minsker (2015) pour une définition, ou la médiane coordonnées par coordonnées).

2.2 Robustesse en haute dimension

2.2.1 Le taux subgaussien en haute dimension : découplage de la complexité et de la déviation

Quels taux est-il possible d'atteindre dans des configurations à haute dimension ? On peut toujours essayer d'atteindre un taux inspiré de la normalité asymptotique de la moyenne empirique des observations à second moment fini, qui est le taux atteint par la moyenne empirique lorsque les données sont un échantillon N de variables gaussiennes i.i.d., que nous appellerons encore *taux subgaussien*. Nous calculons ce taux en utilisant l'inégalité de Borell-TIS (voir (Ledoux, 2001, Théorème 7.1)) : si $Z_1, Z_2, ..., Z_N$ sont des variables gaussiennes indépendantes identiquement distribuées $\mathcal{N}(\mu, \mathrm{Id})$, il découle de cette inégalité qu'avec une probabilité d'au moins $1-\delta$,

$$\left\|\bar{Z}_N - \mu\right\|_2 = \sup_{\|v\|_2 \leq 1} \langle \bar{Z}_N - \mu, v\rangle \leq \mathbb{E} \sup_{\|v\|_2 \leq 1} \langle \bar{Z}_N - \mu, v\rangle + \gamma\sqrt{2\log(1/\delta)},$$

où $\gamma = \sup_{\|v\|_2 \leq 1} \sqrt{\mathbb{E}\langle \bar{Z}_N - \mu, v\rangle^2}$. Il est de connaissance élémentaire sur les distributions gaussiennes multivariées que $\mathbb{E}\sup_{\|v\|_2 \leq 1}\langle \bar{Z}_N - \mu, v\rangle \leq \sqrt{d/N}$ et $\gamma = \sqrt{1/N}$, ce qui conduit au taux subgaussien suivant (où C est une constante absolue),

$$\left\|\bar{Z}_N - \mu\right\|_2 \leq \left(\sqrt{\frac{d}{N}}) + \sqrt{\frac{2\log(1/\delta)}{N}}\right) := Cr_\delta. \tag{2.2}$$

La question de savoir s'il était possible d'atteindre un tel taux avec des données corrompues et à forte queue de distribution a été un problème ouvert pendant quelques années. Les premières tentatives d'adaptation de l'heuristique Median-Of-Mean, utilisant par exemple la médiane géométrique (également appelée point de Fermat) au lieu de la médiane unidimensionnelle, et présentées dans Minsker (2015) ou dans Hsu and Sabato (2016), ont conduit à des estimateurs $\hat{\mu}$ atteignant avec une probabilité supérieure à $1-\delta$,

$$\|\hat{\mu} - \mu\|^2 \leq \frac{d\log(1/\delta)}{N}, \tag{2.3}$$

où $\|\cdot\|$ est la norme euclidienne canonique sur $\mathbb{R}^d$. Cette limite est proportionnelle à $\log(1/\delta)$, donc les estimateurs sont en quelque sorte robustes aux queues lourdes, mais ils n'atteignent pas

tout à fait le taux subgaussien. En effet, dans le taux subgaussien, le terme de complexité d, qui capte le degré de complexité de l'espace ambiant, et le facteur dépendant de la défaillance $\log(1/\delta)$ ne sont pas multipliés, mais ajoutés, et sont donc en quelque sorte *découplés* l'un de l'autre.

Quelques années plus tard, l'article fondateur de Lugosi and Mendelson (2019c) a décrit le premier estimateur permettant d'atteindre le taux subgaussien en supposant uniquement un second moment fini, en utilisant l'heuristique Median-Of-Mean couplée à une procédure de tournoi. L'idée est d'utiliser une généralisation de l'inégalité de Hoeffding utilisée dans le théorème 2.1, appelée *Inégalité de différence bornée.* (également appelée inégalité de McDiarmid ou de Hoeffding/Azuma), que nous rappelons ici (voir par exemple le théorème 6.2 dans Boucheron et al. (2013)) :

Theorem 2.2 (Inégalité de McDiarmid)**.** *Considérons des variables aléatoires indépendantes* $X_1, \cdots, X_n \in E$ *et un mapping* $\psi : E^n \mapsto vers\mathbb{R}$. *Si pour tous* $i \in [\![1, n]\!]$ *et pour tous* $x_1, \cdots, x_n, x_i'$

$$|\psi(x_1, \cdots, x_i, \cdots, x_n) - \psi(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i \ ,$$

alors pour chaque $t > 0$,

$$\mathbb{P}\big(|\psi(X_1, \cdots, X_n) - \mathbb{E}[\psi(X_1, \cdots, X_n)]| \geq t\big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Cette inégalité se trouve dans McDiarmid (1997) et elle permet de traiter des fonctions ϕ plus générales que la simple somme de variables aléatoires. Dans le problème en question, nous utilisons cette inégalité pour lier l'équivalent en haute dimension de la quantité Z définie ci-dessus,

$$\tilde{Z} := \sup_{v \in \mathcal{S}_2^{d-1}} \left(\sum_{k \in [K]} \mathbf{1}(|\langle \overline{X}_k - \mu, v \rangle| > r) \right), \tag{2.4}$$

qui est cette fois un supremum sur toutes les directions possibles à d-dimensions, mais reste un supremum de quantité bornée. Le théorème 2.2 stipule que $\tilde{Z}$ a une probabilité exponentielle d'être proche de son espérance. Les étapes restantes de la preuve, qui sont détaillées dans le chapitre 3, consistent à limiter cette espérance $\mathbb{E}(\tilde{Z})$, en utilisant les outils de la théorie des processus empiriques.

Cet article pionnier, tout en répondant à une question ouverte, ouvre deux directions de recherche principales qui sont au cœur de cette thèse, et qui portent toutes deux sur des notions de complexité. La première concerne la *computational complexity* : pour calculer l'estimateur décrit dans Lugosi and Mendelson (2019c), il faut un nombre d'étapes qui croît exponentiellement avec la dimension, de sorte que cet estimateur ne peut pas être calculé en pratique, même pour des valeurs modérées de d. Notre première question est donc la suivante : *Pouvons-nous trouver des estimateurs subgaussiens traitables en haute dimension ?* Les chapitres 3 et 4 traitent principalement de cette question. La deuxième question traite de la complexité statistique au sens large et demande *s'il est toujours possible de trouver des procédures qui atteignent un taux gaussien* pour d'autres tâches d'estimation, comme l'estimation par rapport à des normes non-euclidiennes par exemple. Les chapitres 5, 6 et 7 suivent cette deuxième direction.

2.2.2 Complexité informatique

Le premier type de complexité que cette thèse explore est une complexité computationnelle. Diakonikolas et al. (2016) est l'un des premiers articles à soulever la question de la tractabilité de l'estimation robuste de la moyenne en haute dimension, et à montrer que l'apprentissage robuste

en haute dimension est *algorithmiquement* possible. Avant cet article pionnier, les considérations calculatoires étaient principalement des résultats de dureté, montrant que les estimateurs robustes traditionnels tels que la médiane de Tukey, bien que prouvés robustes, sont prouvés difficiles à calculer (voir par exemple Johnson and Preparata (1978) ou Bernholt (2006)). En utilisant de nouvelles techniques, Diakonikolas et al. (2016) ouvre la voie à la statistique robuste calculable et traçable et laisse de nombreuses questions passionnantes à explorer. En effet, même si les estimateurs proposés dans Diakonikolas et al. (2016) sont robustes à la contamination adverse, ils échouent avec une probabilité constante (par exemple 1/100), et n'atteignent pas le *taux subgaussien* mentionné ci-dessus. Une observation similaire peut être faite sur les travaux de Minsker (2015) ou Hsu and Sabato (2016). Dans ces deux articles, les estimateurs sont très rapides à calculer : la médiane géométrique de la moyenne peut être calculée aussi rapidement que la moyenne empirique a des facteurs multiplicatifs logarithmiques près en la dimension. Cependant, comme nous l'avons souligné précédemment, les estimateurs proposés ne parviennent pas à atteindre le taux subgaussien. Une grande partie de cette thèse se concentre sur le développement de procédures qui sont efficaces en termes de calcul et qui atteignent le taux subgaussien.

2.2.3 Complexité statistique

Une deuxième question à laquelle ce travail s'attaque est de savoir s'il est encore possible de trouver des estimateurs qui se comportent comme si les données étaient gaussiennes pour d'autres tâches d'estimation, ou si cela n'est possible que dans quelques sous-cas particuliers comme l'estimation de la moyenne par rapport à la norme euclidienne habituelle.

Illustrons cette question par un exemple. Nous voulons maintenant estimer la moyenne par rapport à une norme sparse : nous désignons l'ensemble des vecteurs s-sparse $\mathcal{U}_s$, et considérons la norme suivante :

$$\|a\|_s = \sup_{u \in \mathcal{U}_s, \|u\|_2 = 1} \langle a, u \rangle .$$

En un sens, nous cherchons à remplacer le supremum sur la boule euclidienne $\mathcal{B}_2^d$ par un supremum sur un autre ensemble, ici $\mathcal{U}_s \cap \mathcal{B}_2^d$. On voit, en reprenant les variables gaussiennes Z_i, qu'avec une probabilité d'au moins $1 - \delta$, (toujours d'après (Ledoux, 2001, Theorem 7.1))

$$\begin{aligned} \left\| \bar{Z}_N - \mu \right\|_s &= \sup_{\|v\|_2 \leq 1, v \in \mathcal{U}_s} \langle \bar{Z}_N - \mu, v \rangle \\ &\leq \mathbb{E} \sup_{\|v\|_2 \leq 1, v \in \mathcal{U}_s} \langle \bar{Z}_N - \mu, v \rangle + \gamma \sqrt{2 \log(1/\delta)} \\ &\leq C \left(\frac{s \log(ed/s)}{N} + \frac{\log(1/\delta)}{N} \right)^{1/2} . \end{aligned}$$

Le taux subgaussien pour le problème d'estimation de moyenne sparse de la dernière equation est différent de (2.2) : le "terme d'écart" contenant la probabilité d'échec δ reste inchangé, mais le "terme de complexité" (celui qui ne dépend pas de δ) passe de d à $s \log(ed/s)$. Ce taux peut-il être atteint en supposant un moment du second ordre sur les variables aléatoires ? Bien que cette question soit traitée dans cette thèse pour le cas particulier du supremum sur $\mathcal{U}_s \cap \mathcal{B}_2^d$, sa généralisation à d'autres ensembles V reste une question ouverte, qui n'est que partiellement traitée dans les chapitres 5, 6 et 7. Nous pouvons reformuler cette question d'une autre manière : le terme de complexité issu de la perspective gaussienne est-il atteignable pour des variables non gaussiennes, ou existe-t-il une autre mesure de complexité intrinsèquement liée à la situation non gaussienne ?

2.3 Contributions principales

Après avoir introduit les principaux concepts, outils et questions, nous présentons les différentes contributions des cinq travaux qui composent cette thèse.

2.3.1 Estimation subgaussienne robuste d'un vecteur moyen en temps quasi-linéaire

La première contribution de cette thèse, détaillée dans le chapitre 3, porte sur la complexité computationnelle de l'estimation subgaussienne de la moyenne, par rapport à la norme euclidienne traditionnelle. Au début de cette thèse, deux articles très importants ont été publiés, qui ont été les premiers à proposer des procédures atteignant le taux subgaussien (1.2) tout en fonctionnant en temps polynomial dans les deux variables N et d : Hopkins (2018) et Cherapanamjeri et al. (2019). Ils s'exécutent tous deux en temps polynomial : $\mathcal{O}(N^{24} + Nd)$ pour Hopkins (2018) et $\mathcal{O}(N^4 + N^2 d)$ pour Cherapanamjeri et al. (2019). (voir Cherapanamjeri et al. (2019) pour plus de détails sur ces temps d'exécution). Ils ne considèrent pas une contamination adversariale du jeu de données alors que leurs résultats s'étendent facilement à cette configuration. Le premier chapitre de cette thèse propose donc le troisième algorithme polynomial atteignant le taux subgaussien pour l'estimation de la moyenne, et le premier à traiter explicitement la contamination adverse. De plus, il améliore le temps d'exécution des premières procédures proposées : nous construisons un algorithme s'exécutant en temps $\tilde{\mathcal{O}}(Nd + u\log(1/\delta)d)$ qui produit un estimateur de la vraie moyenne atteignant le taux subgaussien (1.2) avec une confiance de $1 - \delta - (1/10)^u$ (pour $\exp(-c_0 N) \leq \delta \leq \exp(-c_1|\mathcal{O}|)$) sur une base de données corrompue et sous une hypothèse de second moment uniquement. Dans le pire des cas, le temps d'exécution est donc de $\tilde{\mathcal{O}}(N^2 d)$.

Pour ce faire, cet article utilise l'heuristique de la médiane de la moyenne, comme les deux procédures précédentes en temps polynomial. Notre approche reprend en fait des idées de deux communautés : le principe de la médiane des moyennes de la communauté des statistiques et une relaxation de la programmation semi-définie (SDP) utilisée dans la communauté des sciences informatiques par Cheng et al. (2019a) qui peut être théoriquement calculée rapidement. L'amélioration du temps de calcul par rapport à la procédure de Cherapanamjeri et al. (2019) est due à l'utilisation de ce programme semi-défini (SDP) couvrant étudié dans Allen-Zhu et al. (2014), Peng et al. (2012), et Cheng et al. (2019a), et popularisé dans le domaine de la statistique robuste par Cheng et al. (2019a), à chaque itération de l'algorithme de descente du gradient robuste. Alors que dans Cheng et al. (2019a) ce SDP conduit à des procédures échouant avec une probabilité constante et donc à l'impossibilité d'atteindre le taux subgaussien, nous montrons que l'utilisation de ce type de SDP sur les moyennes de blocs plutôt que sur les données elles-mêmes conduit non seulement à atteindre le taux subgaussien, mais aussi à une amélioration du coût de calcul de l'algorithme. L'heuristique de la médiane de la moyenne présente donc dans ce cas un avantage stochastique et un avantage computationnel.

Pour prouver que la procédure proposée atteint effectivement le taux subgaussien, nous avons dû trouver un nouveau lemme stochastique intéressant en soi, généralisant celui de Lugosi and Mendelson (2019c) et Lerasle et al. (2019). Ce lemme a été utilisé depuis dans divers autres travaux et contextes, par exemple dans la régression dans Cherapanamjeri et al. (2020b), ou pour définir la notion de stabilité dans Diakonikolas et al. (2020). La preuve de ce lemme repose sur une technique d'arrondi gaussien similaire à celle utilisée dans l'inégalité de Grothendieck.

Des travaux très récents Lei et al. (2020); Hopkins et al. (2020); Depersin (2020a) obtiennent des résultats similaires à celui de ce travail. Ils ont également réussi à remplacer les SDP par des méthodes spectrales pour le calcul d'une direction de descente robuste à chaque étape. En

effet, même si les SDP de couverture sont d'un point de vue théorique efficaces d'un point de vue informatique (voir Allen-Zhu et al. (2014),Peng et al. (2012)), ils sont notoirement difficiles à mettre en œuvre en pratique alors que les méthodes spectrales utilisées dans Lei et al. (2020); Hopkins et al. (2020); Depersin (2020a) ouvrent la porte à des algorithmes implémentables. Il est intéressant de noter que le temps de calcul proposé dans ce travail est encore à ce jour le meilleur temps d'exécution connu, et on peut conjecturer qu'il pourrait même être le moyen le plus rapide d'atteindre le taux subgaussien.

2.3.2 Un algorithme spectral pour la régression robuste avec des taux sub-gaussiens

La deuxième contribution que nous apportons, détaillée dans le chapitre 4 traite de la question de l'atteinte des limites subgaussiennes en temps polynomial, mais pour la régression au lieu de l'estimation de la moyenne. Nous rappelons rapidement le cadre standard de la régression linéaire où les données sont des couples $(X_i, Y_i)_i \in \mathbb{R}^d \times \mathbb{R}$ et où l'on cherche la meilleure combinaison linéaire des coordonnées d'un vecteur d'entrée X pour prédire la sortie Y, c'est-à-dire que l'on cherche β^* défini comme suit.

$$\beta^* = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \ell(\beta) = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \mathbb{E}(Y_1 - \langle \beta, X_1 \rangle)^2.$$

La question de savoir s'il était même possible d'atteindre des taux sub-gaussiens dans ce cadre, sous des hypothèses de moment faible, est restée longtemps ouverte. En effet, pendant un certain temps, les algorithmes polynomiaux les plus connus étaient ceux de Prasad et al. (2018) ou de Hsu and Sabato (2016). La garantie pour ces deux algorithmes est la suivante : lorsque la covariance de X est l'identité et lorsque le bruit $\xi = Y - \langle \beta^*, X \rangle$ a une variance bornée, $\ell(\hat{f}) - \ell(f^*) \leq \mathcal{O}(\frac{\log(1/\delta)d}{N})$ avec une probabilité de $1-\delta$. Ce taux ne présente pas ce découplage entre complexité et déviation que nous avons appelé sub-gaussien. L'article de Cherapanamjeri et al. (2020a) a été le premier à construire une méthode en temps polynomial permettant d'atteindre le taux subgaussien des MCO dans le cadre gaussien $\ell(\hat{f}) - \ell(f^*) \leq \mathcal{O}(\frac{\log(1/\delta) \vee d}{N})$. Lors de la première publication du chapitre 4, Cherapanamjeri et al. (2020a) était la seule procédure fonctionnant en algorithme polynomial atteignant le taux subgaussien optimal. Cependant, Cherapanamjeri et al. (2020a) utilise la hiérarchie de programmation Sum of Square (SoS) pour concevoir son algorithme. Même si la hiérarchie SoS s'exécute en temps polynomial, sa dépendance à l'égard de la résolution de grands programmes semi-définis la rend peu pratique et reste un résultat théorique, laissant toujours ouverte la question de l'existence d'un algorithme pratique efficace permettant d'obtenir des taux subgaussiens optimaux.

Dans le chapitre 4, nous abordons cette question, en montrant que les techniques de Lei et al. (2020) combinées aux lemmes de Depersin (2020a) peuvent être utilisées pour donner le premier algorithme pratique, presque quadratique (et en fait dans la plupart des cas presque linéaire) qui atteint le taux subgaussien. Nous réalisons également des expériences numériques sur des données simulées avec la procédure que nous proposons pour montrer qu'elle est effectivement pratique et rapide. De plus, comme prévu par nos résultats théoriques, notre analyse de simulation montre qu'elle est robuste à la fois aux données à queue lourde et aux valeurs aberrantes. À notre connaissance, c'est la première fois que des expériences numériques mettant en œuvre la formulation exacte d'un estimateur sous-gaussien sont menées pour un algorithme de régression avec des taux sous-gaussiens et des garanties de temps polynomial.

2.3.3 Estimation subgaussienne robuste avec dimension VC

La troisième contribution que nous apportons est détaillée dans le chapitre 5, et elle concerne la complexité statistique plutôt que la complexité informatique. Nous avons essayé de montrer comment on peut utiliser la dimension VC pour obtenir des limites de pointe dans l'estimation non-euclidienne avec des données à forte queue.

Dans ce travail, nous montrons que l'analyse présentée dans Lugosi and Mendelson (2019c), dans Lecué and Lerasle (2019), Lerasle (2019) ou dans Lecué and Lerasle (2020), et généralisée dans Lugosi and Mendelson (2019b), toutes basées sur le principe de la médiane des moyennes et l'utilisation des complexités de Rademacher, peuvent être modifiées afin d'obtenir des taux sub-gaussiens pour des problèmes épars ou structurés en supposant uniquement des moments d'ordre deux bornés. La méthode développée dans Lerasle (2019) ou dans Lecué and Lerasle (2020) nécessite que les données aient au moins $\log(d)$ de moments finis (où d est la dimension de l'espace) afin d'exploiter la sparsité du problème et n'offre aucune garantie sans cette condition, et est à ce jour la meilleure connue. Nous montrons que nous pouvons abandonner cette condition en introduisant judicieusement la dimension VC dans les différentes preuves, et exploiter la sparsité du problème *avec seulement deux moments*. Nous montrons au chapitre 5 que les approches classiques utilisant les complexités locales de Rademacher ne peuvent pas atteindre ce type de limites subgaussiennes sous l'hypothèse d'un second moment seulement. D'une certaine manière, l'approche classique utilisée jusqu'à présent ne capture pas la complexité statistique correcte des problèmes à haute dimension sous des hypothèses structurelles à basse dimension et sous une hypothèse de second moment seulement : **il semble que la complexité de Rademacher ne soit pas la bonne façon de mesurer la complexité du problème de l'estimation de la moyenne structurée dans une norme quelconque**. Notre approche basée sur la dimension VC permet de surmonter ce problème et d'aller au-delà de cette hypothèse de moments subgaussiens $\log d$ qui est apparue dans tous les travaux sur l'estimation robuste et subgaussienne dans le cadre de la haute dimension Lerasle (2019). Nous montrons également que cette technique générale peut être facilement reproduite et nous donnons de nouveaux estimateurs robustes qui atteignent des limites de pointe pour différentes tâches d'estimation telles que la régression, l'estimation de la moyenne avec des normes non euclidiennes, l'estimation robuste de matrices à faible rang et l'estimation de la covariance.

Ce chapitre n'est pas le premier à introduire la dimension VC dans les problèmes d'estimation robuste : il a été inspiré par Chen et al. (2018) et Gao (2017) par exemple. Dans ces deux articles, l'estimation et la régression avec une éventuelle structure de sparsité et des valeurs aberrantes sont également réalisées avec des taux optimaux, en utilisant des techniques de dimension VC, mais leur hypothèse d'une structure de sparsité et de valeurs aberrantes n'a pas été retenue.

Nous notons que la dimension VC présente certains avantages par rapport à la complexité de Rademacher dans certains cas, mais ce quatrième chapitre montre qu'elle ne conduit pas à des taux optimaux dans tous les cas, et donc qu'elle n'est pas toujours la bonne façon de mesurer la complexité d'une tâche d'estimation robuste. En effet, en utilisant la dimension VC dans l'estimation de la moyenne, nous perdons une dépendance intéressante des bornes de risque dans la structure de covariance : nos taux pour l'estimation de la moyenne (non éparse) dépendent de la dimension ambiante d au lieu du rang effectif $\mathrm{Tr}(\Sigma)/||\Sigma||_{op}$ (qui est capturé par la complexité de Rademacher). En particulier, l'approche générale de la dimension VC ne se généralise pas directement aux espaces de dimension infinie. Dans la dernière section du chapitre 5, nous montrons que ce problème peut être surmonté si nous avons une certaine connaissance de la matrice de covariance, en proposant une nouvelle procédure pour ce cas.

2.3.4 Sur la robustesse de la médiane des moyennes de Stahel-Donoho à la corruption contradictoire et aux données à queue lourde

Dans le chapitre 6, nous traitons de l'estimation par rapport à la norme $x \in \mathbb{R}^d \to \left\|\Sigma^{-1/2}x\right\|_2$, où Σ est la matrice de covariance des données, et nous supposons que cette matrice n'est pas connue au préalable par le statisticien. Ainsi, en un sens, nous essayons d'estimer par rapport à une norme inconnue, alors que les techniques dérivées dans le chapitre 5 s'appliquent principalement à des normes connues. Nous étudions cette norme particulière, dont la boule unitaire est l'ellipsoïde $\Sigma^{1/2}B_2^d$, parce qu'il s'agit de la meilleure métrique – c'est-à-dire celle qui conduit à des ensembles de confiance de volume minimal pour une confiance donnée – dans le cas de référence i.i.d. gaussien.

Alors que le taux subgaussien pourrait être obtenu avec les estimateurs de Lugosi and Mendelson (2019b) ou Lerasle et al. (2019), ces estimateurs nécessiteraient la connaissance de Σ dans leur construction. Il faut donc envisager d'autres techniques. Dans ce chapitre, nous montrons que c'est possible grâce à une notion de profondeur/extrémité introduite au début des années 80 qui utilise une normalisation par une estimation robuste de l'échelle appelée *Stahel-Donoho Outlyingness* (SDO), qui a été introduite pour la première fois dans Donoho and Gasko (1992). Nous couplons cette notion d'excentricité avec l'heuristique Median-Of-Mean pour obtenir des estimateurs subgaussiens par rapport à la métrique $\left\|\Sigma^{-1/2}.\right\|_2$. Sur le chemin de notre objectif, nous complétons les résultats sur la cohérence $\sqrt{n}$ et la normalité asymptotique des estimateurs de Stahel-Donoho que l'on peut trouver dans Maronna and Yohai (1995) et Tyler (1994) en dérivant le premier taux de convergence non-asymptotique pour la médiane SDO (ainsi que sa version médiane des moyennes). Nous montrons également que les propriétés de robustesse de la médiane SDO originale et de sa version MOM vont au-delà du modèle de contamination de Huber et qu'elles persistent toujours dans le modèle de corruption adversariale de Setting 2.1. Nous utilisons également l'échelonnement robuste de l'écart de Stahel-Donoho pour construire des estimateurs de la matrice de covariance sous une certaine hypothèse de régularité.

2.3.5 Estimation robuste optimale de la moyenne et de l'emplacement via des programmes convexes en respectant des pseudo-normes quelconques

Notre dernière contribution, détaillée dans le chapitre 7, consiste à donner une nouvelle borne inférieure pour l'estimation de la moyenne dans toute norme. Lugosi and Mendelson (2019b) donne la borne inférieure suivante pour l'estimation de la moyenne :

Theorem 2.3. *[Théorème 3 de Lugosi and Mendelson (2019b)] Il existe une constante absolue $c > 0$ telle que ce qui suit est vrai. Si $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ est un estimateur tel que pour tout $\mu^* \in \mathbb{R}^d$ et tout $\delta \in (0, 1/4)$,*

$$\mathbb{P}_{\mu^*}^N \left[\|\hat{\mu} - \mu^*\| \leq r^*\right] \geq 1 - \delta$$

où $\mathbb{P}_{\mu^*}^N$ est la distribution de probabilité de $(X_i)_{i\in[N]}$ lorsque les X_i sont i.i.d. $\mathcal{N}(\mu^*, \Sigma)$ alors

$$r^* \geq \frac{c}{\sqrt{N}} \left(\sup_{\eta>0} \eta \sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)} + \sup_{v\in B^\circ} \left\|\Sigma^{1/2}v\right\|_2 \sqrt{\log(1/\delta)} \right)$$

où $N(\Sigma^{1/2}B^\circ, \eta B_2^d)$ est le nombre minimal de translations de ηB_2^d nécessaires pour couvrir $\Sigma^{1/2}B^\circ$.

Le terme de complexité dans cette borne inférieure est donc mesuré en utilisant la borne de Sudakov définie dans le 1.3. Cependant, il existe un écart entre cette borne inférieure et les bornes supérieures dépendant de la largeur moyenne gaussienne dans le cas gaussien. Cet écart

provient du manque de rigueur de l'inégalité de Sudakov présentée dans le théorème 1.3. Pour les ellipsoïdes par exemple, la limite de Sudakov n'est pas nette en général et donc la limite inférieure du théorème 2.3 ne permet pas de retrouver le taux subgaussien classique pour le cas de la norme euclidienne standard (c'est-à-dire pour $S = B_2^d$) qui est donné dans Lugosi and Mendelson (2019c) par :

$$\sqrt{\frac{\mathrm{Tr}\,(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} \log(1/\delta)}{N}}. \tag{2.5}$$

En effet, lorsque $\|\cdot\|$ est la norme euclidienne ℓ_2^d, alors $\mathbb{E}\left\|\Sigma^{1/2}G\right\| = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \sim \sqrt{\mathrm{Tr}(\Sigma)}$. En revanche, l'entropie de $\Sigma^{1/2}B^\circ = \Sigma^{1/2}B_2^d$ par rapport à ηB_2^d peut être calculée en utilisant l'équation (5.45) dans Pisier (1989) que

$$\sup_{\eta>0} \eta\sqrt{\log_2 N(\Sigma^{1/2}B_2^d, \eta B_2^d)} \sim \sup_{n\geq 1, k\in[d]} \frac{\sqrt{n}}{2^{n/k}} \left|\prod_{j=1}^k \sqrt{\lambda_j}\right|^{1/k} \sim \sqrt{\sup_{k\in[d]} k \left|\prod_{j=1}^k \lambda_j\right|^{1/k}} \tag{2.6}$$

où $\lambda_1 \geq \ldots \geq \lambda_d$ sont les valeurs singulières de Σ. En particulier, lorsque $\lambda_j = 1/j$, la borne d'entropie (1.9) est de l'ordre d'une constante alors que la largeur moyenne gaussienne est de l'ordre de $\sqrt{\log d}$. Nous comblons cette lacune dans le chapitre 7 en montrant une limite inférieure où l'entropie est remplacée par la largeur moyenne gaussienne (plus grande). Pour ce faire, nous utilisons le lemme d'Anderson et des arguments analytiques, au lieu des arguments géométriques et volumétriques utilisés par Lugosi and Mendelson (2019b) pour obtenir la limite de Sudakov comme limite inférieure.

Nous montrons également que ce taux peut parfois être atteint par une solution à un problème d'optimisation convexe dans le cadre adversatif et L_2 à queue lourde en considérant le minimum de certaines transformées de Fenchel-Legendre construites en utilisant le principe de la médiane des moyennes.

CHAPTER 3

Robust Subgaussian Estimation of a Mean Vector in Nearly Linear Time

Contents

3.1 Thorough introduction on the robust mean vector estimation problem

Estimating the mean of a random variable in a d-dimensional space when given some of its realizations is arguably the oldest and most fundamental problem of statistics. In the past few years, it has received important attention from two communities: the statistics community, see for instance Catoni (2012); Minsker (2015); Chen et al. (2018); Catoni and Giulini (2017); Lugosi and Mendelson (2019c); Minsker (2018a); M. Lerasle and Lecué (2017); Hopkins (2018); Cherapanamjeri et al. (2019); Lei et al. (2020); Dalalyan and Minasyan (2020) and the computer science one, see Diakonikolas et al. (2016, 2019a, 2018a,b, 2019c); Cheng et al. (2019a); Diakonikolas and Kane (2019); Hopkins et al. (2020). Both communities consider the problem of *robust mean estimation*, focusing mainly on different definitions of robustness.

In recent years, many efforts have been made by the statistics community on the construction of estimators performing in a *subgaussian way* for heavy-tailed data. As seen in the introduction, such estimators achieve the same statistical properties as the empirical mean $\bar{X}_N$ of $(X_1, \cdots, X_N)$, a N-sample of i.i.d. Gaussian variables $\mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ and $\Sigma \succeq 0$ is the covariance matrix. In that case, for a given confidence $1-\delta$, the subgaussian rate as defined in Lugosi and Mendelson (2019c) is (up to an absolute multiplicative constant)

$$r_\delta = \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} \log(1/\delta)}{N}} \tag{3.1}$$

where $\mathrm{Tr}(\Sigma)$ is the trace of Σ and $\|\Sigma\|_{op}$ is the operator norm of Σ. Indeed, it follows from Borell-TIS's inequality (see Theorem 7.1 in Ledoux (2001) or pages 56-57 in Ledoux and Talagrand (2011)) that with probability at least $1-\delta$,

$$\left\|\bar{X}_N - \mu\right\|_2 = \sup_{\|v\|_2\leq 1} \langle \bar{X}_N - \mu, v\rangle \leq \mathbb{E}\sup_{\|v\|_2\leq 1} \langle \bar{X}_N - \mu, v\rangle + \sigma\sqrt{2\log(1/\delta)}$$

where $\sigma = \sup_{\|v\|_2\leq 1}\sqrt{\mathbb{E}\langle \bar{X}_N - \mu, v\rangle^2}$ is the weak variance of the Gaussian process. It is straightforward to check that $\mathbb{E}\sup_{\|v\|_2\leq 1}\langle \bar{X}_N - \mu, v\rangle \leq \sqrt{\mathrm{Tr}(\Sigma)/N}$ and $\sigma = \sqrt{\|\Sigma\|_{op}/N}$, which leads to the rate in (3.1) (up to the constant $\sqrt{2}$ on the second term in (3.1)). In most of the recent works, the effort has been made to achieve the rate r_δ for i.i.d. heavy-tailed data even under the minimal requirement that the data only have a second moment. Under this second-moment assumption only, the empirical mean cannot[1] achieve the rate (3.1) and one needs to consider other procedures. As, recalled in the general introduciton, over the years, some procedures have been proposed to achieve such a goal: it started with Catoni (2012) and Lerasle and Oliveira (2011), then, a Le Cam test estimator, called a tournament estimator in Lugosi and Mendelson (2019c), a minmax median-of-means estimator in M. Lerasle and Lecué (2017) and a PAC-Bayesian estimator in Catoni and Giulini (2017) were constructed. The constructions in Lerasle and Oliveira (2011); Lugosi and Mendelson (2019c); M. Lerasle and Lecué (2017) are based on the median-of-means principle, a technique that we will also use.

On the other side, the computer science (CS) community mostly considers a different definition of robustness and targets a different goal. In many recent CS papers, tractable algorithms (and not only theoretical estimators) have been constructed and proved to be robust with respect to *adversarial contamination* of the dataset seen in the general introduction. We recall now this adversarial contamination model together with the heavy-tailed setup which will serve as our unique assumption in this work.

Assumption 3.1. *There exists N random vectors $(\tilde{X}_i)_{i=1}^N$ in $\mathbb{R}^d$ which are independent with mean μ and covariance matrix $\mathbb{E}(\tilde{X}_i-\mu)(\tilde{X}_i-\mu)^\top \preceq \Sigma$ where Σ is an unknown covariance matrix. The N random vectors $(\tilde{X}_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives the modified dataset $(X_i)_{i=1}^N$ to the statistician. Hence, the statistician receives an "adversarially" contaminated dataset of N vectors in $\mathbb{R}^d$ which can be partitioned into two groups: the modified data $(X_i)_{i\in\mathcal{O}}$, which can be seen as outliers and the "good data" or inliers $(X_i)_{i\in\mathcal{I}}$ such that $\forall i\in\mathcal{I}, X_i = \tilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O}\cup\mathcal{I} = \{1,\dots,N\}$ is unknown to the statistician.*

In the adversarial contamination model from Assumption 3.1, the set $\mathcal{O}$ can depend arbitrarily on the initial data $(\tilde{X}_i)_{i=1}^N$; the corrupted data $(X_i)_{i\in\mathcal{O}}$ can have any arbitrary dependance structure; and the informative data $(X_i)_{i\in\mathcal{I}}$ may also be correlated (for instance, it is the case, in general, when the $|\mathcal{O}|$ data $\tilde{X}_i$ with largest ℓ_2^d-norm are modified by the adversary). The computer science community looks at the problem of robust mean estimation from algorithmic perspectives such as the running time in this contamination model. A typical result in this line of research is Theorem 1.3 from Cheng et al. (2019a) that we recall now.

Theorem 3.1 (Theorem 1.3, Cheng et al. (2019a))**.** *Let $X_1,\dots,X_N$ be a data points in $\mathbb{R}^d$ following Assumption 3.1. We assume that the covariance matrix Σ of the inliers satisfies $\Sigma\preceq\sigma^2 I_d$. We assume that $\epsilon = |\mathcal{O}|/N$ is such that $0<\epsilon<1/3$ and $N\gtrsim d\log(d)/\epsilon$. There exists*

[1] Under only a second-moment assumption, the empirical mean achieves the rate $\sqrt{\mathrm{Tr}(\Sigma)/(\delta N)}$ which can not be improved in general, see Catoni (2012).

an algorithm running in $\tilde{\mathcal{O}}(Nd)/\mathrm{poly}(\epsilon)$ *which outputs* $\hat{\mu}_\epsilon$ *such that with probability at least* $9/10$, $\|\hat{\mu}_\epsilon - \mu\|_2 \lesssim \sigma\sqrt{\epsilon}$.

The notation $\tilde{\mathcal{O}}(Nd)$ stands for the computational running time of an algorithm up to $\log(Nd)$ factors. The first result proving the existence of a polynomial time algorithm robust to adversarial contamination may be found in Diakonikolas et al. (2016) and the first achieving such a result under only a second moment assumption may be found in Diakonikolas et al. (2017). Theorem 3.1 improves upon many existing results since it achieves the optimal information theoretic-lower bound with a (nearly) linear-time algorithm.

Finally, there are two recent papers for which both algorithmic and statistical considerations are important. In Hopkins (2018); Cherapanamjeri et al. (2019), algorithms achieving the subgaussian rate in (3.1) have been constructed. They both run in polynomial time: $\mathcal{O}(N^{24}+Nd)$ for Hopkins (2018) and $\mathcal{O}(N^4+N^2d)$ for Cherapanamjeri et al. (2019) (see Cherapanamjeri et al. (2019) for more details on these running times). They do not consider a contamination of the dataset even though their results easily extend to this setup. Some other estimators which have been proposed in the statistics literature are very fast to compute but they do not achieve the optimal subgaussian rate from (3.1). A typical example is Minsker's geometric median estimator Minsker (2015) which achieves the rate $\sqrt{\mathrm{Tr}(\Sigma)\log(1/\delta)/N}$ in linear time $\tilde{\mathcal{O}}(Nd)$. All the later three papers use the median-of-means principle. We will also use this principle. What we mainly borrow from the literature on MOM estimators is the advantage to work with local block means instead of the data themselves. We will identify two such advantages by doing so: a stochastic one and a computational one (see Remark 3.4 below for more details).

The aim of this work is to show that a single algorithm can answer the three problems: robustness to heavy-tailed data, to adversarial contamination and computational cost. Assumption 3.1 covers the two concepts of robustness considered in the statistics and computer science communities since the *informative data* (data indexed by $\mathcal{I}$) are only assumed to have a second moment and there are $|\mathcal{O}|$ adversarial outliers in the dataset. Our aim is to show that the rate of convergence (3.1) which is the rate achieved by the empirical mean in the ideal i.i.d. Gaussian case can be achieved in the corrupted and heavy-tailed setup from Assumption 3.1 with a fast algorithm: we construct an algorithm running in time $\tilde{\mathcal{O}}(Nd + u\log(1/\delta)d)$ which outputs an estimator of the true mean achieving the subgaussian rate (3.1) with confidence $1-\delta-(1/10)^u$ (for $\exp(-c_0N) \le \delta \le \exp(-c_1|\mathcal{O}|)$) on a corrupted database and under a second moment assumption only. It is therefore robust to heavy-tailed data and to contamination. Our approach takes ideas from both communities: the median-of-means principle which has been recently used in the statistics community and a SDP relaxation from Cheng et al. (2019a) which can be theoretically computed fast. The baseline idea is to construct K equal size groups of data from the N given ones and to compute their empirical means $\bar{X}_k, k = 1, \dots, K$. These K empirical means are used successively to find a robust descent direction thanks to a SDP relaxation from Cheng et al. (2019a). We prove the robust subgaussian statistical property of the resulting descent algorithm under only the Assumption 3.1.

The chapter is organized as follows. In the next section, we give a high-level description of the algorithm and summarize its statistical and computation performance in our main result Theorem 3.2. We also clearly identify how it improves upon existing results on the same subject. In Section 3, we prove its statistical properties and give a precise definition of the algorithm. In Section 4, we study the statistical performance of the SDP relaxation at the heart of the descent direction. In Section 5, we fully characterize its computational cost. In Section 3.6, we construct a procedure achieving the same statistical properties and can automatically adapt to the number of outliers. This latter adaptive procedure is also proved to satisfy estimation results in expectation.

We will use the following notation $[n] = \{1, \dots, n\}$ for any $n \in \mathbb{N}$ and ℓ_2^d stands for the Euclidean space $\mathbb{R}^d$ endowed with its canonical Euclidean norm $\|\cdot\|_2 : x = (x_j)_{j=1}^d \in \mathbb{R}^d \to \left(\sum_j x_j^2\right)^{1/2}$. A ℓ_2^d-ball centered in $x \in \mathbb{R}^d$ with radius $r > 0$ is denoted by $B_2^d(x, r)$, the ℓ_2^d unit ball is denoted by B_2^d and the ℓ_2^d unit sphere is denoted by $\mathcal{S}_2^{d-1}$.

3.2 Construction of the algorithms and main result

The construction of our robust subgaussian descent procedure is using two ideas. The first one comes from the median-of-means (MOM) approach which has recently received a lot of attention in the statistical and machine learning communities, see for instance Bubeck et al. (2013); Lerasle and Oliveira (2011); Devroye et al. (2016); Minsker and Strawn (2017); Minsker (2015); Nemirovsky and Yudin (1983); Alon et al. (1999); Jerrum et al. (1986); Birgé (1984). The MOM approach often yields robust estimation strategies (but usually at a high computational cost). Let us recall the general idea behind that approach already exposed in the introduction: we first randomly split the data into K equal-size blocks $B_1, \dots, B_K$ (if K does not divide N, we just remove some data). We then compute the empirical mean within each block: for $k = 1, \dots, K$,

$$\bar{X}_k = \frac{1}{|B_k|} \sum_{i \in B_k} X_i$$

where we set $|B_k| = \mathrm{Card}(B_k) = N/K$. In the one-dimensional case, we then take the median of the latter K empirical means to construct a robust *and subgaussian* estimator of the mean (see Devroye et al. (2016)). It is more complicated in the multi-dimensional case, where there is no *definitive* equivalent of the one dimensional median but instead there are several candidates: coordinate-wise median, the geometric median (also known as Fermat point), the Tukey Median, among many others (see Small (1990)). The strength of this approach is the robustness of the median operator, which leads to good statistical properties even on corrupted databases. For the construction of our algorithm, we use the idea of grouping the data and compute iteratively some median of the bucketed means $\bar{X}_k, k = 1, \dots, K$.

In Cherapanamjeri et al. (2019), the authors propose to use these block means for a gradient descent algorithm: at the current point x_c of the iterative algorithm, a "robust descent direction" well aligned with $x_c - \mu$ is constructed with high probability. Note that $x_c - \mathbb{E}X$ is the best descent direction towards $\mathbb{E}X$ starting from x_c; we can also re-write that as a matrix problem: a top eigenvector (i.e. an eigenvector associated with the largest singular value) of $(\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top$ is the optimal descent direction $(x_c - \mathbb{E}X)/\|x_c - \mathbb{E}X\|_2$. As a consequence, a top eigenvector of a solution to the optimization problem

$$\underset{M \succeq 0, \mathrm{Tr}(M)=1}{\mathrm{argmax}} \left\langle M, (\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top \right\rangle \tag{3.2}$$

also yields the best descent direction we are looking for (note that $\langle A, B \rangle = \mathrm{Tr}(A^\top B)$ is the inner product between two matrices A and B). Optimization problem (3.2) may be seen as a SDP relaxation for the problem of finding a top eigenvector and it is the reason why we go into SDP optimization techniques. Recently, this SDP relaxation has been bypassed thanks to the power method in Lei et al. (2020) whose aims is also to approximate a top eigenvector.

Of course, we don't know $(\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top$ in (3.2) but we are given a database of N data $X_1, \dots, X_N$ (among which $|\mathcal{I}|$ of them have mean μ). We use these data to estimate in a robust way the unknown quantity $(\mathbb{E}X - x_c)(\mathbb{E}X - x_c)^\top$ in (3.2). Ideally, we would like to identify the *informative data* and their block means $(1/|\mathcal{K}|) \sum_{k \in \mathcal{K}} (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top$, where $\mathcal{K} = \{k : B_k \bigcap \mathcal{O} = \emptyset\}$, to estimate this quantity but this information is not available either.

To address this problem we use a tool introduced in Cheng et al. (2019a); Diakonikolas et al. (2016) adapted to the block means. The idea is to endow each block mean $\bar{X}_k$ with a weight ω_k taken in Δ_K defined as

$$\Delta_K = \left\{ (\omega_k)_{k=1}^K : 0 \leq \omega_k \leq \frac{1}{9K/10}, \sum_{k=1}^K \omega_k = 1 \right\}.$$

Ideally we would like to put 0 weights to all block means $\bar{X}_k$ corrupted by outliers. But, we cannot do it since $\mathcal{K}$ is unknown. To overcome this issue, we learn the optimal weights and consider the following minmax optimization problem

$$\max_{M \succeq 0, \mathrm{Tr}(M)=1} \min_{w \in \Delta_K} \langle M, \sum_{k=1}^K \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle. \tag{E_{x_c}}$$

This is the dual problem from Cheng et al. (2019a) adapted to the block means. The key insight from Cheng et al. (2019a) is that an approximate solution M_c of the maximization problem in (E_{x_c}) can be obtained in a reasonable amount of time using a covering SDP approach Cheng et al. (2019a); Peng et al. (2012) (see Section 3.4). We expect a solution (in M) to (E_{x_c}) to be close to a solution of the minimization problem in (3.2) – which is $M^* = (\mu - x_c)(\mu - x_c)^\top / \|\mu - x_c\|_2^2$ – and the same for their top eigenvectors (up to the sign). We note that in order to find a good descent direction the authors of Cherapanamjeri et al. (2019) also use a (different) SDP relaxation. Theirs costs $\mathcal{O}(N^4 + Nd)$ to be computed.

At a high level description, the robust descent algorithm we perform outputs $\hat{\mu}_K$ after at most $\log d$ iterations of the form $x_c - \theta_c v_1$ where v_1 is a top eigenvector of an approximate solution M_c to the problem (E_{x_c}) and θ_c is a step size. It starts at the coordinate-wise median of the bucketed means $\bar{X}_1, \ldots, \bar{X}_K$. In Algorithm 4, we define precisely the step size and the stopping criteria we use to define the algorithm (it requires too much notation to be defined at this stage). This algorithm outputs the vector $\hat{\mu}_K$ whose running time and statistical performance are gathered in the following result.

Theorem 3.2. *Grant Assumption 3.1. Let $K \in \{1, \ldots, N\}$ be the number of equal-size blocks and assume that $K \geq 300|\mathcal{O}|$. Let $u \in \mathbb{N}^*$ be a parameter of the covering SDP used at each descent step. With probability at least $1 - \exp(-K/180000) - (1/10)^u$, the descent algorithm finishes in time $\tilde{\mathcal{O}}(Nd + Kud)$ and outputs $\hat{\mu}_K$ such that*

$$\|\hat{\mu}_K - \mu\|_2 \leq 808 \left(1200 \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200 \|\Sigma\|_{op} K}{N}} \right).$$

To make the presentation of the proof of Theorem 3.2 as simple as possible we did not optimize the constants (better constants have been obtained in Catoni (2012); Catoni and Giulini (2017)). Theorem 3.2 generalizes and improves Theorem 3.1 in several ways. We first improve the confidence from a constant "9/10" to an exponentially large confidence $1 - \exp(-c_0 K)$ (when $u \sim K$), which was a major technical challenge (note however that the confidence 9/10 in Cheng et al. (2019a) can be increased to any desired confidence at the expense of deteriorating the rate of convergence – see footnote of page 2 in Cheng et al. (2019a)). We obtain the result for any covariance structure Σ and $\hat{\mu}_K$ does not require the knowledge of Σ for its construction. We obtain a result which holds for any N (even in the case where $N \leq d$). The construction of $\hat{\mu}_K$ does not require the knowledge of the exact proportion of outliers ϵ in the dataset, but it requires an upper bound in the number of outlier, so that we can chose $K \gtrsim |\mathcal{O}|$. Moreover, using a Lepskii adaptation method Lepskiĭ (1991, 1990) it is also possible to automatically choose K

and therefore to adapt to the proportion of outliers if we have some extra knowledge on $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$ (see Section 3.6 for more details). Moreover, if we only care about constant $9/10$ confidence, our runtime does not depend on ϵ and is nearly-linear $\tilde{\mathcal{O}}(Nd)$. We also refer the reader to Corollary 3.2 for more comparison with Theorem 3.1.

Remark 3.1 (Nearly-linear time)**.** *We identify two important situations where the algorithm from Theorem 3.2 runs in nearly-linear time, that is, in time $\tilde{\mathcal{O}}(Nd)$. First, when the number of outliers is known to be less than $\sqrt{N}$, we can choose $K \leq \sqrt{N}$ and $u = K$. In that case, the algorithm runs in time $\tilde{\mathcal{O}}(Nd)$ and the subgaussian rate is achieved with probability at least $1 - 2\exp(-c_0 K)$ for some constant c_0 (see also Corollary 3.3 for an adaptive to K version of this result). Another widely investigated situation is when we only want to have a constant confidence like $9/10$ as it is the case in the CS community such as in Theorem 3.1. In that case, one may choose $u = 1$ and any values of $K \in [N]$ can be chosen (so we can have any number of outliers up to a $N/300$) to achieve the rate in Theorem 3.2 with constant probability and in nearly-linear time $\tilde{\mathcal{O}}(Nd)$ (see also Corollary 3.2 for an adaptive to K version of this result). Finally, it is possible to get a subgaussian estimator for the all range of $K \in [N]$ which is also robust to adversarial outliers up to a constant fraction of N when we take $u = K$. In that case, the running time is $\tilde{\mathcal{O}}(Nd + K^2 d)$ which is at worst $\tilde{\mathcal{O}}(N^2 d)$. So algorithm outputs $\hat{\mu}_K$ in time between $\tilde{\mathcal{O}}(Nd)$ and $\tilde{\mathcal{O}}(N^2 d)$ depending on the number of outliers and the probability deviation certifying the result we want.*

Theorem 3.2 improves the result from Hopkins (2018); Cherapanamjeri et al. (2019) since $\hat{\mu}_K$ runs faster than the polynomial times $\mathcal{O}(N^{24} + Nd)$ and $\mathcal{O}(N^4 + Nd)$ in Hopkins (2018) and Cherapanamjeri et al. (2019). The algorithm $\hat{\mu}_K$ also does not require the knowledge of $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$. Finally, Theorem 3.2 provides running time guarantees on the algorithm unlike in Lugosi and Mendelson (2019c); M. Lerasle and Lecué (2017); Catoni and Giulini (2017) and it improves upon the statistical performance from Minsker (2015). The main technical novelty lies in Proposition 3.1, necessary to improve analysis from Cheng et al. (2019a) toward exponentially large confidence $1 - \exp(-c_0 K)$. Proposition 3.1 may be of independent interest. Theorem 3.2 also improves the running time in Cheng et al. (2019a) $\tilde{\mathcal{O}}(Nd/\epsilon^6)$ and the constant probability deviation (see Theorem 3.1 for more details) – both probability estimates and computational time have been improved by using bucketed means in place of the data themselves (see Remark 3.4 below for more details). The computational time improvement from Theorem 3.2 upon the one in Cherapanamjeri et al. (2019) is due to the use of covering SDP Allen-Zhu et al. (2014); Peng et al. (2012); Cheng et al. (2019a) at each iteration of the robust gradient descent algorithm. Very recent works Lei et al. (2020); Hopkins et al. (2020); Depersin (2020b) obtain similar results to the one of Theorem 3.2. They were also able to replace SDPs by spectral methods for the computations of a robust descent direction at each step. Even though cover SDPs are from a theoretical point of view computationally efficient, see Allen-Zhu et al. (2014); Peng et al. (2012), they are notoriously difficult to implement in practice whereas the power methods used in Lei et al. (2020); Hopkins et al. (2020); Depersin (2020b) open the door to implementable algorithms. For more references on robust mean estimation, we refer the reader to the survey from Diakonikolas and Kane (2019).

3.3 Proof of the statistical performance in Theorem 3.2

In this section, we prove the statistical performance of $\hat{\mu}_K$ as stated in Theorem 3.2. We first identify an event $\mathcal{E}$ onto which we will derive the rate of convergence of the order of (3.1). This event is also used to compute the running time of $\hat{\mu}_K$ in the next section as announced in Theorem 3.2.

Proposition 3.1. *Denote by $\mathcal{E}$ the event onto which for all symmetric matrices $M \succeq 0$ such that $\mathrm{Tr}(M) = 1$, there are at least $9K/10$ of the blocks for which $\left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leq 8r$ where*

$$r = 1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200\,\|\Sigma\|_{op}\,K}{N}}. \tag{3.3}$$

If Assumptions 3.1 holds and $K \geq 300|\mathcal{O}|$ then $\mathbb{P}[\mathcal{E}] \geq 1 - \exp(-K/180000)$.

Proposition 3.1 contains all the stochastic arguments we will use in this chapter (constants have not been optimized). In other words, after identifying the event $\mathcal{E}$, all the remaining arguments do not involve any other stochastic tools. The proof of Proposition 3.1 is based on a rounding argument similar to the one used to prove Grothendieck's inequality (see Grothendieck (1953); Pisier (2012)) or in the Goemans and Williamson's analysis of a SDP relaxation of the Max-Cut problem (see Goemans and Williamson (1995)) or in Nesterov's theorem (see Nesterov (1997)). Before proving Proposition 3.1, let us first state a result that is of particular interest beyond our problem.

Corollary 3.1. *On the event $\mathcal{E}$, for all symmetric matrices $M \in \mathbb{R}^{d\times d}$ such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$ there are at least $9K/10$ blocks k for which $\left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leq 8r$ and for all such k's and all $x_c \in \mathbb{R}^d$,*

$$\left\|M^{1/2}(\mu - x_c)\right\|_2 - 8r \leq \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2 \leq \left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r. \tag{3.4}$$

Let us now turn to a proof of Proposition 3.1. We first remark that if we were to only consider matrices M of rank 1, Proposition 3.1 would boil down to showing that for all $v \in \mathcal{S}_2^{d-1}$ (the unit sphere in ℓ_2^d) on more than $9K/10$ blocks $|\langle v, \bar{X}_k - \mu\rangle| \leq 8r$. This is a "classical" result in the MOM literature which has been proved in Lugosi and Mendelson (2019c) and M. Lerasle and Lecué (2017). We recall now this result and the short proof from M. Lerasle and Lecué (2017) adapted to the adversarial contamination setup from Assumption 3.1. We will use it to prove Proposition 3.1.

Lemma 3.1. *Grant Assumption 3.1 and assume that $K \geq 300|\mathcal{O}|$. With probability at least $1 - \exp(-K/180000)$, for all $v \in \mathcal{S}_2^{d-1}$, there are at least $99K/100$ of the blocks k such that $|\langle v, \bar{X}_k - \mu\rangle| \leq r$.*

Proof. We use the notation introduced in Assumption 3.1 and we considered the following bucketed means $\overline{\tilde{X}}_k = |B_k|^{-1}\sum_{i\in B_k}\tilde{X}_i$ for $k \in [K]$. They are the K means constructed on the N independent vectors $\tilde{X}_i, i \in [N]$ before contamination (whereas $\bar{X}_k$ are the ones constructed after contamination).

In the following, we show that with probability at least $1-\exp(-K/180000)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\sum_{k\in[K]} I(|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r) \leq \frac{2K}{300}. \tag{3.5}$$

The result from Lemma 3.1 follows from (3.5) because the adversary is allowed to change at most $|\mathcal{O}|$ data points among the $\tilde{X}_i$'s. Hence, there are at most $|\mathcal{O}|$ bucketed means $\overline{\tilde{X}}_k$ containing an outliers and so $K - |\mathcal{O}| \geq 299K/300$ means $\overline{\tilde{X}}_k$ which are unchanged that is for which $\overline{\tilde{X}}_k = \bar{X}_k$. So, if (3.5) holds then they are at least $298K/300$ means $\overline{\tilde{X}}_k$ for which $|\langle \overline{\tilde{X}}_k - \mu, v\rangle| \leq r$ and so, at least $297K/300 = 99K/100$ means $\bar{X}_k$ for which $|\langle \bar{X}_k - \mu, v\rangle| \leq r$.

As in Koltchinskii et al. (2003), we define $\phi(t) = 0$ if $t \leq 1/2$, $\phi(t) = 2(t - 1/2)$ if $1/2 \leq t \leq 1$ and $\phi(t) = 1$ if $t \geq 1$. We have $I(t \geq 1) \leq \phi(t) \leq I(t \geq 1/2)$ for all $t \in \mathbb{R}$ and so

$$\begin{aligned}
&\sum_{k\in[K]} I(|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r) \\
&\leq \sum_{k\in[K]} I(|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r) - \mathbb{P}[|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r/2] + \mathbb{P}[|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r/2] \\
&\leq \sum_{k\in[K]} \phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right) - \mathbb{E}\phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right) + \mathbb{P}[|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r/2] \\
&\leq \sup_{v\in\mathcal{S}_2^{d-1}} \left(\sum_{k\in[K]} \phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right) - \mathbb{E}\phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right)\right) + \sum_{k\in[K]} \mathbb{P}[|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r/2].
\end{aligned}$$

For all $k \in [K]$, we have

$$\begin{aligned}
\mathbb{P}[|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r/2] &\leq \frac{\mathbb{E}\langle \overline{\tilde{X}}_k - \mu, v\rangle^2}{(r/2)^2} \leq \frac{4Kv^\top \Sigma v}{Nr^2} \\
&\leq \frac{4K \sup_{v\in\mathcal{S}_2^{d-1}} v^\top \Sigma v}{Nr^2} = \frac{4K \|\Sigma\|_{op}}{Nr^2} \leq \frac{1}{300}
\end{aligned}$$

because $r^2 \geq 1200K \|\Sigma\|_{op}/N$.

Next, we use several tools from empirical process theory and in particular, for a symmetrization argument, we consider a family of N independent Rademacher variables $(\epsilon_i)_{i=1}^N$ independent of the $(\tilde{X}_i)_{i=1}^N$. In *(bdi)* below, we use the bounded difference inequality (Theorem 6.2 in Boucheron et al. (2013)). In *(sa-cp)*, we use the symmetrization argument and the contraction principle (Chapter 4 in Ledoux and Talagrand (2011)) – we refer to the supplementary material of M. Lerasle and Lecué (2017) for more details. We have, with probability at least $1 - \exp(-K/180000)$,

$$\begin{aligned}
&\sup_{v\in\mathcal{S}_2^{d-1}} \left(\sum_{k\in[K]} \phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right) - \mathbb{E}\phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right)\right) \\
&\overset{(bdi)}{\leq} \mathbb{E} \sup_{v\in\mathcal{S}_2^{d-1}} \left(\sum_{k\in[K]} \phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right) - \mathbb{E}\phi\left(\frac{|\langle \overline{\tilde{X}}_k - \mu, v\rangle|}{r}\right)\right) + \sqrt{\frac{K^2}{360000}} \\
&\overset{(sa-cp)}{\leq} \frac{4K}{Nr}\mathbb{E} \sup_{v\in\mathcal{S}_2^{d-1}} \langle v, \sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu)\rangle + \frac{K}{600} \\
&= \frac{4K}{\sqrt{N}r}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu)\right\|_2 + \frac{K}{600} \leq \frac{K}{300}
\end{aligned}$$

because $r \geq 1200\mathbb{E}\left\|\sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu^*)\right\|_2/\sqrt{N}$ since

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu)\right\|_2 \leq \sqrt{\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]} \epsilon_i(\tilde{X}_i - \mu)\right\|_2^2} \leq \sqrt{\mathrm{Tr}(\Sigma)}.$$

As a consequence, when $K \geq 300|\mathcal{O}|$, with probability at least $1 - \exp(-K/180000)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\sum_{k\in[K]} I(|\langle \overline{\tilde{X}}_k - \mu, v\rangle| > r) \leq \frac{|\mathcal{K}|}{300} + \frac{K}{300} \leq \frac{2K}{300},$$

which is (3.5). ∎

Proof of Proposition 3.1: Let $M \in \mathbb{R}^{d\times d}$ be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Denote by $\mathcal{A}_M = \{k \in [K] : \left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \geq 8r\}$ and assume that $|\mathcal{A}_M| \geq 0.1K$. Let G be a Gaussian vector in $\mathbb{R}^d$ with mean 0 and covariance matrix M (and independent from $X_1, \ldots, X_N$). We consider the random variable $Z = \sum_{k\in[K]} I\left(|\langle \bar{X}_k - \mu, G\rangle| > 5r\right)$. We work conditionally to $X_1, \ldots, X_N$ in this paragraph.

For all $k \in [K]$, $\langle \bar{X}_k - \mu, G\rangle$ is a centered Gaussian variable with variance $\sigma_k^2 := \left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2^2$. In particular, for all $k \in \mathcal{A}_M$, if we denote by g a standard real-valued Gaussian variable, we have $\mathbb{P}_G\left[|\langle \bar{X}_k - \mu, G\rangle| > 5r\right] \geq \mathbb{P}_G\left[|\langle \bar{X}_k - \mu, G\rangle| > 5\sigma_k/8\right] = 2\mathbb{P}[g > 5/8] \geq 0.528$ (where $\mathbb{P}_G$ (resp. $\mathbb{E}_G$) denotes the probability (resp. expectation) w.r.t. G conditionally on $X_1, \ldots, X_N$). Hence, $\mathbb{E}_G Z \geq 0.528|\mathcal{A}_M| \geq 0.0528K$. Since $|Z| \leq K$ a.s., it follows from Paley-Zygmund inequality (see Proposition 3.3.1 in de la Peña and Giné (1999)) that

$$\mathbb{P}_G[Z > 0.01K] \geq \frac{(\mathbb{E}_G Z - 0.01K)^2}{\mathbb{E}_G Z^2} \geq (0.0428)^2 = 0.0018.$$

Moreover, it follows from the Borell-TIS inequality (see Theorem 7.1 in Ledoux (2001) or pages 56-57 in Ledoux and Talagrand (2011)) that with probability at least $1 - \exp(-8)$, $\|G\|_2 \leq \mathbb{E}\|G\|_2 + 4\sqrt{\|M\|_{op}}$. Moreover, $\mathbb{E}\|G\|_2 \leq \sqrt{\mathrm{Tr}(M)} \leq 1$ and $\|M\|_{op} \leq \mathrm{Tr}(M) \leq 1$, so $\|G\|_2 \leq 5$ with probability at least $1 - \exp(-8) \geq 0.9996$. Since $0.9996 + 0.0018 > 1$ there exists a vector $G_M \in \mathbb{R}^d$ such that $\|G_M\|_2 \leq 5$ and $\sum_{k\in[K]} I\left(|\langle \bar{X}_k - \mu, G_M\rangle| > 5r\right) > 0.01K$. We recall that this latter result holds when we assume that $|\mathcal{A}_M| \geq 0.1K$.

Next, we denote by Ω_0 the event onto which for all $v \in \mathcal{S}_2^{d-1}$, there are at least $99K/100$ blocks such that $|\langle \bar{X}_k - \mu, v\rangle| \leq r$. We know from Lemma 3.1 that $\mathbb{P}[\Omega_0] \geq 1 - \exp(-K/180000)$. Let us place ourselves on the event Ω_0 up to the end of the proof. Let $M \in \mathbb{R}^{d\times d}$ be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$ and assume that $|\mathcal{A}_M| \geq 0.1K$. It follows from the first paragraph of the proof that there exists $G_M \in \mathbb{R}^d$ such that $\|G_M\|_2 \leq 5$ and $\sum_{k\in[K]} I\left(|\langle \bar{X}_k - \mu, G_M\rangle| > 5r\right) > 0.01K$. Given that we work on the event Ω_0, we have for $v_M = G_M/\|G_M\|_2$, that for more than $99K/100$ blocks $|\langle \bar{X}_k - \mu, v_M\rangle| \leq r$ and so $|\langle \bar{X}_k - \mu, G_M\rangle| \leq \|G_M\|_2 r \leq 5r$ which contradicts the fact that $\sum_{k\in[K]} I\left(|\langle \bar{X}_k - \mu, G_M\rangle| > 5r\right) > 0.01K$. Therefore, we necessarily have $|\mathcal{A}_M| \leq 0.1K$, which concludes the proof. ∎

Proof of Corollary 3.1: Let us assume that the event $\mathcal{E}$ holds up to the end of the proof. Let $M \in \mathbb{R}^{d\times d}$ be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Let $\mathcal{K}_M = \{k \in [K] : \left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leq 8r\}$. On the event $\mathcal{E}$, we have $|\mathcal{K}_M| \geq 9K/10$. Let $x_c \in \mathbb{R}^d$. For all $k \in \mathcal{K}_M$, we have $\left\|M^{1/2}(\mu - \bar{X}_k)\right\|_2 \leq 8r$ and so

$$\begin{aligned}\left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2 &\in \left\|M^{1/2}(\mu - x_c)\right\|_2 + \left[-\left\|M^{1/2}(\mu - \bar{X}_k)\right\|_2, \left\|M^{1/2}(\mu - \bar{X}_k)\right\|_2\right] \\ &\subset \left\|M^{1/2}(x_c - \mu)\right\|_2 + [-8r, 8r]\,.\end{aligned}$$

∎

Let us now turn to the study of the optimization problem (E_{x_c}) on the event $\mathcal{E}$. Like in Cheng et al. (2019a), we denote by OPT_{x_c} the optimal value of (E_{x_c}) and by

$$h_{x_c} : M \to \min_{w\in\Delta_K} \langle M, \sum_{k\in[K]} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle$$

its objective function to be minimized over $\{M \in \mathbb{R}^{d\times d} : M \succeq 0, \mathrm{Tr}(M) = 1\}$.

Remark 3.2. *For a given M, the optimal choice of $w \in \Delta_K$ in the definition of $h_{x_c}(M)$ is straightforward: one just have to put the maximum possible weight on the $9K/10$ smallest $\langle M, (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle, k \in [K]$. Formally, we set $\mathcal{S}_M = \sigma(\{1, 2, \cdots, 9K/10\})$, where σ is a permutation on $[K]$ that arranges the $(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c), k \in [K]$ in ascending order:*

$$\left\|M^{1/2}(\bar{X}_{\sigma(1)} - x_c)\right\|_2 \leq \left\|M^{1/2}(\bar{X}_{\sigma(2)} - x_c)\right\|_2 \leq \cdots \leq \left\|M^{1/2}(\bar{X}_{\sigma(K)} - x_c)\right\|_2 .$$

Then we get $h_{x_c}(M) = (1/|\mathcal{S}_M|) \sum_{k \in \mathcal{S}_M} (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)$.

The first lemma deals with the optimal value of (E_{x_c}) when the current point x_c is far from the mean μ.

Lemma 3.2. *On the event $\mathcal{E}$, for all $x_c \in \mathbb{R}^d$, if $\|x_c - \mu\|_2 > 16r$ then*

$$(8/9)(\|x_c - \mu\|_2 - 8r)^2 \leq OPT_{x_c} \leq (\|x_c - \mu\|_2 + 8r)^2.$$

Proof. Let M be a matrix such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Set $\mathcal{K}_M = \{k \in [K] : \left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leq 8r\}$. On the event $\mathcal{E}$, we have $|\mathcal{K}_M| \geq 9K/10$ and it follows from Corollary 3.1 that for all $k \in \mathcal{K}_M$ and all $x_c \in \mathbb{R}^d$,

$$\left\|M^{1/2}(\mu - x_c)\right\|_2 - 8r \leq \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2 \leq \left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r. \tag{3.6}$$

Then we define a weight vector $\tilde{\omega} \in \Delta_K$ by setting for all $k \in [K]$

$$\tilde{\omega}_k = \begin{cases} 1/|\mathcal{K}_M| & \text{if } k \in \mathcal{K}_M \\ 0 & \text{else.} \end{cases}$$

It follows from the definition of h_{x_c} and (3.6) that

$$\begin{aligned} h_{x_c}(M) &\leq \sum_{k \in [K]} \tilde{\omega}_k (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c) \qquad (3.7) \\ &= \frac{1}{|\mathcal{K}_M|} \sum_{k \in \mathcal{K}_M} \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2 \leq \left(\left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r\right)^2 . \end{aligned}$$

Taking the maximum over all $M \in \mathbb{R}^d$ such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$ on both side of the latter inequality yields the right-hand side inequality of Lemma 3.2.

For the left-hand side inequality of Lemma 3.2, we let $x_c \in \mathbb{R}^d$ be such that $\|x_c - \mu\|_2 > 16r$ and let M be such that $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. We use the notation and observation from Remark 3.2: we note that $|\mathcal{K}_M \bigcap \mathcal{S}_M| \geq 8K/10$ so that it follows from Corollary 3.1 that

$$\begin{aligned} h_{x_c}(M) &= \frac{1}{9K/10} \sum_{k \in \mathcal{S}_M} \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2 \geq \frac{1}{9K/10} \sum_{k \in \mathcal{A}_M \bigcap \mathcal{S}_M} \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2 \\ &\geq \frac{8K/10}{9K/10} \left(\left\|M^{1/2}(\mu - x_c)\right\|_2 - 8r\right)^2 . \end{aligned}$$

Then, taking the maximum over all $M \succeq 0$ such that $\mathrm{Tr}(M) = 1$ on both sides, finishes the proof. ∎

The next lemma shows that the top eigenvector of an approximate solution to (E_{x_c}) is aligned with the best possible descent direction $(\mu - x_c)/\|\mu - x_c\|_2$. It is taken from the proof of Lemma 3.3 in Cheng et al. (2019a). We reproduce here a short proof for completeness.

Proposition 3.2. *On the event $\mathcal{E}$, if M is a matrix such that $M \succeq 0$, $\mathrm{Tr}(M) = 1$ and $h_{x_c}(M) \geq (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $1/\sqrt{2} \leq \beta \leq 1$, then any top eigenvector v_1 of M satisfies*

$$\left|\langle v_1, \frac{x_c - \mu}{\|x_c - \mu\|_2}\rangle\right| > \sqrt{2\beta^2 - 1}.$$

Proof. Let M be a matrix such that $M \succeq 0$, $\mathrm{Tr}(M) = 1$ and $h_{x_c}(M) \geq (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $1/\sqrt{2} \leq \beta \leq 1$. We use the same argument as in the proof of Lemma 3.2: on the event $\mathcal{E}$, $|\mathcal{K}_M| \geq 9K/10$ where $\mathcal{K}_M = \{k \in [K] : \left\|M^{1/2}(\bar{X}_k - \mu)\right\|_2 \leq 8r\}$ and so $\tilde{\omega} \in \Delta_K$ where for all $k \in [K], \tilde{\omega}_k = 1/|\mathcal{K}_M|$ if $k \in \mathcal{K}_M$ and $\tilde{\omega}_k = 0$ if $k \notin \mathcal{K}_M$. It follows from the definition of h_{x_c} that

$$h_{x_c}(M) \leq \sum_{k\in[K]} \tilde{\omega}_k (\bar{X}_k - x_c)^\top M (\bar{X}_k - x_c) = \frac{1}{|\mathcal{K}_M|} \sum_{k\in\mathcal{K}_M} \left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2$$

and so from Corollary 3.1, $h_{x_c}(M) \leq \left(\left\|M^{1/2}(\mu - x_c)\right\|_2 + 8r\right)^2$. Since, we assumed that $h_{x_c}(M) \geq (\beta \|x_c - \mu\|_2 + 8r)^2$, it follows that $\left\|M^{1/2}(\mu - x_c)\right\|_2^2 \geq \beta^2 \|\mu - x_c\|_2^2$.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ denote the eigenvalues of M and let $v_1, \ldots, v_d$ denote corresponding eigenvectors. The conditions on M imply that $\sum_j \lambda_j = 1$ and $\mathcal{B}_M = (v_1, \ldots, v_d)$ is an orthonormal basis of $\mathbb{R}^d$. We denote $v = (\mu - x_c)/\|\mu - x_c\|_2$. We decompose v in $\mathcal{B}_M$ as $v = \sum_j \alpha_j v_j$ with $\sum_j \alpha_j^2 = 1$. Using this decomposition, we have $v^\top M v = \sum_j \lambda_j \alpha_j^2$. We have $\lambda_1 = \lambda_1 \sum_j \alpha_j^2 \geq \sum_j \lambda_j \alpha_j^2 \geq \beta^2$, so $\lambda_1 \geq \beta^2$. Moreover, since $\sum_j \lambda_j = 1$, we have $\beta^2 \sum_j \alpha_j^2 \leq \sum_j \lambda_j \alpha_j^2 \leq \lambda_1 \alpha_1^2 + (1 - \lambda_1)(1 - \alpha_1^2) \leq \alpha_1^2 + (1 - \beta^2) \sum_j \alpha_j^2$, so we have $\alpha_1^2 \geq (2\beta^2 - 1)$. As we know that $\alpha_1 = \langle v_1, v\rangle$, we get the result. ∎

Proposition 3.2 is the first tool we need to construct a descent algorithm since it provides a descent/ascent direction (depending on the sign of the top eigenvector of an approximate solution to (E_{x_c})). It remains to specify three other quantities to fully characterize our algorithm: a starting point, a step size and a stopping criteria. We start with the starting point. Here we simply use the coordinate-wise median-of-means. The following statistical guarantee on the coordinate-wise median-of-means is known or folklore but we want to put forward that in our case it holds on the event $\mathcal{E}$. This again shows that $\mathcal{E}$ is the only event we need to fully analyze all the building blocks of the algorithm. We recall that the coordinate-wise median-of-means is the estimator $\hat{\mu}^{(0)} \in \mathbb{R}^d$ whose coordinates are for all $j \in [d], \hat{\mu}_j^{(0)} = \mathrm{med}(\bar{X}_{k,j} : k \in [K])$ where $\bar{X}_{k,j}$ is the j-th coordinate of the block mean $\bar{X}_k$ for all $k \in [K]$.

Proposition 3.3. *On the event $\mathcal{E}$, we have $\left\|\hat{\mu}^{(0)} - \mu\right\|_2 \leq 8\sqrt{d}r$.*

Proof. Let us place ourselves on the event $\mathcal{E}$ during all the proof. For all directions, $v \in \mathcal{S}_2^{d-1}$, there are at least $9K/10$ blocks k such that $|\langle \bar{X}_k - \mu, v\rangle| \leq 8r$. In particular, for all $j \in [d], |\langle \bar{X}_k - \mu, e_j\rangle| \leq 8r$ where $(e_1, \ldots, e_d)$ is the canonical basis of $\mathbb{R}^d$. That is for at least $9K/10$ blocks $|\bar{X}_{k,j} - \mu_j| \leq 8r$. In particular, the latter result is true for the median of $\{\bar{X}_{k,j} : k \in [K]\}$, that is, for $\hat{\mu}_j^{(0)}$. We therefore have $\left\|\hat{\mu}^{(0)} - \mu\right\|_\infty \leq 8r$ and so $\left\|\hat{\mu}^{(0)} - \mu\right\|_2 \leq 8r\sqrt{d}$. ∎

Proposition 3.3 guarantees that starting from the coordinate-wise median-of-means we are off by a $\sqrt{d}$ proportional factor from the optimal rate r. This will play a key role to analyze the number of steps we need to reach μ within the optimal rate r. Indeed, if we prove a geometric decay of the distance to μ along the descent algorithm then only $\log d$ steps (up to a multiplicative constants) would be enough to reach μ by a distance at most of the order of r.

Let us now specify the step size we use at each iteration. At the current point x_c we compute a top eigenvector v_1 of an approximate solution M to (E_{x_c}) (i.e. M such that $h_{x_c}(M) \geq (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $1/\sqrt{2} \leq \beta \leq 1$). The next iteration is $x_{c+1} = x_c - \theta_c v_1$ where the step size is

$$\theta_c = -\text{Med}\left(\langle \bar{X}_k - x_c, v_1 \rangle : k \in [K]\right). \tag{3.8}$$

In particular, since $\theta_c v_1$ does not depend on the sign of v_1 (the product $\theta_c v_1$ is the same if we replace v_1 by $-v_1$), we do not care which top eigenvector of M we choose.

Let us now prove a geometric decay of the algorithm while x_c is far from μ. Again, this result is proved on the event $\mathcal{E}$.

Proposition 3.4. *On the event $\mathcal{E}$, the following holds. Let $x_c \in \mathbb{R}^d$ (be the current point of the algorithm). Assume that M is an approximate solution of (E_{x_c}): M is such that $h_{x_c}(M) \geq (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $0.78 \leq \beta \leq 1$ and let v_1 be one of its top eigenvectors. Then, we have*

$$\|x_{c+1} - \mu\|_2^2 \leq 0.8 \|x_c - \mu\|_2^2 + 64r^2$$

when $x_{c+1} = x_c - \theta_c v_1$ for θ_c defined in (3.8).

Proof. Let us assume that the event $\mathcal{E}$ holds up to the end of the proof. Let M be an approximate solution to (E_{x_c}) such that $h_{x_c}(M) \geq (\beta \|x_c - \mu\|_2 + 8r)^2$ for some $0.78 \leq \beta \leq 1$ and let v_1 be a top eigenvector of M.

In direction v_1, there are at least $9K/10$ blocks such that $|\langle \bar{X}_k - \mu, v_1 \rangle| \leq 8r$ (see Lemma 3.1). Hence, on these blocks, we also have

$$\begin{aligned} |\theta_c - \langle x_c - \mu, v_1 \rangle| &= |\text{Med}\left(\langle \mu - \bar{X}_k, v_1 \rangle : k \in [K]\right)| \\ &\leq \text{Med}\left(|\langle \mu - \bar{X}_k, v_1 \rangle| : k \in [K]\right) \leq 8r. \end{aligned} \tag{3.9}$$

Let $v = (\mu - x_c)/\|\mu - x_c\|_2$ denote the optimal normalized descent direction. We write $v = \lambda_1 v_1 + \lambda_2 v_1^\perp$ where $v_1^\perp$ is a normalized orthogonal vector to v_1. We have $\lambda_1^2 + \lambda_2^2 = 1$ and it follows from Proposition 3.2 that $|\lambda_1| = |\langle v_1, v \rangle| > \sqrt{2\beta^2 - 1}$. We conclude that

$$\begin{aligned} \|x_{c+1} - \mu\|_2^2 = \|x_c - \mu - \theta_c v_1\|_2^2 &= \left\|(\langle x_c - \mu, v_1 \rangle - \theta_c) v_1 + \langle x_c - \mu, v_1^\perp \rangle v_1^\perp\right\|_2^2 \\ &= (\langle x_c - \mu, v_1 \rangle - \theta_c)^2 + \langle x_c - \mu, v_1^\perp \rangle^2 \leq (8r)^2 + \lambda_2^2 \|x_c - \mu\|_2^2 \end{aligned}$$

As $\lambda_2^2 = 1 - \lambda_1^2 < 2 - 2\beta^2 < 0.8$ we get the result. ∎

We now have almost all the building blocks to fully characterize the algorithm. The last and final step is to find a stopping rule. The idea we use to design such a rule is based on Proposition 3.4: we know that when the current point x_c is not in a ℓ_2^d-neighborhood of μ with a radius $80r$ then the ℓ_2^d-distance between the next iteration x_{c+1} and μ should be less than $\sqrt{0.81}$ times the ℓ_2^d-distance between x_c and μ – that is a geometric decay of the distance to μ. Moreover, if the current iteration x_c is in a ℓ_2^d-ball centered in μ with the radius $80r$ then, it follows from Proposition 3.4 that the next iteration x_{c+1} will also be in a ℓ_2^d-ball centered in μ with radius at most $80r$. So once the algorithm enters the ball $B_2^d(\mu, 80r)$ it never leaves it. We therefore have a geometric decay of the distance to μ along the iterations until we reach the ball $B_2^d(\mu, 80r)$. Starting from the coordinate-wise median(-of-means) which is in a $8\sqrt{d}r$ neighborhood of μ (see Proposition 3.3), we only have to do $\log(8\sqrt{d})/\log(1/\sqrt{0.81})$ iterations to output a current point which at most $80r$-close to μ w.r.t. the ℓ_2^d-norm.

We are now in a position to write an "almost final" pseudo-code of our algorithm. In the next section, we will dive a bit deeper in this pseudo-code (and in particular on the covering SDP algorithm used to construct an approximate solution to (E_{x_c})) in order to provide a final pseudo-code together with its total running time.

input : $X_1, \dots, X_N$ and a number K of blocks
output : A robust subgaussian estimator of μ

1 Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$
2 Construct the K empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
3 Compute $\hat{\mu}^{(0)}$ the coordinate-wise median-of-means and put $x_c \leftarrow \hat{\mu}^{(0)}$
4 **for** $T = 1, 2, \cdots, \log(8\sqrt{d})/\log(1/\sqrt{0.81})$ **do**
5 Compute M_c an approximate solution to (E_{x_c}) such that

$$h_{x_c}(M_c) \geq (0.78 \|x_c - \mu\|_2 + 8r)^2$$

6 Compute v_1 a top eigenvector of M_c
7 Compute a step size $\theta_c = -\operatorname{Med}\left(\langle \bar{X}_k - x_c, v_1 \rangle : k \in [K]\right)$
8 Update $x_c \leftarrow x_c - \theta_c v_1$
9 **end**
10 **Return** x_c

Algorithm 1: "Almost final" pseudo-code of the robust sub-gaussian estimator of μ

Algorithm 1 is "almost" our final algorithm. There is one last step we need to check carefully: given a current point x_c we need to find a way to construct M_c satisfying "$h_{x_c}(M_c) \geq (0.78 \|x_c - \mu\|_2 + 8r)^2$" without knowing r or μ. This is the last issue we need to address in order to explain how step **5** from Algorithm 1 can be realized in a fully data-dependent way in a good time. This issue is answered in the next section together with the computation of its running time.

3.4 Approximately solving the SDP (E_{x_c})

The aim of this section is to show that, on the event $\mathcal{E}$, it is possible to construct in a reasonable amount of time a matrix M_c such that "$h_{x_c}(M_c) \geq (0.78 \|x_c - \mu\|_2 + 8r)^2$" without any extra information than the data. To that end we construct in an efficient way an approximate solution to the optimization problem (E_{x_c}) using covering SDP as in Cheng et al. (2019a). The main result of this section is the following.

Theorem 3.3. *Let $u \in \mathbb{N}^*$. On the event $\mathcal{E}$, for every $x_c \in \mathbb{R}^d$ such that $\|x_c - \mu\|_2 \geq 800r$, given input x_c, we can either compute, in time $\tilde{\mathcal{O}}(Kud)$, with probability $> 1 - (1/10)^{u+5}/\sqrt{d}$:*

- *A matrix M_c such that*

$$h_{x_c}(M_c) \geq (0.78 \|x_c - \mu\|_2 + 8r)^2$$

- *Or directly a subgaussian estimate of μ, using only the block means $\bar{X}_1, \dots, \bar{X}_K$ as inputs.*

Theorem 3.3 answers the last issue raised at the end of Section 3.3 and provides the running time for step **5** of Algorithm 1. It therefore concludes the statement that there exists a fully data-driven robust subgaussian algorithm for the estimation of a mean vector under the only Assumption 3.1 (the total running time of Algorithm 1 is studied in Section 3.5).

Remark 3.3. *Theorem 3.3 states that we either find an approximate solution M_c to (E_{x_c}) or a good estimate of μ (at the current point x_c). As we will see in this section, this second case is degenerate as it is not the typical situation.*

Before turning to the proof of Theorem 3.3, we recall the definition of the following quantities to ease the reading of the proof:

$$OPT_{x_c} = \min_{M \succeq 0 : \mathrm{Tr}(M)=1} h_{x_c}(M) \text{ where } h_{x_c} : M \to \min_{w \in \Delta_K} \langle M, \sum_{k \in [K]} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle$$

and (E_{x_c}) refers to the optimization problem $\min_M (h_{x_c}(M) : M \succeq 0, \mathrm{Tr}(M) = 1)$.

We now turn to the proof of Theorem 3.3. It is decomposed into several lemmas adapted from techniques developed by Cheng et al. (2019a) to approximately solve the SDP problem (E_{x_c}) in time $\tilde{\mathcal{O}}(Kud)$ as announced in Theorem 3.1. To that end, we first introduce the following covering SDP

$$\begin{aligned} \underset{M', y'}{\text{minimize}} \quad & \mathrm{Tr}(M') + \|y'\|_1 \\ \text{subject to} \quad & M' \succeq 0,\ y' \geq 0, \\ & \forall k \in [K],\ \rho(\bar{X}_k - x_c)^\top M'(\bar{X}_k - x_c) + 9K/10\ y'_k \geq 1 \end{aligned} \tag{C_ρ}$$

where $\rho > 0$ is some parameter that we will show how to fine-tune later. Then, we show that, for a good choice of ρ, we can turn a good approximate solution for (C_ρ) into a good approximate solution for (E_{x_c}).

We denote by $g(\rho)$ the optimal objective value of (C_ρ). We begin with a first lemma that shows how to link the two optimization problems (E_{x_c}) and (C_ρ). The proof can be found in Lemma 4.2 from Cheng et al. (2019a). We adapt it here for our purpose.

Lemma 3.3. *Let $\rho > 0$. From a feasible solution (M', y') for (C_ρ) that achieves $\mathrm{Tr}(M') + \|y'\|_1 \leq 1$, we can construct a feasible solution M for (E_{x_c}) with objective value $h_{x_c}(M) \geq 1/\rho$. The reverse is also true. In particular, if $g(\rho)$ (resp. OPT_{x_c}) denotes the optimal value achieved by the objective function in (C_ρ) (resp. (E_{x_c})), we have $g(\rho) \leq 1$ iff $1/\rho \geq OPT_{x_c}$.*

Proof. We first note that the optimization problem (E_{x_c}) is equivalent to the following one:

$$\begin{aligned} \underset{M, y, z}{\text{maximize}} \quad & z - \frac{\|y\|_1}{9K/10} \\ \text{subject to} \quad & M \succeq 0,\ \mathrm{Tr}(M) = 1,\ y \geq 0,\ z \geq 0 \\ & \forall k \in [K],\ (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c) + \ y_k \geq z \end{aligned} \tag{$\tilde{E}_{x_c}$}$$

Indeed, for a given $M \succeq 0$ such that $\mathrm{Tr}(M) = 1$, one can notice that the optimal value is achieved in $(\tilde{E}_{x_c})$ for $y_k = \max(0, z - (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c))$, $k \in [K]$ and $z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right)$ the 9/10-th quantile of $\{(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c) : k \in [K]\}$, so that $z - \|y\|_1 / (9K/10) = h_{x_c}(M)$ which gives the equivalence between (E_{x_c}) and $(\tilde{E}_{x_c})$.

Then, let a feasible solution (M', y') for (C_ρ) be such that $\mathrm{Tr}(M') + \|y'\|_1 \leq 1$. We define

$$M = \frac{M'}{\mathrm{Tr}(M')}, z = \frac{1}{\rho\,\mathrm{Tr}(M')} \text{ and } y = \frac{(9K/10)}{(\rho\,\mathrm{Tr}(M'))} y'.$$

We can check that (M, y, z) is feasible for $(\tilde{E}_{x_c})$ and $z - \|y\|_1 / (9K/10) \geq 1/\rho$. Hence, given the equivalence between (E_{x_c}) and $(\tilde{E}_{x_c})$, we obtain that M is feasible for (E_{x_c}) and that $h_{x_c}(M) \geq 1/\rho$.

Conversely, if M is feasible for (E_{x_c}) such that $h_{x_c}(M) \geq 1/\rho$ then we define y and z such that for all $k \in [K]$, $y_k = \max(0, z - (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)), k \in [K]$ and $z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right)$. We check that (M, y, z) is feasible for $(\tilde{E}_{x_c})$ with objective values equals to $h_{x_c}(M)$ and so it is larger than $1/\rho$. Next, by defining

$$M' = \frac{M}{\rho z} \text{ and } y' = \frac{y}{(9K/10)z},$$

we see that (M', y') is feasible for (C_ρ) and its objective values is less than 1. ∎

From Lemma 3.3, it is enough to solve (C_ρ) – for a good choice of ρ – to find a good approximate solution for (E_{x_c}). It therefore remains to find such a good ρ. To do so, we rely on the next two lemmas. The first one is adapted from Lemma 4.3 in Cheng et al. (2019a); we recall that $g(\rho)$ is the optimal value achieved by the objective function in (C_ρ).

Lemma 3.4. *For every $\rho > 0$ and $\alpha \in (0, 1)$, $g((1-\alpha)\rho) \geq g(\rho) \geq (1-\alpha)g((1-\alpha)\rho)$.*

Proof. A feasible pair (M', y') for $(C_{(1-\alpha)\rho})$ is also feasible for (C_ρ), which gives the first inequality. If (M', y') is a feasible pair for (C_ρ), then $(M'/(1-\alpha), y'/(1-\alpha))$ is a feasible pair for $(C_{(1-\alpha)\rho})$, which gives the second inequality. ∎

It follows from Lemma 3.4 that g is continuous, non increasing and $g(1/OPT_{x_c}) = 1$ (this follows from Lemma 3.3 since we have that $g(\rho) \leq 1$ iff $1/\rho \geq OPT_{x_c}$ and the continuity of g). So in order to find a good solution, we must find a ρ such that $g(\rho)$ is as close to 1 as possible. Unfortunately, we do not know how to solve (C_ρ) exactly for a given $\rho > 0$, but we can compute efficiently a good approximation (M', y') and a top eigenvector of M' thanks to the following result which can be found in Peng et al. (2012) or Allen-Zhu et al. (2015) and is detailed in Cheng et al. (2019a) (see Section 4 and Remark 3.4).

Lemma 3.5. *[Peng et al. (2012), Allen-Zhu et al. (2015)] Let $u \geq 1$ be an integer. For every $\rho > 0$ and every fixed $\eta > 0$, we can find with probability $> 1 - (1/10)^{u+10}/d$ a feasible solution to (C_ρ) that is η-close to the optimal, that is to say a feasible pair (M', y') so that $\mathrm{Tr}(M') + \|y'\|_1 \leq (1+\eta)g(\rho)$ in time $\tilde{\mathcal{O}}(uKd)$. Moreover, it is possible to find an approximate top eigenvector of M' in $\tilde{\mathcal{O}}(Kd)$.*

We compute $(u + 3\log(d) + 10)$ times independently the (randomized) algorithm from Peng et al. (2012) (or the one from Allen-Zhu et al. (2015)) that has a runtime of $\tilde{\mathcal{O}}(Kd)$ and that outputs an η-close feasible solution with probability 9/10. By taking the largest of the output's objective value, we have an η-close feasible solution with probability $1 - (1/10)^{u+3\log(d)+10}$, in time $\tilde{\mathcal{O}}(uKd)$, proving Lemma 3.5.

Let us call $\texttt{ALG}_\rho$ the algorithm from Lemma 3.5, that takes as input $((\bar{X}_k)_{k=1}^K, x_c, \rho, \eta, u)$ and returns a feasible pair (M', y') for (C_ρ) satisfying $\mathrm{Tr}(M') + \|y'\|_1 \leq (1+\eta)g(\rho)$ in $\tilde{\mathcal{O}}(uKd)$, with probability $> 1 - (1/10)^{u+10}/d$. Next, in order to find a good ρ, we have to get some additional information on the function g. We will get it on the event $\mathcal{E}$.

Lemma 3.6. *On the event $\mathcal{E}$, for all $x_c \in \mathbb{R}^d$, if $\|x_c - \mu\|_2 > 8r$ then*

$$g(\rho) \leq \frac{1}{\rho\ OPT_{x_c}}\left(1 + \rho OPT_{x_c}\left(\frac{9(\|x_c - \mu\|_2 + 8r)^2}{8(\|x_c - \mu\|_2 - 8r)^2} - 1\right)\right).$$

Proof. We use the same notation as in the proof of Lemma 3.3. For any $\nu > 0$, we can choose a triplet (M, y, z) feasible for $(\tilde{E}_{x_c})$ such that $z - \|y\|_1/(9K/10) > OPT_{x_c} - \nu$ and z and y are the optimal solutions of the problem $(\tilde{E}_{x_c})$ given by $y_k = \max(0, z - (\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)), k \in [K]$

and $z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right)$ the 9/10-th quantile of $\{(\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c) : k \in [K]\}$.

On the event $\mathcal{E}$, Lemma 3.2 yields $OPT_{x_c} > (8/9)(\|x_c - \mu\|_2 - 8r)^2$ and we have from Corollary 3.1 that

$$z = \mathcal{Q}_{9/10}\left((\bar{X}_k - x_c)^\top M(\bar{X}_k - x_c)\right) = \mathcal{Q}_{9/10}\left(\left\|M^{1/2}(\bar{X}_k - x_c)\right\|_2^2\right)$$
$$\leq \left(\left\|M^{1/2}(x_c - \mu)\right\|_2 + 8r\right)^2 \leq (\|x_c - \mu\|_2 + 8r)^2$$

because $M \succeq 0$ and $\mathrm{Tr}(M) = 1$. Let $M' = M/(\rho z), y' = y/[z(9K/10)]$. Since (M', y') is feasible for (C_ρ), we have

$$g(\rho) \leq \mathrm{Tr}(M') + \|y'\|_1 \leq \frac{1 + \rho\,\|y\|_1/(9K/10)}{\rho z}$$
$$< \frac{1 + \rho(z - OPT_{x_c} + \nu)}{\rho z} \leq \frac{1 + \rho\nu + \rho OPT_{x_c}\left(\frac{9(\|x_c-\mu\|_2+8r)^2}{8(\|x_c-\mu\|_2-8r)^2} - 1\right)}{\rho(OPT_{x_c} - \nu)}.$$

By taking $\nu \to 0$, we get the result. ∎

Proof of Theorem 3.3. Let us place ourselves on the event $\mathcal{E}$ so that we can apply Lemma 3.6. Let $x_c \in \mathbb{R}^d$ and assume that $\|x_c - \mu\|_2 > 800r$. It follows from Lemma 3.6 that $g(\rho) \leq 1/(\rho\ OPT_{x_c}) + 0.171$. Therefore, if we can find a ρ such that $g(\rho) \geq 1 - \epsilon + 0.171$ for some $0 < \epsilon < 1$, then necessarily $1/\rho \geq OPT_{x_c}(1 - \epsilon)$. Let us take $\epsilon = 0.173$, and $\eta = 0.0001$. Then if $\texttt{ALG}_\rho$ returns a feasible pair (M', y') for (C_ρ) so that $0.9981 \leq \mathrm{Tr}(M') + \|y'\|_1 \leq 1$, then, since $0.9981 > 1.0001 \times 0.998 = (1 + \eta)(1 - \epsilon + 0.171)$ we will know that, with probability $> 1 - (1/10)^{u+10}/d$,

$$(1 + \eta)g(\rho) \geq \mathrm{Tr}(M') + \|y'\|_1 \geq (1 + \eta)(1 - \epsilon + 0.171)$$

hence $1/\rho \geq OPT_{x_c}(1 - \epsilon)$, and by Lemma 3.3, we can construct a feasible solution M_c for (E_{x_c}) with objective value satisfying $h_{x_c}(M_c) \geq OPT_{x_c}(1 - \epsilon)$. Next, using Lemma 3.2, we obtain that when $\|x_c - \mu\|_2 \geq 800r$,

$$h_{x_c}(M_c) \geq OPT_{x_c}(1 - \epsilon) \geq (1 - \epsilon)(8/9)\,(\|x_c - \mu\|_2 - 8r)^2 \geq (0.78\,\|x_c - \mu\|_2 + 8r)^2$$

for $\epsilon = 0.173$, solving step **5** from Algorithm 1.

Therefore, it only remains to show how to find a ρ such that $\texttt{ALG}_\rho$ returns a pair (M', y') (feasible for (C_ρ)) satisfying $0.9981 \leq \mathrm{Tr}(M') + \|y'\|_1 \leq 1$. We do it first by assuming that we have access to an initial ρ_0 such that $\texttt{ALG}_{\rho_0}$ returns a feasible pair (M', y') for (C_ρ) (for $\rho = \rho_0$) so that $\mathrm{Tr}(M') + \|y'\|_1 \leq 1$ and to a maximal number T of iterations (we will also see later how to choose such ρ_0 and T). The following algorithm (which is a binary search) taking as input $(\bar{X}_1, \ldots, \bar{X}_K, x_c, \rho_0, u, T)$ returns a feasible pair (M', y') for (C_ρ) so that $0.9981 \leq \mathrm{Tr}(M') + \|y'\|_1 \leq 1$ (when T is large enough). This is simply due to the fact that g is continuous, non increasing, $g(0) = 10/9 > 1$ and $g(\rho) \leq 2/8$ when $\rho \to +\infty$ and $\|x_c - \mu\|_2 > 800r$ (because of Lemma 3.6). For this to work, we need that for each iteration, $\texttt{ALG}_\rho$ returns a feasible pair (M', y') for (C_ρ) (for $\rho = \rho_0$) so that $\mathrm{Tr}(M') + \|y'\|_1 \leq (1 + 0.0001)g(\rho)$. We will suppose that it is the case for the rest of the proof. By union bound, this happens with probability at least $> 1 - T(1/10)^{u+10}/d$

input : $\bar{X}_1, \dots, \bar{X}_K$, x_c, ρ_0, u, T
output : A feasible pair (M', y') for (C_ρ) satisfying $0.9981 \leq \mathrm{Tr}(M') + \|y'\|_1 \leq 1$

$\rho_m \leftarrow 0$, $\rho_M \leftarrow \rho_0$, $V \leftarrow \mathtt{ALG}_{\rho_0}(x_c, u, \eta = 0.0001)$, $i \leftarrow 0$
while $V \notin [0.9981, 1]$ *and* $i < T$ **do**
  if $V < 0.9981$ **then**
    $\rho_M \leftarrow (\rho_M + \rho_m)/2$
  end
  else
    $\rho_m \leftarrow (\rho_M + \rho_m)/2$
  end
  $V \leftarrow objective(\mathtt{ALG}_{\frac{\rho_m+\rho_M}{2}}(x_c, u, \eta = 0.0001))$, $i \leftarrow i + 1$
end
Return $\mathtt{ALG}_{\frac{\rho_m+\rho_M}{2}}(x_c, u, \eta = 0.0001)$

Algorithm 2: The `BinarySearch` algorithm to find a ρ so that $\mathtt{ALG}_\rho$ returns a pair (M', y') feasible for (C_ρ) satisfying $0.9981 \leq \mathrm{Tr}(M') + \|y'\|_1 \leq 1$.

If we can find a ρ_0 (such that $\mathtt{ALG}_{\rho_0}$ returns a feasible pair (M', y') for (C_ρ) so that $\mathrm{Tr}(M') + \|y'\|_1 \leq 1$) and a large enough number of iterations T in `BinarySerach`, Algorithm 2 returns a feasible pair (M', y') for (C_ρ) from which we can construct an approximating solution M_c for (E_{x_c}) with objective value $h_{x_c}(M_c)$ larger than $(0.78\|x_c - \mu\|_2 + 8r)^2$ whenever $\|x_c - \mu\|_2 \geq 800r$. This is exactly what we expect in step **5** of Algorithm 1. Next, the last and final step that remains to be explained is to show how one can get such a ρ_0 and T using only the block means $(\bar{X}_k)_{k=1}^K$ in $\tilde{\mathcal{O}}(Nd + uKd)$.

Let us consider $\hat{\mu}^{(0)}$ the coordinate-wise median(-of-means) and let us define $\delta = \mathrm{Med}(\left\|\bar{X}_k - \hat{\mu}^{(0)}\right\|_2 : k \in [K])$ – both quantities can be computed in time $\tilde{\mathcal{O}}(Kd)$. On the event $\mathcal{E}$, it follows from Corollary 3.1 (for $M = I_d/d$) and Proposition 3.3 that $\delta \leq 16\sqrt{d} \times r$. So if one takes $\rho_0 = d/\delta^2 \geq 1/[(16)^2 r^2]$, and if $\|x_c - \mu\|_2 > 800r$, Lemma 3.2 and Lemma 3.6 guarantee that $OPT_{x_c} \geq (8/9)\left(\|x_c - \mu\|_2 - 8r\right)^2 \geq (8/9)(792)^2 r^2$ and so

$$g(\rho_0) \leq \frac{1}{\rho\, OPT_{x_c}} + 0.171 \leq \frac{16^2}{(8/9)(792)^2} + 0.171 < 0.18$$

so $\mathtt{ALG}_{\rho_0} \leq (1+\eta)g(\rho) < 1.0001 \times 0.18 < 1$ (for the same choice of $\eta = 0.0001$).

Now we tackle the question of the number T of iterations, which is crucial for the runtime. We know from Lemma 3.4 and Lemma 3.6 that the interval I of all ρ's such that $0.9981 \leq objective(\mathtt{ALG}_\rho) \leq 1$ is at least of size $0.001/OPT_{x_c}$ when $\|x_c - \mu\|_2 > 800r$. Indeed, since $g(\rho) \leq objective(\mathtt{ALG}_\rho) \leq (1+\eta)g(\rho)$, if ρ is such that $0.9981 \leq g(\rho) \leq 1/(1+\eta)$ then $0.9981 \leq objective(\mathtt{ALG}_\rho) \leq 1$. Now, if we let $\rho_1 > 0$ and $0 < \alpha < 1$ be such that $g(\rho_1) = 0.9981$ and $g((1-\alpha)\rho_1) = 1/(1+\eta)$ the interval I is at least of size $\alpha\rho_1$. Moreover, from Lemma 3.4 we have $1/(1+\eta) \leq g((1-\alpha)\rho_1) \leq g(\rho_1)/(1-\alpha)$ and so $0.9981 = g(\rho_1) \geq (1-\alpha)/(1+\eta)$, i.e. $\alpha \geq 1 - 0.9981(1+\eta) > 0.001$. Finally, since $g(\rho_1) \leq 1$, $g(1/OPT_{x_c}) = 1$ and g is non-increasing, we conclude that $\rho_1 \geq 1/OPT_{x_c}$ and so the length of I is at least $\alpha\rho_1 \geq 0.001/OPT_{x_c}$.

So, in the case where $\|x_c - \mu\|_2 > 800r$, $\log_2(\rho_0 \times OPT_{x_c}/0.001)$ iterations are enough to ensure that `BinarySearch` outputs (M', y') (from $\mathtt{ALG}_\rho$ for a well-chosen ρ) feasible for (C_ρ) and such that $0.9981 \leq \mathrm{Tr}(M') + \|y'\|_1 \leq 1$. Moreover, on the event $\mathcal{E}$ it is possible to show that for all iterations x_c of the algorithm we have $\|x_c - \mu\|_2 < C\sqrt{d}r$ for a constant $C \leq 800$ (we may take that as an induction hypothesis for the firsts iterates x_c, and the proof of Theorem 3.2

below in Section 3.5 shows that it will still holds for x_{c+1}). So if $\delta > r/d$ then $\rho_0 < d^3/r^2$, and since $OPT_{x_c} < (C^2d+8)r^2$ (this follows from Lemma 3.2), the binary search ends in time $T = \log_2(\tilde{C}d^4)$ with $\tilde{C} < 10^6$.

Thus, if the binary search has not ended in that time, we have either $\delta < r/d$ (which is a degenerate case) or $\|x_c - \mu\|_2 < 800r$ (or both). If $\|x_c - \mu\|_2 > 800r$ and $\delta < r/d$, then, taking $\rho_1 = 1/(d\delta)^2$, we have, by Lemma 3.6, $\texttt{ALG}_{\rho_1} < 1/2$. So, if we can not end our binary search in time $\log_2(\tilde{C}d^4)$, we compute $\texttt{ALG}_{1/(d\delta)^2}$: if this gives something smaller than 1, that means that $1/(d\delta)^2 > 1/OPT_{x_c} \Rightarrow \delta < \sqrt{(C^2d+8)}r/d < (C+1)r/\sqrt{d}$. We notice that on $\mathcal{E}$, $\left\|\hat{\mu}^{(0)} - \mu\right\|_2 < \delta + 8r$, so if $\texttt{ALG}_{1/(d\delta)^2} < 1$, then $\hat{\mu}^{(0)}$ is a good estimate for μ. If on the contrary we have $\texttt{ALG}_{\rho_1} > 1$, it means that $\|x_c - \mu\|_2 < 800r$, so we stop the algorithm and return x_c.

Let us write now in pseudo-code the procedure we just described. This is an algorithm, named `SolveSDP`, running in $\tilde{\mathcal{O}}(Kud)$ which takes as inputs $\bar{X}_1, \dots, \bar{X}_K$, x_c, u and which outputs, on the event $\mathcal{E}$, with probability $> 1 - \log(\tilde{C}d^4)(1/10)^{u+10}/d$, for every $x_c \in \mathbb{R}^d$ such that $\|x_c - \mu\|_2 \geq 800r$ either a matrix M_c such that

$$h_{x_c}(M_c) \geq (0.78\|x_c - \mu\|_2 + 8r)^2$$

or a subgaussian estimate of μ. It therefore describes step **5** from Algorithm 1.

input : $\bar{X}_1, \dots, \bar{X}_K$, x_c and u
output : A feasible solution for (E_{x_c})

Compute the coordinate wise MOM $\hat{\mu}^{(0)}$ and $\delta = \text{Med}(\left\|\bar{X}_k - \hat{\mu}^{(0)}\right\|_2 : k \in [K])$
$T \leftarrow \log(\tilde{C}d^4)$, $\rho_0 \leftarrow d/\delta^2$
$(M', y') \leftarrow \text{BinarySearch}(T, \rho_0, u, x_c)$
if $\text{Tr}(M') + \|y\|_1 \in [0.9981, 1]$ **then**
  $M \leftarrow M'/\text{Tr}(M')$
  Return (True, M)
end
else
  if $\texttt{ALG}_{1/(d\delta)^2}(x_c, u, \eta = 0.0001) < 1$ **then**
    Return (False, $\hat{\mu}^{(0)}$)
  end
  else
    Return (False, x_c)
  end
end

Algorithm 3: SolveSDP

Remark 3.4. *[Two advantages of block means] During the whole algorithm, we solve the program (C_ρ) up to a factor $(1+\eta)$ where η is* fixed *(here we take it equal to* 0.0001*). This differs crucially from the work of Cheng et al. (2019a) where η depends on the fraction of outliers, which decreases the performance of the algorithm in Lemma 3.5, the true running time being $\tilde{\mathcal{O}}(Kd/Poly(\eta))$. This is a first advantage of using bucketed means instead of the data themselves: we work with a constant fraction of corrupted blocks (we took it equal to* $1/10$*). The second advantages is of stochastic nature, it is revealed by Proposition 3.1 or Lemma 3.1: most of the bucketed means have a nice subgaussian behavior in all directions. Working with bucketed means has therefore two advantages: a stochastic one, which is to exhibit a subgaussian behavior for* $9K/10$ *blocks*

even under a L_2-moment assumption and a computational one, which is to make the proportion of corrupted blocks constant.

3.5 The final algorithm and its computational cost: proof of Theorem 3.2

We are now in a position to fully describe our robust subgaussian descent algorithm running in $\tilde{\mathcal{O}}(Nd + uKd)$. One may check that its construction is fully data-dependent, in particular, we do not need to know the value of r or the proportion of outliers.

input : $X_1, \dots, X_N$, $K \in [N]$ and $u \in \mathbb{N}^*$
output : A robust subgaussian estimator of μ

Construct an equipartition $B_1 \sqcup \dots \sqcup B_K = \{1, \dots, N\}$
Construct the K empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
Compute $\hat{\mu}^{(0)}$ the coordinate-wise median
$x_c \leftarrow \hat{\mu}^{(0)}$, Bool $\leftarrow$ True, $T \leftarrow 0$
while *Bool and* $T < \log(8\sqrt{d})/\log(1/0.81)$ **do**
    Bool, $A \leftarrow$SolveSDP$(\bar{X}_1, \dots, \bar{X}_K, x_c, u)$
    if *Bool* **then**
        $M_c \leftarrow A$
        Compute v_1 a top eigenvector of M_c
        Compute a step size $\theta_c = -\operatorname{Med}\left(\langle \bar{X}_k - x_c, v_1 \rangle : k \in [K]\right)$
        Update $x_c \leftarrow x_c - \theta_c v_1$
        $T \leftarrow T + 1$
    end
    else
        $x_c \leftarrow A$
    end
end
Return x_c

Algorithm 4: Final Algorithm: covSDPofMeans

Proof of Theorem 3.2. From Theorem 3.3, we know that on $\mathcal{E}$, when, $\|x_c - \mu\|_2 > 800r$, we get, with probability $> 1 - (1/10)^{u+5}/\sqrt{d}$, an M_c so that $h_{x_c}(M_c) \geq (0.8\|x_c - \mu\|_2 + 8r)^2$ (or directly a subgaussian estimate, in which case our work is done). Proposition 3.4, states that in that case $\|x_{c+1} - \mu\|_2^2 \leq 0.8\|x_c - \mu\|_2^2 + 64r^2 \leq 0.81\|x_c - \mu\|_2^2$. So we have a geometric decays and Proposition 3.3 guarantees that our starting point is at most $8\sqrt{d}r$ far away from the mean so that in at most $\log(8\sqrt{d})/\log(1/0.81))$ steps the algorithm outputs its current point which is r-close to μ, with probability $> 1 - (1/10)^{u+5}\log(8\sqrt{d})/(\log(1/0.81))\sqrt{d}) > 1 - (1/10)^u$ (by union bound).

The last thing to do is to control what happens when $\|x_c - \mu\|_2 < 800r$. Then, we have no guarantees on v_1, but using the similar argument as in the proof of Proposition 3.4 we know that

$$|\theta_c - \langle x_c - \mu, v_1 \rangle| = |\operatorname{Med}\left(\langle \mu - \bar{X}_k, v_1 \rangle : k \in [K]\right)| \leq \operatorname{Med}\left(|\langle \mu - \bar{X}_k, v_1 \rangle| : k \in [K]\right) \leq 8r \tag{3.10}$$

and (for some $v_1^\perp$ a normalized orthogonal vector to v_1)

$$\|x_{c+1}-\mu\|_2^2 = \|x_c-\mu-\theta_c v_1\|_2^2 = \left\|(\langle x_c-\mu, v_1\rangle-\theta_c)v_1+\langle x_c-\mu, v_1^\perp\rangle v_1^\perp\right\|_2^2$$
$$= (\langle x_c-\mu, v_1\rangle-\theta_c)^2+\langle x_c-\mu, v_1^\perp\rangle^2 \le (8r)^2+\|x_c-\mu\|_2^2 .$$

Hence, $\|x_{c+1}-\mu\|_2 \le (8r)+\|x_c-\mu\|_2$. Therefore, in the worst case scenario where $\|x_c-\mu\|_2 > 800r$ at the last iteration, the algorithm outputs the next iteration $\hat{\mu}_K = x_{c+1}$ so that $\|\hat{\mu}_K-\mu\|_2 \le 808r$.

We end this proof with the computation of the running time of Algorithm 4. We detail the computation cost for each line of Algorithm 4: line **1** cost N, line **2** costs Nd, line **3** costs $\mathcal{O}(dK\log(K))$. The while loop in line **5** is running at least $\log d$ times (up to constant) so that the computational cost of all remaining lines of Algorithm 4 are at worst to be multiplied by $\log d$. Line **6** costs $\log(\tilde{C}d^4)$ steps, each of cost $\tilde{\mathcal{O}}(Kud)$ (that comes from Lemma 3.5). Line **9** can be computed in $\tilde{\mathcal{O}}(Nd)$ thanks to Lemma 3.5. Finally, line **10** costs $\mathcal{O}(Kd)$. Other lines take time at most d. We thus recover the running time announced in Theorem 3.2. ∎

3.6 Adaptive choice of K and results in expectation

Given a number of blocks $K \in \{1,\dots,N\}$, a parameter $u \ge 1$ (so that the covering SDPs from Peng et al. (2012) (used in Lemma 3.5) run in $u+3\log d+10$ times) and the (adversarially corrupted and heavy-tailed) dataset $\{X_1,\dots,X_N\}$, Algorithm 4 returns a vector $\hat{\mu}_K$ in $\mathbb{R}^d$ and Theorem 3.2 ensures that $\hat{\mu}_K$ estimates the true mean μ at the subgaussian rate (**??**) with large probability as long as $K \ge 300|\mathcal{O}|$. As a consequence, we have certified statistical guarantees for $\hat{\mu}_K$ only when some a priori knowledge on the number $|\mathcal{O}|$ of outliers is provided (such as "the corruption of this database is less than 5%") or if we choose K like N- but, in this later case the rate (**??**) may be too pessimistic. The aim of this section is to overcome this issue by constructing a procedure which can automatically adapt to the number of outliers. The resulting procedure (denoted later by $\hat{\mu}^{(\hat{\jmath})}$) satisfies the same statistical bounds as $\hat{\mu}_K$ for all $K \ge 300|\mathcal{O}|$ without knowing $|\mathcal{O}|$ (up to constants). We also show that it satisfies results in expectation.

The adaptation method we use is based on the Lepski method Lepskiĭ (1990, 1991) which is another tool used by the "statistical community" working on robustness issues since Lugosi and Mendelson (2019c); Catoni (2012). The price we pay for this adaptation is the a priori knowledge of the rate (**??**) for all K which means that we know in advance $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$ – this is for instance the case when it is known that Σ is the identity matrix I_d. Of course, one can design robust estimators for $\mathrm{Tr}(\Sigma)$ and $\|\Sigma\|_{op}$ but this requires stronger assumptions (more than four moments) that we want to avoid at this stage.

Lepski's method proceeds as follows. We set for all $K \in \{1,\dots,N\}$ and all $j \in \{0,1,\dots,\log_2 N\}$

$$r_K^* = 808\left(1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}}+\sqrt{\frac{1200\,\|\Sigma\|_{op}K}{N}}\right) \text{ and } r^{(j)} = r^*_{\lceil N/2^j\rceil}$$

the rate of convergence from Theorem 3.2. For a given parameter $u_j \in \mathbb{N}^*$, we construct from Algorithm 4

$$\hat{\mu}^{(j)} \leftarrow covSDPofMeans(X_1,\dots,X_N, K=\lceil N/2^j\rceil, u=u_j). \tag{3.11}$$

Classical Lepski's method considers the largest J such that $\bigcap_{j=0}^J B_2^d(\hat{\mu}^{(j)}, r^{(j)})$ is none-empty and then take any point $\hat{\mu}$ in this none-empty intersection. Standard analysis of Lepski's method shows that $\hat{\mu}$ estimates μ at the rate r_K^* (up to an absolute constant) simultaneously for all

$K \in \{300|\mathcal{O}|, \dots, N\}$ without knowing $|\mathcal{O}|$. Given that checking that the intersection of several ℓ_2^d-balls may not be straightforward, we use a slightly modified version of Lepski's method as described in the following algorithm.

input : $X_1, \dots, X_N$ and $\{u_j : j = 0, 1, 2, \dots, \log_2 N\} \subset \mathbb{N}^*$
output: A robust subgaussian estimator of μ with adaptive choice of K
init : $J = 0$ and $\hat{\mu}^{(0)} = covSDPofMeans(X_1, \dots, X_N, K = N, u = u_0)$

while $\left\|\hat{\mu}^{(J)} - \hat{\mu}^{(j)}\right\|_2 \leq r^{(J)} + r^{(j)}, j = J-1, J-2, \dots, 0$ **do**
  $J \leftarrow J + 1$
  $\hat{\mu}^{(J)} \leftarrow covSDPofMeans(X_1, \dots, X_N, K = \lceil N/2^J \rceil, u = u_J)$
end
Return $\hat{\mu}^{(J)}$

Algorithm 5: Adaptive choice of K in covSDPofMeans

Unlike for the traditional Lepski's method we check that $\hat{\mu}^{(J)}$ is in $\bigcap_{j=0}^{J-1} B_2^d(\hat{\mu}^{(j)}, r^{(J)} + r^{(j)})$ instead of checking that $\bigcap_{j=0}^{J} B_2^d(\hat{\mu}^{(j)}, r^{(j)})$ is none-empty – this simplifies the adaptation step. It is also possible to speed up the whole procedure by constructing iteratively the bucketed means. Indeed, given that we consider a dyadic grid for K, i.e. $K \in \{N, \lceil N/2 \rceil, \lceil N/4 \rceil, \dots\}$, for all $j \in \mathbb{N}$, we can construct the block means $\{\bar{X}_k^{(j+1)}, k = 1, \dots, \lceil N/2^{j+1} \rceil\}$ at step $K = \lceil N/2^{j+1} \rceil$ using the block means from the previous step $K = \lceil N/2^j \rceil$ by simply averaging two successive block means: $\bar{X}_k^{(j+1)} \leftarrow (\bar{X}_{2k}^{(j)} + \bar{X}_{2k+1}^{(j)})/2$.

Let us now turn to the statistical analysis of the output $\hat{\mu}^{(\hat{J})}$ from Algorithm 5 where

$$\hat{J} = \max\left(J \in \{0, 1, \dots, \log_2 N\} : \hat{\mu}^{(J)} \in \bigcap_{j=0}^{J-1} B_2^d(\hat{\mu}^{(j)}, r^{(J)} + r^{(j)}) \right).$$

Theorem 3.4. *Let $\{u_j : j = 0, 1, 2, \dots, \log_2 N\} \subset \mathbb{N}^*$ be the family of parameters used to construct the family of estimators $\{\hat{\mu}^{(j)}, j = 0, 1, \dots\}$ in Algorithm 5 (see also (3.11)). For all $K \in \{600|\mathcal{O}|, \dots, N\}$, with probability at least*

$$1 - 2\exp(-K/360000) - \sum_{j=0}^{\log_2(N/(K-1))} (1/10)^{u_j} \tag{3.12}$$

the output $\hat{\mu}^{(\hat{J})}$ of Algorithm 5 is such that $\left\|\hat{\mu}^{(\hat{J})} - \mu\right\|_2 \leq 3r_K^*$.

Proof. For all $j \in \{0, 1, \dots, \log_2 N\}$ denote by $\mathcal{E}_j$ the event onto which Theorem 3.2 is valid for $K = \lceil N/2^j \rceil$ and for $u = u_j$: that is on $\mathcal{E}_j$, if $\lceil N/2^j \rceil \geq 300|\mathcal{O}|$, $\left\|\hat{\mu}^{(j)} - \mu\right\|_2 \leq r^{(j)}$ and $\mathbb{P}[\mathcal{E}_j] \geq 1 - \exp(-\lceil N/2^j \rceil/180000) - (1/10)^{u_j}$. Let $K \in \{600|\mathcal{O}|, \dots, N\}$ and $J \in \{0, 1, \dots, \log_2 N\}$ be such that $\lceil N/2^J \rceil \leq K < \lceil N/2^{J-1} \rceil$. On the event $\bigcap_{j=0}^{J} \mathcal{E}_j$, we have $\left\|\hat{\mu}^{(j)} - \mu\right\|_2 \leq r^{(j)}$ for all $j = 0, 1, \dots, J$, in particular, for all $j = 0, 1, \dots, J-1$, $\left\|\hat{\mu}^{(J)} - \hat{\mu}^{(j)}\right\|_2 \leq r^{(J)} + r^{(j)}$ and so $\hat{\mu}^{(J)} \in \bigcap_{j=0}^{J-1} B_2^d(\hat{\mu}^{(j)}, r^{(J)} + r^{(j)})$. As a consequence $\hat{J} \geq J$ therefore $\left\|\hat{\mu}^{(\hat{J})} - \hat{\mu}^{(J)}\right\|_2 \leq r^{(\hat{J})} + r^{(J)} \leq$

$2r^{(J)} \leq 2r_K^*$. Finally, we have

$$\mathbb{P}\Big[\bigcap_{j=0}^{J} \mathcal{E}_j\Big] \geq 1 - \sum_{j=0}^{J} \exp(-\lceil N/2^j \rceil/180000) - (1/10)^{u_j}$$
$$\geq 1 - 2\exp(-K/360000) - \sum_{j=0}^{\log_2(N/(K-1))} (1/10)^{u_j}.$$

■

We can see in Algorithm 5 that $\hat{\mu}^{(\hat{J})}$ does not use any information on the number of outliers $|\mathcal{O}|$ for its construction but it can still estimate μ at the optimal rate r_K^* for all deviation parameters K in $\{600|\mathcal{O}|, \dots, N\}$. The maximum total running time of Algorithm 5 is achieved when $\hat{J} = \log_2 N$; in that case, it is at most $\tilde{\mathcal{O}}(Nd + \sum_{j=0}^{\log_2 N} \lceil N/2^j \rceil u_j d)$. In particular, if one chooses $u_j = 2^j$ for all $j = 0, 1, \dots, \log_2 N$ then the total running time for the construction of $\hat{\mu}^{(\hat{J})}$ is nearly-linear $\tilde{\mathcal{O}}(Nd)$. For this choice of u_j, the probability deviation in (3.12) is constant and so one should choose the smallest possible K allowed in Theorem 3.4, that is $K = 600|\mathcal{O}|$. Let us write formally this result.

Corollary 3.2. *If one takes $u_j = 2^j$ for all $j = 0, 1, \dots, \log_2 N$ in Algorithm 5 then, in nearly-linear time $\tilde{\mathcal{O}}(Nd)$, with probability at least $1 - 2\exp(-600|\mathcal{O}|/360000) - 1/11$, the output $\hat{\mu}^{(\hat{J})}$ from Algorithm 5 satisfies*

$$\left\|\hat{\mu}^{(\hat{J})} - \mu\right\|_2 \leq 2r^*_{600|\mathcal{O}|} = 1616\left(1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + 850\sqrt{\frac{\|\Sigma\|_{op}|\mathcal{O}|}{N}}\right).$$

In particular, considering the setup from Theorem 3.1, if $|\mathcal{O}| = \epsilon N$ for some $\epsilon \leq 1/600$ then the rate achieved by $\hat{\mu}^{(\hat{J})}$ in Corollary 3.2 is of the order of

$$\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\|\Sigma\|_{op}\,\epsilon} \tag{3.13}$$

which is like $\sqrt{\|\Sigma\|_{op}\,\epsilon}$ when $N \geq (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/\epsilon$. As a consequence, the result from Corollary 3.2 improves the one from Theorem 3.1 by removing an extra $\log d$ factor in the sample complexity in the case considered in Theorem 3.1 that is when $\Sigma \preceq \sigma^2 I_d$. Moreover, Corollary 3.2 also shows that the sample complexity depends on the *effective rank* $\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op}$ of Σ. This ratio can be much smaller than d if the spectrum of Σ decays sufficiently fast. Finally, Corollary 3.2 also covers the case where the sample size N is less than the sample complexity – that is when $N \leq (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/\epsilon$. In that case, the estimation rate is given by $\sqrt{\mathrm{Tr}(\Sigma)/N}$ which is the complexity coming from the estimation of μ in the none corrupted case. As a consequence, Corollary 3.2 exhibits a phase transition happening at $N \sim (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/\epsilon$ above which corruption is the main source of estimation mistakes and below which corruption does not play any role.

Corollary 3.2 covers the case where $\hat{\mu}^{(\hat{J})}$ is computed in nearly-linear time and with statistical guarantees happening with constant probability. In the following final result, we show that $\hat{\mu}^{(\hat{J})}$ can estimate μ at the optimal rate r_K^* for all $K \geq 600|\mathcal{O}|$ with a subgaussian deviation $1 - 2\exp(-K/360000)$ if we perform more iterations u_j of the covering SDP from Lemma 3.5. The price we pay for this subgaussian behavior of $\hat{\mu}^{(\hat{J})}$ is on the total running time which goes from nearly-linear time $\tilde{\mathcal{O}}(Nd)$ to $\tilde{\mathcal{O}}(N^2 d)$ by taking $u_j = \lceil N/2^j \rceil$ for $j = 0, 1, \dots, \log_2 N$ ($u_j = N$ would do as well). We write formally this statement in the next corollary which follows directly from Theorem 3.4.

Corollary 3.3. *If one takes $u_j = \lceil N/2^j \rceil$ for all $j = 0, 1, \dots, \log_2 N$ in Algorithm 5 then, in time $\tilde{\mathcal{O}}(N^2 d)$, for all $K \geq 600|\mathcal{O}|$, with probability at least $1 - 4\exp(-K/360000)$, the output $\hat{\mu}^{(\hat{J})}$ from Algorithm 5 satisfies*

$$\left\|\hat{\mu}^{(\hat{J})} - \mu\right\|_2 \leq 2r_K^* = 1616\left(1200\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{1200\,\|\Sigma\|_{op} K}{N}}\right).$$

As a consequence $\hat{\mu}^{(\hat{J})}$ is a subgaussian estimator of μ for all range of K from $600|\mathcal{O}|$ to N which can handle up to $|\mathcal{O}|$ outliers in the database (even when $|\mathcal{O}| \sim N$) and that can be constructed in time $\tilde{\mathcal{O}}(N^2 d)$. It does not require any knowledge on $|\mathcal{O}|$ for its construction.

Let us now show that the algorithm $\hat{\mu}^{(\hat{J})}$ constructed in Corollary 3.3 also satisfies estimation results in expectation. So far all the statistical properties have been given with large probability; for $\hat{\mu}^{(\hat{J})}$ it is also possible to obtain a result in expectation.

The benchmark result we use here is the rate achieved by the empirical mean in a non-corrupted setup but unlike the result in deviation we don't need i.i.d. Gaussian variables since $\mathbb{E}\left\|\overline{\tilde{X}}_n - \mu\right\|_2 \leq \sqrt{\mathrm{Tr}(\Sigma)/N}$ where $\overline{\tilde{X}}_n = n^{-1}\sum_i \tilde{X}_i$ and $\tilde{X}_1, \dots, \tilde{X}_N$ are the non-corrupted data points from Assumption 3.1. Hence, $\sqrt{\mathrm{Tr}(\Sigma)/N}$ is the rate we aim to achieve but we also may expect a price to pay for the adversarial corruption, in particular, when $\epsilon = |\mathcal{O}|/N$ is above the phase transition exhibited in (3.13), that is for $\epsilon \geq (\mathrm{Tr}(\Sigma)/\|\Sigma\|_{op})/N$.

Theorem 3.5. *Under Assumption 3.1, and if $N \geq 600|\mathcal{O}|$, the following holds. If one takes $u_j = \lceil N/2^j \rceil$ for all $j = 0, 1, \dots, \log_2 N$ in Algorithm 5 then, in time $\tilde{\mathcal{O}}(N^2 d)$, Algorithm 5 outputs $\hat{\mu}^{(\hat{J})}$ satisfying*

$$\mathbb{E}\left\|\hat{\mu}^{(\hat{J})} - \mu\right\|_2 \leq (3 + 16c_0^2) r^*_{600|\mathcal{O}|} \leq (3 + 16c_0^2) 808 \times 1200\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op}|\mathcal{O}|}{2N}}\right)$$

as long as and $N \geq 4c_0 \log(c_0 d + c_0)$ where $c_0 = 360000$.

Proof. We denote $\tilde{\mu} = \hat{\mu}^{(\hat{J})}$ and $c_0 = 360000$. We know from Corollary 3.3 that for all $600|\mathcal{O}| \leq K \leq N$, with probability at least $1 - 4\exp(-K/c_0)$, $\|\tilde{\mu} - \mu\|_2 \leq 2r_K^*$. So we know how to control the estimation property of $\tilde{\mu}$ up to an event of probability measure at most $4\exp(-N/c_0)$. On that event, we only need a crude upper bound on $\|\tilde{\mu} - \mu\|_2$ to get the result. This is what we do now.

We know that by construction that $\tilde{\mu} \in B_2^d(\hat{\mu}^{(N)}, 2r_N^*)$. Moreover, $\hat{\mu}^{(N)}$ starts from $\hat{\mu}_0^{(N)}$, the coordinate wise median of the data X_i (because $K = N$ blocks here) and makes at most $T = \log(8\sqrt{d})/\log(1/0.81)$ descent iterations like $x_{c+1} = x_c - \theta_c v_1$ where $v_1 \in \mathcal{S}_2^{d-1}$ and $\theta_c = -\,\mathrm{Med}\left(\langle X_i - x_c, v_1\rangle : i \in [N]\right)$. In particular, one has at every iteration

$$\|x_{c+1} - \mu\|_2 \leq 2\|x_c - \mu\|_2 + \mathrm{Med}(\|X_i - \mu\|_2 : k \in [K]).$$

Hence, $\hat{\mu}^{(N)}$ satisfies

$$\begin{aligned}\left\|\hat{\mu}^{(N)} - \mu\right\|_2 &\leq 2^{T+1}\left(\left\|\hat{\mu}_0^{(j)} - \mu\right\|_2 + \mathrm{Med}(\|X_i - \mu\|_2 : i \in [N])\right)\\ &\leq 16d\left(\left\|\hat{\mu}_0^{(N)} - \mu\right\|_\infty + \mathrm{Med}(\|X_i - \mu\|_\infty : i \in [N])\right). \qquad (3.14)\end{aligned}$$

In the adversarial contamination model from Assumption 3.1, as we assumed that $N \geq 600|\mathcal{O}|$, there are at least $N - |\mathcal{O}| \geq (599/600)N$ indices i such that $X_i = \tilde{X}_i$, hence for at least $(599/600)N$

i's we have, for all $p \in [d]$,

$$|X_{i,p} - \mu_p| \leq \max_{i\in[N]} |\tilde{X}_{i,p} - \mu_p| \text{ and } \|X_i - \mu\|_\infty \leq \max_{i\in[N]} \left\|\tilde{X}_i - \mu\right\|_\infty$$

where $X_{i,p}$ (resp. μ_p) denotes the p-th coordinate of X_i (resp. μ). Hence, in (3.14), we get

$$\left\|\hat{\mu}^{(N)} - \mu\right\|_2 \leq 32d \max_{i\in[N]} \max_{p\in[d]} |X_{i,p} - \mu_p|.$$

Let us now turn to the stochastic argument to upper bound the right-hand side in the last inequality.

$$\mathbb{E}(\max_{i\in[N]} \max_{p\in[d]} |X_{i,p} - \mu_p|^2) \leq \mathbb{E}(\max_{i\in[N]} ||X_i - \mu||_2^2) \leq N \operatorname{Tr}(\Sigma).$$

Hence,

$$\mathbb{E}(\|\tilde{\mu} - \mu\|_2^2) \leq 2048 d^2 N \operatorname{Tr}(\Sigma) + 8(r_N^*)^2. \tag{3.15}$$

We are now in a position to obtain an estimation result in expectation for $\tilde{\mu}$. We denote $K_{\mathcal{O}} = 600|\mathcal{O}|$:

$$\begin{aligned}
\mathbb{E}\,\|\tilde{\mu} - \mu\|_2 &= \sum_{k=K_{\mathcal{O}}}^{N-1} \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(2r_k^* \leq \|\tilde{\mu} - \mu\|_2 \leq 2r_{k+1}^*)\right] \\
&\quad + \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \leq 2r_{K_{\mathcal{O}}}^*)\right] + \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \geq 2r_N^*)\right] \\
&\leq 2r_{K_{\mathcal{O}}}^* + \sum_{k=K_{\mathcal{O}}}^{N-1} 2r_{k+1}^* \times 4\exp(-k/c_0) + \mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \geq 2r_N^*)\right] \\
&\leq 2r_{K_{\mathcal{O}}}^* + 16c_0^2 r_{K_{\mathcal{O}}}^* \exp(-K_{\mathcal{O}}/c_0) + 25c_0 d\sqrt{N \operatorname{Tr}(\Sigma)} \exp(-N/(2c_0))
\end{aligned}$$

where, in the last inequality, we used that

$$\begin{aligned}
&\mathbb{E}\left[\|\tilde{\mu} - \mu\|_2 I(\|\tilde{\mu} - \mu\|_2 \geq 2r_N^*)\right] \leq \left(\mathbb{E}\left[\|\tilde{\mu} - \mu\|_2^2\right]\right)^{1/2} \left(\mathbb{P}\left[\|\tilde{\mu} - \mu\|_2 \geq 2r_N^*\right]\right)^{1/2} \\
&\leq (64d\sqrt{N \operatorname{Tr}(\Sigma)} + 3r_N^*) \times 2\exp(-N/(2c_0)) \leq 25c_0 d\sqrt{N \operatorname{Tr}(\Sigma)} \exp(-N/(2c_0))
\end{aligned}$$

from (3.15). When $N \geq 4c_0 \log(c_0 d + c_0)$, then $N \geq 2c_0 \log[c_0 dN]$, so $\mathbb{E}\,\|\tilde{\mu} - \mu\|_2 \leq (3 + 16c_0^2) r_{K_{\mathcal{O}}}^*$. ∎

We therefore recover the same rate of convergence in expectation in Theorem 3.5 as the one in deviation in Corollary 3.3 for the adaptive estimator $\hat{\mu}^{(\hat{J})}$, it is also the rate achieved by the non adaptive estimator $\hat{\mu}_K$ for the minimal value of $K = 600|\mathcal{O}|$. In particular, the same phase transition phenomena occurs in expectation as in the discussion following Equation (3.13).

CHAPTER 4

A Spectral Algorithm for Robust Regression with Subgaussian Rates

Contents

4.1 Introduction

Much work concerning the prototypical problem of regression focuses on the study of error rates of a given statistical procedure while making strong assumptions on the underlying distributions of samples, assuming for instance that they are i.i.d. and subgaussian or bounded (see for instance, Koltchinskii (2011); Massart (2007); Lecué and Mendelson (2013)). It is however of fundamental importance to understand what happens when the data violates such strong assumptions, for instance, when the underlying distribution of samples is *heavy-tailed* and/or when the dataset is corrupted by outliers. In such cases – which are everyday cases for real-world datasets – classical estimators such as OLS or MLE exhibit, at best, far-from-optimal statistical behaviours and at worst completely non-sens outputs. In this work, we study the statistical properties (non-asymptotic estimations and predictions results) of algorithms coming with actual working code constructed on this type of real-word datasets. We want to put forward that it is an algorithm and not only a purely theoretical estimator and that this algorithm can be

coded efficiently (we provide a simulation study in the following) since its most time consuming fundamental building block is to find a top singular vector of a reasonable size matrix. On top of these practical considerations, our theoretical results show that even though the dataset is far from the ideal i.i.d. subgaussian framework and even though we study an actually codable algorithm, the resulting estimator achieves the very same minimax bounds with (exponentially) high probability as the MLE/OLS does in the ideal i.i.d. Gaussian framework (i.e. Gaussian design and independent Gaussian noise), (see Lecué and Mendelson (2013) for deviation optimal result in the ideal framework). On top of that, we prove a theoretical running time for that algorithm which can be linear $\mathcal{O}(Nd)$ (where N is the sample size and d is the number of features) and at most quadratic $\mathcal{O}(N^2 d)$.

For a statistical problem such as mean estimation, regression or covariance estimation, we are given a loss function and an associated risk function ℓ (for instance, for the problem of estimation of the mean vector $\mu^* := \mathbb{E}(X) \in \mathbb{R}^d$ tackled in the general introduction, the loss function is $\ell_\mu(X) = \|\mu - X\|_2^2, \forall \mu \in \mathbb{R}^d$ and the associated risk is $\ell(\mu) = \mathbb{E}\ell_\mu(X)$). For robust estimators the emphasis is not put on the expected risk $\mathbb{E}(\ell(\hat{\mu}))$ – where the expectation is taken w.r.t. the data – but rather on the dependence of the risk bound r_δ on the confidence level $1-\delta \in [0,1]$: we want to find the smallest r_δ so that $\mathbb{P}(\ell(\hat{\mu}) > r_\delta) \leq \delta$ and the way r_δ depends on δ is paramount in this approach (this is a key property of the estimator $\hat{\mu}$ that cannot be revealed when its expected risk is studied). An estimator, as we have seen, is robust to heavy-tailed data if the rate r_δ does not grow "too quickly" when δ goes to 0: we look for the optimal "subgaussian-rate" defined in the Introduction. Here we consider the standard linear regression setting where data are couples $(X_i, Y_i)_i \in \mathbb{R}^d \times \mathbb{R}$ and we look for the best linear combination of the coordinates of an input vector X to predict the output Y, that is we look for β^* defined as follows.

$$\beta^* = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}}\, \ell(\beta) = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathbb{E}(Y_1 - \langle \beta, X_1 \rangle)^2.$$

The theoretical question of finding robust to heavy-tailed estimators reaching optimal rates for the regression problem has attracted much attention during the last ten years. It first started with the study the standard procedures in this heavy-tailed framework, such as Empirical Risk Minimization or its regularized versions Lecué and Mendelson (2016); van de Geer and Muro (2014); Lecué and Mendelson (pear); Oliveira (2016). Several results showed the negative but unavoidable impact of heavy-tailed data on these classical procedures Lecué and Mendelson (2016). In the mean time, new estimators have been introduced. For instance, the pioneer work of Audibert and Catoni (2011) has considered weak moment conditions, such as a $L_2 - L_4$ norm equivalence, under which the subgaussian rate could be reached. It was then followed by a rich literature such as Lugosi and Mendelson (2016); Lecué and Lerasle (2019); Lecué and Lerasle (2020); Oliveira (2016); van de Geer and Muro (2014). The remaining issue is that naive methods to compute these new theoretically-optimal estimators take exponential time in the number of dimension d, partly because some of them are based on non-convex optimization.

Recent advances have shown that, for the problem of mean estimation, one could find computationally efficient procedures (that is to say polynomial in both the dimension d and the number of data N) that are statistically nearly optimal, meaning that they reach -up to universal constants- the optimal radius $r_\delta = \sqrt{\frac{\mathrm{Tr}(\Sigma) + \|\Sigma\|_{op} \log(1/\delta)}{N}}$ for every confidence level $\delta \in [0,1]$ (see Hopkins (2018); Cherapanamjeri et al. (2019); Depersin and Lecué (2019)). More recently, Lei et al. (2020) introduce a spectral method reaching the optimal sub-gaussian rates without using Semi-Definite Programming, making somehow robust mean estimation easier to understand, easier to interpret and easier to code while still keeping optimal statistical and computational results.

The question of whether reaching similar bounds (matching the one of the OLS in the Gaussian setting without the Gaussian and i.i.d. assumptions – thus allowing for corrupted and heavy-tailed datasets) in polynomial time was possible for other statistical problems such as regression or covariance matrix estimation had been open for a long time. Indeed, up to recently, the best known polynomial algorithms were the one from Prasad et al. (2018) or from Hsu and Sabato (2016). The guarantee is the same for those two algorithms: when the covariance of X is the identity and when the noise $\xi = Y - \langle \beta^*, X \rangle$ has bounded variance $\ell(\hat{f}) - \ell(f^*) \leq \mathcal{O}(\frac{\log(1/\delta)d}{N})$ with probability $1-\delta$, and they need a number of sample of order $N \gtrsim \log(1/\delta)d$. The article Cherapanamjeri et al. (2020a) has been the first to construct a polynomial-time method achieving the rate of the OLS in the Gaussian setting $\ell(\hat{f}) - \ell(f^*) \leq \mathcal{O}(\frac{\log(1/\delta)\vee d}{N})$. To the date, it is the only procedure running in polynomial algorithm achieving the optimal subgaussian rate. However, Cherapanamjeri et al. (2020a) uses the Sum of Square (SoS) programming hierarchy to design their algorithm. Even if SoS hierarchy runs in polynomial time, its reliance on solving large semi-definite programs makes it impractical and remains a theoretical result leaving still open the question on the existence of a practical efficient algorithm achieving optimal subgaussian rates.

In this chapter, we tackle this issue, showing that techniques from Lei et al. (2020) can be used to give the first practical, nearly quadratic (and in fact in most cases nearly-linear) algorithm that reaches the subgaussian rate. We also conduct numerical experiments on simulated data with our proposed procedure to show that it is indeed practical and fast. Moreover, as predicted by our theoretical findings, our simulation analysis shows that it is robust both to heavy-tailed data and to outliers. To the best of our knowledge, this is the first time that numerical experiments are conducted for a regression algorithm with sub-gaussian rates and polynomial time guarantees.

From a theoretical point of view, our main result (that we will prove later) can be stated as follows (see Setting 4.1 for the precise set of assumptions and next sections for the construction of the algorithm).

Theorem 4.1. *There are universal constants A, B, C so that the following hold. Let $\delta \geq e^{-AN}$ and $K \geq B(\lfloor \log(1/\delta) \rfloor \vee d \vee |\mathcal{O}|)$ where $|\mathcal{O}|$ is the number of outliers. Given $N \geq K$ points, there is an algorithm running in time*

$$\mathcal{O}\left((Nd + K^2 d) \times \log(||\beta^*||_\Sigma) \times \text{polylog}(K, d)\right)$$

that outputs an estimate $\hat{\beta} \in \mathbb{R}^d$ such that with probability at least $1-\delta$

$$\ell(\hat{\beta}) - \ell(\beta^*) \leq C \frac{\sup_{u \in B_\Sigma} \mathbb{E}(\xi_1^2 \langle u, X_1 \rangle^2) K}{N}.$$

So for $K = B(\lfloor \log(1/\delta) \rfloor \vee d \vee |\mathcal{O}|)$, we get, up to universal constants the (deviation minimax optimal) subgaussian rate achieved by OLS in the Gaussian framework (see Lecué and Mendelson (2013)). This rate was achieved previously under similar assumptions by Median-of-means estimators in Lugosi and Mendelson (2016); Lerasle (2019); Lecué and Lerasle (2020); Lecué and Lerasle (2019) but none of them come with computational time guarantees.

To construct estimator $\hat{\beta}$ from Theorem 4.1 and to prove its theoretical properties as stated in Theorem 4.1, we outline now the role of the following key tools:

- The **Median of Means** framework, that has already been explained in the general introduction, and which is still a very important tool in this chapter.

- The **Furthest hyperplane problem** was first adapted to compute median-of-mean estimators very recently by Lei et al. (2020). Authors from Lei et al. (2020) adapt to the problem of robust mean estimation a procedure initially proposed by Karnin et al. (2012) to find the approximate furthest hyperplane, that is to say the hyperplane that separate 0 from *most of the data* and that is the furthest possible from 0. The method from Karnin et al. (2012) is based on the multiplicative weight update method (see Arora et al. (2012) for a survey), a technique which allows to compute efficiently approximations of quantities such as $\inf_{w_i \in \Delta} \sup_u \sum_i w_i \langle u, x_i \rangle^2$ where Δ is a convex set of positive weights.

The combination of these two techniques is at the heart of both the construction and the statistical and computational time studies of the algorithm satisfying Theorem 4.1.

In Section 4.2 we present the assumptions we make on the data and provide all the stochastic lemmas that will be needed for the algorithm. In Section 4.3 we will present our descent algorithm, give its precise statistical performance and make some connexion with the furthest hyperplane problem. In Section 4.4 we present some empirical results on simulated data.

4.2 Assumptions and preliminary stochastic results

4.2.1 Assumptions

As explained in the previous section, the observed dataset $(\tilde{X}_i, \tilde{Y}_i)_{i=1}^N \in \mathbb{R}^d \times \mathbb{R}$ is a corrupted version of the i.i.d. dataset $\{(X_i, Y_i)_i, i \in \{1, \dots, N\}\}$ in a possibly adversarial way. The assumptions made on good data $(X_i, Y_i)_i$ are gathered in the following setting: (see also Lerasle (2019) or Audibert and Catoni (2011)).

Setting 4.1. *We assume that the following "heavy-tailed setting" holds:*

1. *X_1 is centered and has finite second moments; we write its L^2-moments matrix $\Sigma = \mathbb{E}(X_1 X_1^T)$ and we assume that Σ is known. Let also $B_\Sigma = \{x \in \mathbb{R}^d | \langle x, \Sigma x \rangle \leq 1\}$ be the ellipsoid associated with this L_2 structure and, for $u \in \mathbb{R}^d$ $\|u\|_\Sigma^2 = \langle u | \Sigma u \rangle$.*

2. *Let $\xi_1 = Y_1 - \langle \beta^*, X_1 \rangle$ and assume that $\sigma^2 := \sup_{u \in B_\Sigma} \mathbb{E}(\xi_1^2 \langle u, X_1 \rangle^2)$ is such that $\sigma^2 < \infty$.*

3. *There exists an universal constant γ such that, for all $u \in \mathbb{R}^d$, $\gamma \mathbb{E}(\langle u, X \rangle^2) \geq \sqrt{\mathbb{E}(\langle u, X \rangle^4)}$.*

We assume adversarial contamination *on the data: $(X_1, Y_1), \cdots, (X_N, Y_N)$ denote N i.i.d. random vectors in $\mathbb{R}^d \times \mathbb{R}$. The vectors $(X_1, Y_1), \cdots, (X_N, Y_N)$ are not observed, instead, there exists a (possibly random) set $\mathcal{O}$ such that, for any $i \in \mathcal{O}^c$, $(\tilde{X}_i, \tilde{Y}_i) = (X_i, Y_i)$. The set of indices of outliers $\mathcal{O}$ can be arbitrarily correlated with the data (X_i, Y_i) – for instance, only the $9N/10$ data with the largest $\|X_i\|_2$ are observed – and the outliers $(\tilde{X}_i, \tilde{Y}_i)_{i \in \mathcal{O}}$ can be anything (they can be arbitrarily correlated between themselves and with the non-corrupted data $(X_i, Y_i), i = 1, \dots, N$). The only constraint on $\mathcal{O}$ is on its size: we suppose that we know an upper bound of $|\mathcal{O}|$ (even though, this constraint may be dropped out if we use an adaptive scheme on K such as Lepski's method in the end). The observed dataset is therefore $\{(\tilde{X}_i, \tilde{Y}_i) : i = 1, \cdots, N\}$, and we want to recover β^* out of it.*

Let us now comment on Setting 4.1. The first three assumptions deal with the heavy-tailed setup. It involves at most the existence of a fourth moment on the noise ζ and the functions class $\{u \in \mathbb{R}^d \to \langle u, X \rangle\}$. The strongest assumption among them is the third one which is a L_2/L_4 norm equivalence assumption. This type of assumption has been used from the beginning for the statistical study of ERM and other classical methods in the heavy-tailed scenario for instance in Oliveira (2016); van de Geer and Muro (2014); Lecué and Mendelson (pear) or in Audibert and

Catoni (2011). It is also related to the small ball assumption from Koltchinskii and Mendelson (pear). It has been systematically used for the study of Median-of-means estimators (see Lerasle (2019)). The remaining of Setting 4.1 deals with the adversarial contamination model, that has been presented in the introduction.

4.2.2 Bounds on three stochastic processes

In this section, we introduce three stochastic processes that play a central role in our analysis. We provide a high probability control for the supremum of the three of them into three lemmas. All the stochastic tools that we will need later will be related to one of the three processes. So that all the stochastic part of this work is gather into this section and in the end we will identify a single event onto which the study of the algorithm will be using purely deterministic arguments.

We now state the three lemmas. The two first one deal with the classical quadratic and multiplier processes which already appeared in the study of ERM in Lecué and Mendelson (2013). They naturally show up when the quadratic loss is used. The last one is new and is related to the descent algorithm we are studying below.

We split the data in K blocks denoted by $B_k, k \in \{1, \dots, K\}$, in agreement with the Median-of-Mean framework. We note $m = N/K$ the number of data in each blocks, and we set $\mathbf{X}_k = (X_i)_{i \in B_k}$ and $\tilde{\mathbf{X}}_k = (\tilde{X}_i)_{i \in B_k}$. $\mathbf{Y}_k$ and $\tilde{\mathbf{Y}}_k$ are defined the same way. We start with (Depersin, 2020a, Lemma 2), presented in greater details later in Chapter **??**, that we will use several times in what follows. We use the definition of VC-dimension presented in the general introduction for what follows.

Lemma 4.1. *Let $\mathcal{F}$ be a set of Boolean functions satisfying the following assumptions.*

- *For all $f \in \mathcal{F}$, $\mathbb{P}\left(f(\boldsymbol{X}_1, \boldsymbol{Y}_1) = 0\right) \geq 31/32$.*
- *$K \geq C(\mathrm{VC}(\mathcal{F}) \vee |\mathcal{O}|)$ where C is a universal constant.*

Then, with probability at least $1 - \exp(-K/512)$, for all $f \in \mathcal{F}$, there are at least $19K/20$ blocks B_k on which $f(\tilde{\boldsymbol{X}}_k, \tilde{\boldsymbol{Y}}_k) = 0$.

This lemma is used as a baseline to prove the three following lemmas that will define the three stochastic events $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{E}$ that are needed for our algorithm to give a good estimate. We state in this section that all three fail with exponentially low probability. We introduce the rate

$$r = 8\sigma\sqrt{\frac{K}{N}}. \tag{4.1}$$

Lemma 4.2 (**Multiplier process**). *There is a universal constant C_1 so that the following hold. If $K \geq C_1(d \vee |\mathcal{O}|)$, then the following event $\mathcal{E}$ holds with probability at least $1 - \exp(-K/512)$: for all $u \in B_\Sigma$, there exist more than $19/20K$ blocks B_k so that*

$$\frac{1}{m}\Big|\sum_{i \in B_k} \left(\tilde{Y}_i - \langle \beta^*, \tilde{X}_i \rangle\right) \langle u, \tilde{X}_i \rangle \Big| \leq r.$$

This can also be also written as: for all $u \in \mathbb{R}^d$ there exist more than $19/20K$ blocks B_k so that :

$$\frac{1}{m}\Big|\sum_{i \in B_k} \left(\tilde{Y}_i - \langle \beta^*, \tilde{X}_i \rangle\right) \langle u, \tilde{X}_i \rangle \Big| \leq r||u||_\Sigma.$$

Lemma 4.3 (**Quadratic process**). *There is C_1 a universal constant so that the following hold. If $K \geq C_1(d \vee |\mathcal{O}|)$ then the following event $\mathcal{B}$ holds with probability at least $1 - \exp(-K/512)$: for all $u, v \in \mathbb{R}^d$, there exists more than $19/20K$ blocks B_k so that*

$$|\frac{1}{m}\sum_{i\in B_k} \langle u, \tilde{X}_i\rangle \langle v, \tilde{X}_i\rangle - \langle u, \Sigma v\rangle| \leq 6\gamma\sqrt{\frac{1}{m}} \|u\|_\Sigma \|v\|_\Sigma$$

In particular, when $m \geq 360\ 000\gamma^2$, on the event $\mathcal{B}$, for all $u \in R^d$

$$99/100 \langle u, \Sigma u\rangle \leq \frac{1}{m}\sum_{i\in B_k} \langle u, \tilde{X}_i\rangle^2 \leq 101/100 \langle u, \Sigma u\rangle .$$

Lemma 4.4. *There is C_1 a universal constant so that the following hold. If $K \geq C_1(d \vee |\mathcal{O}|)$ and $m \geq 128\gamma$, then the following event $\mathcal{A}$ holds with probability at least $\exp(-K/512)$. For all $\beta_c \in \mathbb{R}^d$, there are more than $19/20K$ blocks B_k so that*

$$\left\|\tilde{Z}_k(\beta_c)\right\|_2 \leq 8\sqrt{\frac{\mathbb{E}(||(\xi_1\Sigma^{-1/2}X_1||_2^2)}{m}} + \sqrt{d}\|\beta_c - \beta^*\|_\Sigma \leq \sqrt{d}(r + \|\beta_c - \beta^*\|_\Sigma)$$

where

$$\tilde{Z}_k(\beta_c) = \frac{1}{m}\sum_{i\in B_k} (\tilde{Y}_i - \beta_c\tilde{X}_i)\Sigma^{-1/2}\tilde{X}_i$$

with r defined as in (4.1).

We assume for the rest of this work, that $K \geq C_1(d \vee |\mathcal{O}|)$ and $m \geq 360\ 000\gamma^2$. We moreover assume that events $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{E}$ hold.

4.3 Analysis of the algorithm

Starting from β_t, the ideal descent direction is $u^* = (\beta_t - \beta^*)/||\beta_t - \beta^*||_\Sigma$, and the associated step size is $||\beta_t - \beta^*||_\Sigma$. Of course, none of those two quantities can be exactly computed, but they give a sense of what one should look for : a good descent direction v should check $\langle v, \Sigma\beta_t - \beta^*\rangle \geq c_0||\beta_t - \beta^*||_\Sigma$ and $||v||_\Sigma = 1$ for some constant $c_0 < 1$, and a good step size should check $d_t \in [c_1\|\beta_t - \beta^*\|_\Sigma, c_0\|\beta_t - \beta^*\|_\Sigma]$ with $0 < c_1 < c_0$ so that, taking $\beta_{t+1} = \beta_t + d_t v_t$,

$$||\beta_{t+1} - \beta^*||_\Sigma^2 \leq (1 - 2c_0c_1 + c_1^2)||\beta_t - \beta^*||_\Sigma^2 \leq \alpha||\beta_t - \beta^*||_\Sigma^2$$

with $\alpha < 1$. In order to find a good descent direction, we will be using the central quantity

$$Z_k(\beta_c) = \frac{1}{m}\sum_{i\in B_k} (Y_i - \beta_c X_i)\Sigma^{-1/2}X_i$$

already mentioned in the previous section (see Lemma 4.4). Remember that we assume Σ to be known. We decompose Z_k as $Z_k(\beta_c) = \frac{1}{m}\sum_{i\in B_k}\xi_i\Sigma^{-1/2}X_i + \sum_{i\in B_k}\langle\beta^* - \beta_c, X_j\rangle\Sigma^{-1/2}X_i$. The first term has mean zero by definition of β^*, but the expectation of the second one is $\Sigma^{1/2}(\beta^* - \beta_c)$, so one might hope the $\Sigma^{-1/2}Z_k(\beta_c)$ to point toward the right direction.

In fact we can be a little more precise using the previous section. For any u such that $||u||_2 = 1$, we can see using Lemma 4.2 and 4.3 that, for at least $9/10K$ of the block, $\langle Z_k(\beta_c), u\rangle \simeq \langle\Sigma^{1/2}(\beta^* - \beta_c), u\rangle$ (up to errors of magnitude $\max(r, \frac{||\Sigma^{1/2}(\beta^* - \beta_c)||_2}{100})$). So if one were to find the vector u which maximise $\mathcal{Q}_{1/10}(\langle Z_k(\beta_c), u\rangle)_k$, where we denote by $\mathcal{Q}_{1/10}$ the

first decile of a sequence, we could find a vector u well aligned with $\Sigma^{1/2}(\beta^* - \beta_c)$. This unusual and non-convex maximisation has been tackled under the name of *furthest hyperplane problem* in Karnin et al. (2012), where it shown that one can approximately solve such a maximization using spectral methods. This method has been first used in the context of robust estimation by Lei et al. (2020), where authors use the spectral algorithm from Karnin et al. (2012) to solve efficiently the problem of robust mean estimation. The introduction of the quantity Z_k, which allows to adapt the procedures of Lei et al. (2020) to the regression problem, is one of the novelty of this work. We will see in the rest of this section that finding such a direction indeed leads to a nice descent, and we will show how to find it efficiently.

The general algorithm, as in Lei et al. (2020), is a basic descent procedure :

input : $\tilde{X}_1, \tilde{Y}_1 \ldots, \tilde{X}_N, \tilde{Y}_N$, $K \geq C_1(d \vee |\mathcal{O}|)$, and T_{des}.
output : A robust estimator of β^*

Initialize $\beta_0 = 0$
for $t = 1, ..., T_{des}$ **do**
  $d_t = \texttt{stepSize}(\tilde{X}, \tilde{Y}, K, \beta_t, T_{des})$
  $g_t = \texttt{descentDirection}(\tilde{X}, \tilde{Y}, K, \beta_t, d_t, T_{des})$
  $\beta_{t+1} = \beta_t - d_t g_t$
end
Return $\beta_{T_{des}}$.

Algorithm 6: Meta descent algorithm for robust linear regression

More precisely, we will show that the algorithms `stepSize` and `descentDirection` are good step size and descent direction. The main tool is a modification of the algorithm `APPROXBREGMAN` from Lei et al. (2020) (which is in turn an adaptation from Karnin et al. (2011)), that we called `BregmanRegression`

We summarize the properties of this descent direction in the following theorem :

Theorem 4.2. *On the event $\mathcal{E}, \mathcal{A}, \mathcal{B}$, each iteration of Algorithm 6 satisfies the following with probability at least $1 - \exp(-K)/T_{des}$*

- *Whenever $||\beta_c - \beta^*||_\Sigma \geq 100r$,*

$$\|\beta_{c+1} - \beta^*\|_\Sigma \leq (1 - 2/100.000) \|\beta_c - \beta^*\|_\Sigma$$

- *Whenever $||\beta_c - \beta^*||_\Sigma \leq 100r$,*

$$\|\beta_{c+1} - \beta^*\|_\Sigma \leq 102r$$

Moreover, each iteration runs in time $\mathcal{O}((Nd + K^2 d) \times \text{polylog}(d, K))$

Note that even if we are on the right set of event, our bound holds with a high probability, but not with probability 1. This is because our algorithm is stochastic in itself, and it has some chance to fail even if $\mathcal{E}, \mathcal{A}$, and $\mathcal{B}$ hold.

To prove this theorem, we need a few intermediate lemma and algorithms. All the results presented hold on the event $\mathcal{A} \cap \mathcal{B} \cap \mathcal{E}$. We first state some essential remarks about pruning. Because $\mathcal{A}$ holds, we know that $9/10K$ blocks check $\left\|\tilde{Z}_k(\beta_c)\right\|_2 \leq \sqrt{d}(r + \|\beta_c - \beta^*\|_\Sigma)$. For

simplicity, we will just denote $\tilde{Z}_k(\beta_c) = \tilde{Z}_i$. We set $K' = \lfloor 9/10K \rfloor$, and we note $Z'_1, ..., Z'_{K'}$ the K' smallest $\tilde{Z}_i$, as returned by algorithm 7. For the rest of this part we will mainly work with the pruned data, so that, on $\mathcal{A}$, $R := \max_{k \leq K'} ||Z'_k||_2 < \sqrt{d}(r + \|\beta_c - \beta^*\|_\Sigma)$.

input : $\tilde{Z}_1, ..., \tilde{Z}_K$
output : Pruned $\tilde{Z}_{\sigma(1)}, ..., \tilde{Z}_{\sigma(K')}$

1 Compute the norms $||\tilde{Z}_i||_2$ and sort them $\tilde{Z}_{\sigma(1)} < \tilde{Z}_{\sigma(2)} < ... < \tilde{Z}_{\sigma(K)}$
2 Remove the top 1/20
3 **Return** $\tilde{Z}_{\sigma(1)}, ..., \tilde{Z}_{\sigma(K')} := (Z'_k)_{k \in \{1,...,K'\}}$.

Algorithm 7: Pruning algorithm

The first lemma of this section states that if $\mathcal{Q}^{8/10}$ is the 8/10 quantile of a serie, $\max_{u \in \mathcal{B}_2^d} \mathcal{Q}^{8/10}(\langle Z'_i, u \rangle)$ is a good estimate of the distance $||\beta_c - \beta^*||_\Sigma$ ($\mathcal{B}_2^d$ denote the unit ball for the canonical euclidean distance on $\mathbb{R}^d$)

Lemma 4.5. *There is $u \in \mathcal{B}_2^d$ so that, for at least 8/10 of the $k \in \{1, ..., K'\}$*

$$\langle Z'_k, u \rangle \geq \theta_1$$

with $\theta_1 := 99/100||\beta_c - \beta^*||_\Sigma - r$

Moreover, for any $u \in \mathcal{B}_2^d$, at least 8/10 of the pruned blocks check, $\langle Z'_i, u \rangle \leq r + 101/100||\beta_c - \beta^*||_\Sigma$.

Now we recall the main lemma from Lei et al. (2020), that states that it is possible to approximate $\max_{u \in \mathcal{B}_2} \mathcal{Q}^{8/10}(\langle Z'_k, u \rangle)$ with exponentially high probability in polynomial time.

Lemma 4.6 (Lemma 5.2 of Lei et al. (2020))**.** *There is a universal constant C such that the following holds. Suppose there is $u \in \mathcal{B}_2^d$ so that, for at least 8/10 of the k*

$$\langle Z'_k, u \rangle \geq \theta > 0$$

and that, for all k, $Z'_k < R$. Then, when $T \geq 2\log(K')R^2/\theta^2$, with probability at least $1 - \exp(-T/C)$, algorithm 8 applied with T and θ outputs a vector $\tilde{u} \in \mathcal{B}_2^d$ so that, for at least 2/10 blocks, $\langle \tilde{u}, Z'_k \rangle \geq \theta/10$ (and returns "fail" with probability $\exp(-T/C)$). Moreover, each of the T iteration of algorithm 8 costs at most $K \times d + \text{polylog}(d)$ operations.

Remark: Algorithm 8 always return either a vector $u \in \mathcal{B}_2^d$ so that, for at least 2/10 of the k, $\langle u, Z'_k \rangle \geq \theta/10$ or "fail". If there is no u so that for at least 2/10 blocks $\langle u, Z'_k \rangle \geq \theta/10$, then it will always return "fail"

input : $Z'_1, \dots, Z'_K$, θ and T.
output : A good descent direction or "Fail".

$R = \max(||Z'_i||_2)$
Initialize weights $\omega_1 = (1, ..., 1)/K \in \mathbb{R}^K$
for $t = 1, ..., T$ **do**
  Let A_t be the $K \times d$ matrix whose i^{th} row is $\sqrt{\omega_t(i)} Z'_i$ and u_t be the approximate top right singular vector of $A_t \times \Sigma^{-1/2}$, computed with a `PowerMethod` (see Lei et al. (2020)).
  Set $\sigma_i = \langle Z'_i, u_t \rangle^2$.
  $\omega_{t+1}(i) = \omega_t(i) \times (1 - \sigma_i/2)$
  Normalize $a = \sum_i \omega_{t+1}(i)$, $\omega_{t+1} = \omega_{t+1}/a$
  Compute the Bregmann projection $\omega_{t+1} = \texttt{Bregmann}(\omega_{t+1})$
end
Return $\texttt{ROUND}(Z', \theta, (u_t)_t)$.

Algorithm 8: `BregmanRegression`

Remark 4.1. *Lemma 4.6 has a failure probability even if $\mathcal{A} \cap \mathcal{B} \cap \mathcal{E}$ holds: it is because Algorithm 8 calls two random algorithms,* `PowerMethod` *(see Lei et al. (2020)), which fails with constant probability, and* `ROUND`*, which fails with exponentially low probability $\propto \exp(-cT)$ with c a constant (Karnin et al. (2012); Lei et al. (2020)). Algorithm 8 can tolerate at most $0.1T$ among T mistakes in the computation of the top eigenvectors of the matrices A_t, and the event where more than $0.1T$ of the power methods fail happens with probability exponentially low in T. The failure probability of algorithm 8 and the algorithm itself are explained in depth in Lei et al. (2020).*

The computation of the Bregman projection is described in Barak et al. (2009), and appears to be a building block in Lei et al. (2020), the rounding algorithm is also given in Lei et al. (2020).

The following lemma states that finding a direction "aligned" with most of the Z'_k grants a good descent direction.

Lemma 4.7. *If for at least $2/10$ blocks , $\langle u, Z'_k \rangle \geq \theta/10$, then $v = \Sigma^{-1/2} u$ satisfies $\langle v, \Sigma\beta_c - \beta^* \rangle \geq \theta/10 - r - ||\beta_c - \beta^*||_\Sigma/100$ (and of course $||v||_\Sigma = 1$).*

Proof of Theorem 4.2. We now have all the right tools to perform our analysis.

- Whenever $||\beta_c - \beta^*||_\Sigma \geq 100r$, then by Lemma 4.5, there exists u so that for at least $8/10K'$ of the (pruned) blocks $\langle Z'_i, u \rangle \geq 98/100||\beta_c - \beta^*||_\Sigma$. So algorithm 8 with $\theta \in [49/100 \, \|\beta_c - \beta^*\|_\Sigma \, , 98/100 \, \|\beta_c - \beta^*\|_\Sigma]$, and with $T \geq 6 \log(K')K \geq 6 \log(K')d \geq 2 \log(K')R^2/\theta^2$ does not output "Fail" (Lemma 4.6).

  We also recall that if there is no u so that for at least $4/10$ blocks , $\langle u, Z_i \rangle \geq \theta/10$, then it will always return "Fail". Thus whenever $\theta \geq 10(101/100 \, \|\beta_c - \beta^*\| + r)$, by Lemma 4.5 , the algorithm returns "Fail".

  So our binary search `stepSize` returns a $\theta \in [49/100||\beta_c - \beta^*||_\Sigma \, , \, 10(102/100 \, \|\beta_c - \beta^*\|_\Sigma)] \times 2/100 \times (1/10) \times (100/102)$, in less than $\log(R/||\beta_c - \beta^*||_\Sigma) \lesssim \log(d)$ iterations. The vector u returned by `descentDirection` is so that $v = \Sigma^{-1/2} u$ checks $\langle v, \Sigma\beta_c - \beta^* \rangle \geq 2||\beta_c - \beta^*||_\Sigma/100$, with high probability (Lemma 4.7).

So we have, if $c_1 = 49/100 \times 1/10 \times 2/100 \times 100/102$ and $c_0 = 2/100$

$$\|\beta_{c+1} - \beta^*\|_\Sigma \leq (1 - 2c_0c_1 + c_1^2) \|\beta_c - \beta^*\|_\Sigma \leq (1 - 2/100.000) \|\beta_c - \beta^*\|_\Sigma$$

- Whenever $||\beta_c - \beta^*||_\Sigma \leq 100r$, whenever $\theta \geq 10(101/100 \|\beta_c - \beta^*\| + r)$, by Lemma 4.5 , the algorithm returns "Fail", so our binary search `stepSize` returns a $\theta \leq 10(102/100 \|\beta_c - \beta^*\|_\Sigma) \times 2/100 \times (1/10) \times (100/102) = 2/100 \|\beta_c - \beta^*\|_\Sigma)$. We have

$$\|\beta_{c+1} - \beta^*\|_\Sigma \leq 102/100 \|\beta_c - \beta^*\|_\Sigma \leq 102r$$

■

Once again, we recall that there is no effort made here to optimize the constants.

4.4 Experiments

In this section, we present the results of some synthetic numerical experiments. Our first aim is to show that our algorithm comes with actual code and that it can be computed efficiently. This is a important feature of our approach that we want to put forward because, even though there are polynomial time algorithms (even linear time ones for the problem of mean estimation) they usually do not come with efficient code. Our second aim is to show the robustness (to heavy-tailed and outliers) properties of our algorithms as predicted by our theoretical findings in Theorem 4.2.

4.4.1 Experiments with heavy-tailed data and outliers

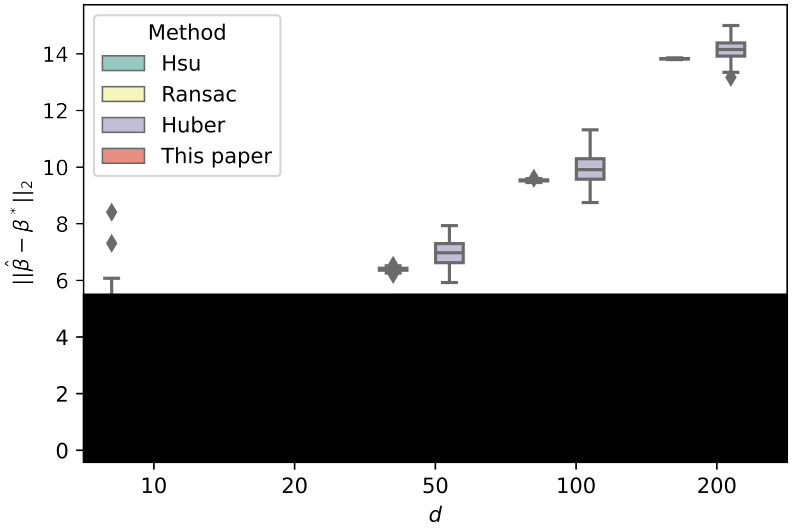
Data generating process. We fix the contamination level $\epsilon = |\mathcal{O}|/N$. Then, we generate $(1-\epsilon)N$ "clean" input vectors X_i following a multivariate Student's standard t-distribution with parameter 3 and we generate the corresponding "clean" responses following the linear model $Y = \langle \beta^*, X \rangle + \sigma\xi$ where $\beta^* = [1, \ldots, 1] \in \mathbb{R}^d$ and where ξ also follows Student's t-distribution and is independent from the feature vector X, and σ is the inverse signal to noise ration (SNR). We simulate an outliers attack by adding on the ϵN remaining data an arbitrary large number (10^9) to some cordinates of the input vectors, or multiplying them by 10^9. We also set some responses to 0 and some other to 10^9. The total number of samples is set to be $N = 50d$. We note that the sample size we choose increases with the dimension. We conduct 200 independent simulations.

Metric. We measure the parameter error in ℓ_2^d norm, which is also the estimation norm $\|.\|_\Sigma$ as we take $\Sigma = Id$.

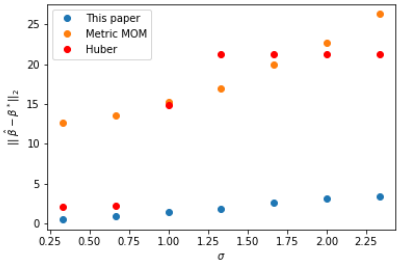
Baselines. As our baselines, we use the Ordinary Least Square, the Huber-loss M-estimator, RANdom SAmple Consensus (RANSAC) and the MOM-estimator from Hsu and Sabato (2016), that we name *metric MOM*. The first three are implemented in the python library sci-kit learn, and we coded the last one.

Results. We summarize our main findings here.

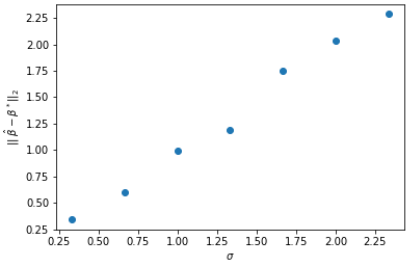
- Error vs dimension d: We fix $\epsilon = 0.005$, and we choose, for both our algorithm and the one from Hsu and Sabato (2016) to take $K = d$. We do not include the OLS in our graphic



(a) Parameter error vs d



(b) Mean parameter error vs σ



(c) Mean parameter error vs σ

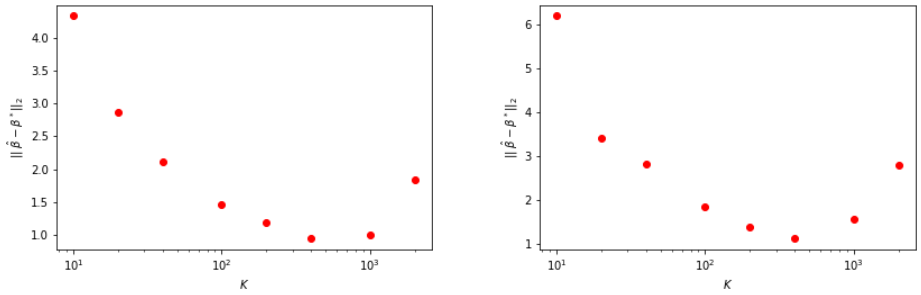
Figure 4.1: Parameter error variations

because its very poor performance (due to the presence of contamination) would prevent us to compare the four others. We notice that for all the algorithms but the one presented in this chapter, the prediction error grows quickly with the dimension. On the opposite, for our algorithm, the performance does not depend on the dimension. This does not come as a surprise, as the error is $\propto \sigma K/N$, which we chose to be d/N, which is a fixed quantity in this setup.

- Error vs the inverse SNR σ: We fix $\epsilon = 0.005$, $d = 200$, we still choose $K = d$ and we study how the algorithms perform for a range of SNR σ. We do not include OLS and we do not include RANSAC, because its error explodes for large σ. We notice that our algorithm's error depends linearly on σ, which is predicted by Theorem 4.2.

4.4.2 Which choice of K ?

From a theoretical point of view, we answered the question of how one should choose the parameter K in the previous section: K should me at least $K \geq C_1(d \vee |\mathcal{O}| \vee \log(1/\delta))$ for our algorithm to work with probability $\geq 1 - \delta$, but it should not be too large because we do not want our bound $\propto K/N$ to explode.



(a) Mean parameter error vs K ($d = 100$) (b) Maximum parameter error vs K

Figure 4.2: Choice of K

Setup. In Figure 4.2, we fix the contamination level $\epsilon = |\mathcal{O}|/N$ to be 0 (there is no outlier). Then, we generate the covariates of dimension $d = 100$ from a multivariate Student's t-distribution with parameter 3 and we generate the corresponding clean responses using $y = \langle \beta^*, x \rangle + \xi$ where $\beta^* = [1, \cdots, 1]$ and where ξ follows Student's t-distribution and is independent from the covariates. The number of samples is set to 10000. We conduct 50 independent simulations.

Results. We can recover a kind of trade-off from numerical experiment. It seems indeed that when $K \ll d$, our algorithm can not seize the complexity of the regression task, and that when $K \gg d$, there are not enough data per block and thus the block are "not informative enough". Those two opposite phenomenons lead to a sort of bias-variance trade-off.

4.5 Conclusion

We can outline the main benefits and limitations of our algorithm. On the practical side, the main benefit is its low computational complexity and that it comes with efficient actual code. On the theoretical side, the algorithm is robust to adversarial outliers and robust to heavy-tailed data and it achieves the subgaussian rate. It avoids the pitfall of SOS or SDPs since it uses spectral methods. This makes our algorithm both easy to understand easy to code, and that is the reason why this work comes with a simulation study unlike many other works in this literature.

The main limitation for now is that we need to know the variance matrix Σ of the co-variates (whereas sub optimal algorithms such as Hsu and Sabato (2016) do not require knowledge of Σ). An other limitation of this work lies in the choice of K: we need prior knowledge on the number of outliers for our procedure to work. It might be possible to improve this with a Lepski-type procedure Lerasle (2019).

A final comment is that, while we choose the descent procedure from Lei et al. (2020) for its simplicity and practical performances, the procedures from Depersin and Lecué (2019) or from Cherapanamjeri et al. (2020a) applied with our $\tilde{Z}_k$'s would probably work just as well and give similar rates but may be harder to code efficiently in practice.

An interesting perspective would be to extend this work to other estimation problems such as covariance estimation, as presented in Cherapanamjeri et al. (2020a). To do so, one would have

to find an efficient way to compute $\sup_{u\in\mathcal{B}_2}\sum_i \langle u, A_i u\rangle^2$ for any symmetric matrices A_i. While it is simple to compute $\sup_{u\in\mathcal{B}_2}\sum_i \langle u, v_i\rangle^2$ with the power method, this other problem seems harder. We may also wonder if it is possible to adapt this kind of spectral procedure in order to recover sparse signals or, more generally, if it is possible to introduce any regularisation to recover structured signal.

4.6 Proofs

4.6.1 Stochatic proofs

We state a theorem and its direct corollary that will be useful to bound the different VC-dimensions at stake.

Theorem 4.3 (Warren (1968)). *Let $P = \{P_1, ..., P_m\}$ denote a set of polynomials of degree at most ν in n real variables with $m > n$, then the number of sign assignments consistent for P is at most $(4e\nu m/n)^n$.*

We denote by $\mathbb{R}^n_\nu[X]$ the set of polynomials of dergree at most ν in n real variables.

Corollary 4.1. *Assume that the set of functions $\mathcal{F}$ can be written $\mathcal{F} = \{P \in \mathbb{R}^n_\nu[X] \to 1_{P(x)\geq 0}, x \in \mathbb{R}^n\}$, then $\mathrm{VC}(\mathcal{F}) \leq 2n\log_2(4e\nu)$.*

Let us also recall that, if $g : \mathcal{Y} \to \mathcal{X}$ is a function and $\mathcal{F}\circ g = \{f \circ g \mid f \in \mathcal{F}\}$, then $\mathrm{VC}(\mathcal{F}\circ g) \leq \mathrm{VC}(\mathcal{F})$.

Proof of Lemma 4.2. Let $\mathcal{F} = \{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{(d+1)\times m} \to \mathbf{1}_{\langle u, \sum_i (y_i - \langle \beta^*, x_i\rangle)x_i\rangle^2 \geq m^2r^2}, u \in B_\Sigma\}$. This is not a set of indicators of half-spaces, but $\mathcal{F}$ is the composition of $g : (\mathbf{x},\mathbf{y}) \in R^{(d+1)\times m} \to (u \to \langle u, \sum_i (y_i - \langle \beta^*, x_i\rangle)x_i\rangle^2 - m^2r^2) \in \mathbb{R}^d_2[X]$ and of $\{P \in \mathbb{R}^d_2[X] \to 1_{P(u)\geq 0}, u \in \mathbb{R}^d\}$. By Corollary 4.1 , there exists an absolute constant c such that $\mathrm{VC}(\mathcal{F}) \leq cd$.

For all $u \in B_\Sigma$,

$$\mathbb{P}\left(\frac{1}{m}\Big|\sum_{i\in B_1}(Y_i - \langle\beta^*, X_i\rangle)\langle u, X_i\rangle\Big| \geq r\right) \leq \frac{\mathbb{E}(\xi_1^2\langle u, X_1\rangle^2)}{mr^2} \leq \frac{1}{32}.$$

By Lemma 4.1 applied with $\mathcal{F}$, it follows that the following event $\mathcal{E}$ has probability $\geq 1 - \exp(-K/512)$: for all $u \in B_\Sigma$, there exist more than $3/4K$ blocks k where

$$\Big|\sum_{i\in B_k}(\tilde{Y}_i - \langle a, \tilde{X}_i\rangle)\langle u, \tilde{X}_i\rangle\Big| \leq mr.$$

■

Proof of Lemma 4.3. We note that, by bilinearity, it is enough to prove this result when $||u||_\Sigma = ||v||_\Sigma = 1$.

Let $\mathcal{G} = \{(\mathbf{x}_i) \in \mathbb{R}^{d\times m} \to \mathbf{1}_{|\sum\langle x_i,u\rangle\langle x_i,v\rangle - u\Sigma v|^2 \geq c||u||^2_\Sigma||v||^2_\Sigma}, u, v \in \mathbb{R}^d\}$. Once again, $\mathcal{G}$ is a composition of $g : (\mathbf{x},\mathbf{y}) \in R^{(d+1)\times m} \to (u, v \to |\sum\langle x_i, u\rangle\langle x_i, v\rangle - u\Sigma v|^2 - c||u||^2_\Sigma||v||^2_\Sigma) \in \mathbb{R}^{2d}_4[X]$ and of $\{P \in \mathbb{R}^{2d}_4[X] \to 1_{P(u)\geq 0}, u \in \mathbb{R}^d\}$, so there exists an absolute constant c such that $\mathrm{VC}(\mathcal{G}) \leq cd$ (Corollary 4.1).

Let $r_1 = 6\gamma\sqrt{\frac{1}{m}}\,\|u\|_\Sigma\,\|v\|_\Sigma$.

$$\mathbb{P}\left(|\frac{1}{m}\sum_{i\in B_1}\langle u,X_i\rangle\langle v,X_i\rangle - \langle u,\Sigma v\rangle| \geq r_1\right) \leq \frac{\mathbb{E}(\langle u,X_1\rangle^2\langle v,X_1\rangle^2)}{mr_1^2} \leq \frac{1}{32}$$

because $\mathbb{E}(\langle u,X_1\rangle^2\langle v,X_1\rangle^2) \leq \mathbb{E}(\langle u,X_1\rangle^4)^{1/2}\mathbb{E}(\langle v,X_1\rangle^4)^{1/2} \leq \gamma^2\|u\|_\Sigma^2\|v\|_\Sigma^2$ (this is from the L_2-L_4 norm equivalence). We conclude with Lemma 4.1.

■

Proof of Lemma 4.4. We define $Z_k(\beta_c) = \sum_{j\in B_k}(Y_j - \beta X_j)\Sigma^{-1/2}X_j$.

We can write $\|Z_k(\beta_c)\|_2 \leq \left\|\frac{1}{m}\sum_{j\in B_k}(Y_j-\beta^*X_j)\Sigma^{-1/2}X_j\right\|_2 + \left\|\frac{1}{m}\sum_{j\in B_k}((\beta^*-\beta_c)X_j)\Sigma^{-1/2}X_j\right\|_2$, we will bound those two quantities :

First $E((Y_j-\beta^*X_j)\Sigma^{-1/2}X_j) = 0$, so, if $a = 8\sqrt{\frac{\mathbb{E}(||(\xi_1\Sigma^{-1/2}X_1||_2^2)}{m}}$

$$\mathbb{P}(||\frac{1}{m}\sum_{j\in B_1}(Y_j-\beta_cX_j)\Sigma^{-1/2}X_j||_2 \geq a) \leq \frac{\mathbb{E}(||(\xi_1\Sigma^{-1/2}X_1||_2^2)}{ma^2} \leq \frac{1}{64}$$

Then, if we note $V_k = \langle(\beta^*-\beta_c)X_j\rangle\Sigma^{-1/2}X_j$ we notice that $\mathbb{E}(V_k) = \Sigma^{1/2}(\beta^*-\beta_c)$, and that $\mathbb{E}(||V_k||^2) \leq \mathbb{E}(\langle\Sigma^{-1/2}X,\Sigma^{-1/2}X\rangle^2)^{1/2}\,\mathbb{E}(\langle\beta_c-\beta^*,X\rangle^2)^{1/2}$. As X checks the L_4-L_2 norm equivalence, $\Sigma^{-1/2}X$ checks the same equivalence, so $\mathbb{E}(\left\|\Sigma^{-1/2}X\right\|_2^4)^{1/2} \leq \gamma\mathbb{E}(\left\|\Sigma^{-1/2}X\right\|_2^2) = \gamma d$, and $\mathbb{E}(\langle\beta_c-\beta^*,X\rangle^2)^{1/2} = \|\beta_c-\beta^*\|_\Sigma$, so

$$\mathbb{E}(||\frac{1}{m}\sum_{i\in B_1}V_i||_2^2) = ||\mathbb{E}(V_1)||_2^2 + \frac{1}{m}\mathbb{E}(||V_i-\mathbb{E}(V_i)||_2^2) \leq ||\mathbb{E}(V_1)||_2^2 + \frac{1}{m}\mathbb{E}(||V_i||_2^2) \leq \|\beta_c-\beta^*\|_\Sigma^2 + \frac{1}{m}\gamma d\,\|\beta_c-\beta^*\|_\Sigma^2$$

So, as $m \geq 128\gamma$, if $b = \sqrt{d}\,\|\beta_c-\beta^*\|_\Sigma$

$$\mathbb{P}(||\frac{1}{m}\sum_{i\in B_1}V_i|| \geq b) \leq \frac{1}{64}$$

So the probability that one of the two bounds fails is $\leq 1/32$. We then just use lemma 4.1, with the functions $\mathcal{F} = \{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{(d+1)\times m} \to \mathbf{1}_{||\sum_i(y_i-\langle\beta,x_i\rangle)x_i||^2 \geq d(r^2+\|\beta_c-\beta^*\|_\Sigma^2)}, \beta \in \mathbb{R}^d\}$. Again, we use Corollary 4.1 to state that there exists an absolute constant c such that $\mathrm{VC}(\mathcal{G}) \leq cd$. ■

4.6.2 Algorithmic proofs

Proof of Lemma 4.5. In fact, we just know that, if we take $u = \frac{\Sigma^{1/2}(\beta_c-\beta^*)}{(||\beta_c-\beta^*||_\Sigma)}$, and $v = \frac{(\beta_c-\beta^*)}{(||\beta_c-\beta^*||_\Sigma)} \in B_\Sigma$

$$\langle\tilde{Z}_i,u\rangle = \sum_{i\in B_k}(\tilde{Y}_i - \langle\beta^*,\tilde{X}_i\rangle)\langle v,\tilde{X}_i\rangle + \sum_{i\in B_k}(\langle\beta^*-\beta_c,\tilde{X}_i\rangle)\langle v,\tilde{X}_i\rangle \tag{4.2}$$

So for at least 9/10 blocks, $\langle\tilde{Z}_i,u\rangle \geq 99/100||\beta_c-\beta^*||_\Sigma - r := \theta_1$. This is true for at least 9/10 of the blocks $(\tilde{Z}_i)$, it is true for at least $17/19 > 8/10$ of the "pruned blocks" (Z_i').

The same way, for any $u \in \mathcal{B}_2$, we take $v = \Sigma^{-1/2}u \in \mathcal{B}_\Sigma$

$$\begin{aligned}\langle\tilde{Z}_i,u\rangle &= \sum_{i\in B_k}(\tilde{Y}_i - \langle\beta^*,X_i\rangle)\langle v,X_i\rangle + \sum_{i\in B_k}(\langle\beta^*-\beta_c,X_i\rangle)\langle v,X_i\rangle\\ &\leq r + \langle\beta^*-\beta_c,\Sigma v\rangle + 1/100||\beta_c-\beta^*||_\Sigma\\ &\leq r + 101/100||\beta_c-\beta^*||_\Sigma\end{aligned}$$

for at least 9/10 of the blocks. Again, as this is true for at least 9/10 of the blocks, it is true for at least $17/19 > 8/10$ of the "pruned blocks"

■

Proof of Lemma 4.7.

$$\begin{aligned}\langle \tilde{Z}_i, u\rangle &= \sum_{i\in B_k} (\tilde{Y}_i - \langle \beta^*, X_i\rangle) \langle v, X_i\rangle + \sum_{i\in B_k} (\langle \beta^* - \beta_c, X_i\rangle) \langle v, X_i\rangle \\ &\leq r + \langle \beta^* - \beta_c, \Sigma v\rangle + 1/100 ||\beta_c - \beta^*||_\Sigma\end{aligned}$$

for at least 9/10 of the blocks $\tilde{Z}_i$. Again, as this is true for at least 9/10 of the blocks, it is true for at least $17/19 > 8/10$ of the "pruned blocks" Z_i'.

Their is at least one block that checks both $\langle u, Z_i'\rangle \geq \theta/10$ and $\langle u, Z_i'\rangle \leq r + \langle \beta^* - \beta_c, \Sigma v\rangle + 1/100||\beta_c - \beta^*||_\Sigma$ (as $2/10 + 17/19 > 1$), so

$$\langle \beta^* - \beta_c, \Sigma v\rangle \geq \theta/10 - r - ||\beta_c - \beta^*||_\Sigma/100$$

■

4.7 Appendix

input : $\tilde{Z}_1, \ldots, \tilde{Z}_K$, θ and $u_1, ..., u_T$.
output : u.

while $\langle Z_i, u\rangle \leq \theta/10$ *for more than* $0.6K$ *blocks* **do**
  $g_j \sim \mathcal{N}(0,1)$ for $j \in \{1, ..., T\}$
  $u = \sum_j g_j u_i / ||\sum_j g_j u_i||$
  Report "Fail" and exit if more than T trials have been performed
end
Return u.

Algorithm 9: `Round`

input : $\tilde{X}_1, \tilde{Y}_1 \ldots, \tilde{X}_N, \tilde{Y}_N$, β_c, $K \geq |\mathcal{O}|$, T_{des}
output : A good distance estimation, d_t

Let, for $i \leq K$, $\tilde{Z}_i = \frac{1}{m} \sum_{j \in B_i} (\tilde{Y}_j - \beta_c \tilde{X}_j) \Sigma^{-1/2} \tilde{X}_j$
$Z' = \texttt{prune}(\tilde{Z})$
$R = \max(\tilde{Z}'_i)$
$d_{high} = R,\ d_{low} = 0$
for $j \in \{1, 2, ..., \lfloor \log(K) \rfloor\}$ **do**
    $d_m = (d_{high} + d_{low})/2$
    if *BregmanRegression*$(Z', d, \log(T_{des}) + \log(K)K)$ *returns "Fail"* **then**
        $d_{high} \leftarrow d_m$
    end
    else
        $d_{low} \leftarrow d_m$
    end
end
Return $d_{low} \times 2/100 \times (1/10) \times (100/102)$.

Algorithm 10: `stepSize`

input : $\tilde{X}_1, \tilde{Y}_1 \ldots, \tilde{X}_N, \tilde{Y}_N$, β_c, $K \geq |\mathcal{O}|$, T_{des}, θ.
output : u.

Let, for $i \leq K$, $\tilde{Z}_i = \frac{1}{m} \sum_{j \in B_i} (\tilde{Y}_j - \beta_c \tilde{X}_j) \Sigma^{-1/2} \tilde{X}_j$
`prune`$(\tilde{Z})$
$u =$ `BregmanRegression`$(\tilde{Z}, \theta \times 100/2 \times (10) \times (102/100), \log(T_{des}) + \log(K)K)$
Return $\Sigma^{-1/2} u$.

Algorithm 11: `descentDirection`

CHAPTER 5

Robust subgaussian estimation with VC-dimension

Contents

5.1 Introduction

We stated in the general introduction that the subgaussian rate for the sparse mean estimation problem (5.2) described below is different from (5.1): the "complexity term" (the one that does not depend on δ) goes from d to $s\log(d/s)$. Can this rate be reached only assuming second order moment on the random variables at stake ?

$$\left(\frac{d}{N}+\frac{\log(1/\delta)}{N}\right)^{1/2} \tag{5.1}$$

$$\left(\frac{s\log(d/s)}{N} + \frac{\log(1/\delta)}{N}\right)^{1/2} \tag{5.2}$$

In this work, we show that the analysis presented in Lugosi and Mendelson (2019c), in Lecué and Lerasle (2019), Lerasle (2019) or in Lecué and Lerasle (2020), all based on the Median-of-mean principle and the use of Rademacher complexities, can be modified in order to achieve subgaussian rates for sparse or structured problems assuming only bounded two-order moments. The method developed in Lerasle (2019) or in Lecué and Lerasle (2020) requires data to have at least $\log(d)$ finite moments (where d is the dimension of the space) in order to exploit the sparsity of the problem and offers no guarantees without that requirement, and to the date is the best known. We show that we can drop this condition by judiciously introducing VC-dimension in the different proofs, and exploit the sparsity of the problem *with only two moments*. Classical approaches using local Rademacher complexities cannot achieve this type of subgaussian bounds under only a second moment assumption in this setup. Indeed, as shown in the counter-example from Section 3.2.3 of Chinot et al. (2018), local Rademacher complexities may scale like $d^{1/8}$ whereas the right Gaussian bound should be of the order of $\sqrt{\log d}$. Somehow the classical approach used so far does not capture the right statistical complexity of high-dimensional problems under low-dimensional structural assumptions and under only a second moment assumption : **it seems that the Rademacher complexity is not the right way to measure the complexity of the problem of mean estimation in any norm**. Our VC-dimension based approach allows to overcome this issue and to go beyond this $\log d$ subgaussian moments assumption that has appeared in all works on robust and subgaussian estimation in the high-dimensional framework Lerasle (2019). We also show that this general technique can be easily replicated and give new robust estimators that achieve state-of-the-art bounds for different estimation tasks such as:

- Regression, already studied in Lecué and Lerasle (2020) where our estimator's rate match the one from Lecué and Lerasle (2020), and sparse regression where our estimator's rate is the first to match the one from Lecué and Lerasle (2020) *with only two moments*.
- Mean estimation with non-Euclidean norms, studied in Lugosi and Mendelson (2018), where our analysis gives a different rate that is better for some norms.
- Robust low-rank matrix estimation.
- Covariance estimation, studied in Mendelson and Zhivotovskiy (2018) under $L_4 - L_2$ norm equivalence: we do not need this assumption with our analysis, thus we give the first subgaussian estimator without this assumption.

This paper is not the first to introduce VC-dimension in robust estimation problems: we have been inspired by Chen et al. (2018) and Gao (2017) for instance. In those two papers, estimation and regression with possible sparsity and outliers are also achieved with optimal rates, using VC dimension techniques. The main differences lie in the model assumptions. For example, Chen et al. (2018) estimates the center of *symmetric distributions* without moment assumption. In comparison, our estimators is for mean and covariance, and thus moment assumption is needed, but *we do not need the distributions at stake to be symmetric.*

Using VC-dimension in mean estimation, we lose a nice dependence of the risk bounds in the covariance structure: our rates for (non-sparse) mean estimation depend on the ambient dimension d instead of the effective rank $\mathrm{Tr}(\Sigma)/||\Sigma||_{op}$. In particular, the general approach does not generalize directly to infinite dimensional spaces. In the last section, we show that this issue can be overcome if we have some knowledge on the covariance matrix.

5.2 Warm-up: MOM principle, VC-dimension and mean estimation

We start with the mean estimation problem in $\mathbb{R}^d$ that illustrates our technique: the goal is to estimate the mean $\mathbb{E}[Y]$ of a random vector Y in $\mathbb{R}^d$ given a possibly corrupted dataset of i.i.d. copies of Y. The precise setting is the following:

Setting 5.1. *Let $(Y_1, ..., Y_N)$ denote N independent and identically distributed random vectors in $\mathbb{R}^d$. We want to estimate $\mathbb{E}(Y_1) = \mu$, assuming that Y_1 has finite second moment. Let $\Sigma = \mathbb{E}((Y_1 - \mu)(Y_1 - \mu)^T)$ denote the unknown covariance matrix of Y_1.*

The vectors $Y_1, \dots, Y_N$ are not observed, instead, this dataset may have been corrupted, and this corruption may be adversarial: there exists a (possibly random) set $\mathcal{O}$ such that, for any $i \in \mathcal{O}^c$, $X_i = Y_i$. The $\mathcal{O}$ satisfies $|\mathcal{O}| \leq \lfloor \varepsilon N \rfloor$

The observed dataset is $\{X_i : i = 1, ..., N\}$, and we want to recover μ.

Notice that there are no assumption on the data $\{X_i, i \in \mathcal{O}\}$. In particular these may be dependent of $\{Y_i : i = 1, ..., N\}$, and the $\{X_i : i \in \mathcal{O}\}$ may have arbitrary dependence structure.

5.2.1 VC-dimension

We start this part by recalling some basic facts about VC-dimension that appear for instance in Ahsen and Vidyasagar (2019).

Definition 5.1. *Let $\mathcal{F}$ be a set of Boolean functions on any space $\mathcal{X}$. We say that a finite set $S \subset \mathcal{X}$ is shattered by $\mathcal{F}$ if, for every subset $B \subset S$, there exists $f \in \mathcal{F}$ such that $S \cap f^{-1}(\{1\}) = B$. We call VC-dimension of $\mathcal{F}$ (and note $\mathrm{VC}(\mathcal{F})$) the largest integer n such that there exists a set S of cardinality n that is shattered by $\mathcal{F}$.*

Whenever E is a Euclidean space, we will sometimes abusively call VC-dimension of a set $C \subset E$ and note $\mathrm{VC}(C)$ the VC-dimension of the set of half-spaces generated by the vectors of C:

$$\mathrm{VC}(C) = \mathrm{VC}(\{x \in E \to \mathbf{1}_{\langle x, v \rangle \geq 0},\ v \in C\}).$$

Let us recall some basic facts about VC dimensions.

1. $\mathrm{VC}(\mathbb{R}^d) = d + 1$. More generally, if F a set of real-valued functions in a k-dimensional linear space, then $\mathrm{Pos}(F) = \{x \to \mathbf{1}_{f(x) \geq 0}, f \in F\}$ has VC-dimension $k+1$ (see for instance Dudley (1978), Theorem 7.2).

2. For a function $g : \mathcal{Y} \to \mathcal{X}$, if we note $\mathcal{F} \circ g = \{f \circ g \mid f \in \mathcal{F}\}$, then we have $\mathrm{VC}(\mathcal{F} \circ g) \leq \mathrm{VC}(\mathcal{F})$.

3. For any $r > 0$, $\mathrm{VC}(\{x \in E \to \mathbf{1}_{\langle x, v \rangle \geq r}, v \in C\}) \leq \mathrm{VC}(C - C) \lesssim \mathrm{VC}(C)$, see Section 5.6.

4. Sauer's Lemma Sauer (1972): Let $\mathcal{F}$ denote a set a functions with VC-dimension ν and let S be a set of $n \geq s$ points. Let $\mathcal{F} * S = \{S \cap f^{-1}(\{1\}), f \in \mathcal{F}\}$, then

$$\mathrm{Card}(\mathcal{F} * S) \leq \left(\frac{en}{\nu}\right)^{\nu}.$$

This last lemma can be used to prove the following result that is useful to bound the VC dimension of the set of sparse vectors.

Lemma 5.1. *Let $\mathcal{F}_1$, ..., $\mathcal{F}_n$ denote n sets of boolean functions, each having VC-dimension $\leq \nu$. Then,*

$$\mathrm{VC}(\mathcal{F}_1 \cup \mathcal{F}_2 \cup ... \cup \mathcal{F}_n) \leq 4\nu + 2\log_2(n).$$

Lemma 5.1 is a straightforward extension of Theorem 3 in Ahsen and Vidyasagar (2019).

Proof. Let S be a set shattered by $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup ... \cup \mathcal{F}_n$, and $s = \mathrm{Card}(S)$. Because S is shattered, we have $\mathrm{Card}(\mathcal{F} * S) = 2^s$. But we also have $\mathcal{F} * S = \mathcal{F}_1 * S \cup \mathcal{F}_2 * S \cup ... \cup \mathcal{F}_n * S$, so $\mathrm{Card}(\mathcal{F} * S) \leq n \left(\frac{es}{\nu}\right)^\nu$.

It follows that $2^s \leq n \left(\frac{es}{\nu}\right)^\nu$ or $s \leq \nu \log_2(esn^{1/\nu}/\nu)$. By technical Lemma 4.6 in Vidyasagar (1997) if $x \leq a\log_2(bx)$, then $x \leq 2a\log_2(ab)$. Hence, $s \leq 2\nu\log_2(en^{1/\nu})$, which implies Lemma 5.1. ∎

Corollary 5.1. *Fix $v_1, ..., v_d \in \mathbb{R}^n$ and note $\mathcal{U}_s = \{\sum_i \lambda_i v_i \mid \lambda_i \in \mathbb{R} \ \& \ \sum_i \mathbf{1}_{\lambda_i \neq 0} \leq s\}$ the set of s-sparse vectors, then*

$$\mathrm{VC}(\mathcal{U}_s) \leq 4s\log_2(ed/s).$$

To prove Corollary 5.1, just write the set $\mathcal{U}_s$ as a union of $\binom{d}{s}$ s-dimensional subspaces. As a side remark, we note that Ahsen and Vidyasagar (2019) also shows that this bound is tight up to multiplicative constants: there exists an absolute constant c such that $\mathrm{VC}(\mathcal{U}_s) \geq cs\log_2(ed/s)$. Besides, the result holds even if the set of vectors $(v_1, ..., v_d)$ is not an orthogonal family or if it is not a base. Let us now recall an important theorem that will be very useful in regression and covariance estimation. Let $P = \{P_l, ..., P_m\}$ denote a set of multivariate polynomials. A sign assignment is an element s of $\{+, -\}^m$. The sign assignment s is consistent with P if there exists $x \in \mathbb{R}^n$ such that $P_i(x) \geq 0 \Leftrightarrow s_i = +$.

Theorem 5.1 (Warren, Warren (1968))**.** *Let $P = \{P_1, ..., P_m\}$ denote a set of polynomials of degree at most ν in n real variables with $m > n$, then the number of sign assignments consistent for P is at most $(4e\nu m/n)^n$.*

Corollary 5.2. *Assume that the set of functions $\mathcal{F}$ can be written $\mathcal{F} = \{P \in \mathbb{R}^n_\nu[X] \to 1_{P(x)\geq 0}, x \in \mathbb{R}^n\}$, then $\mathrm{VC}(\mathcal{F}) \leq 2n\log_2(4e\nu)$.*

The following example will be useful in some applications (we note that this is not a novelty, a similar result can be found, for instance, in Wolf et al. (2007), Theorem 2).

Proposition 5.1. *Let $r \geq 0$ and call $\mathcal{M}_d^k(\mathbb{R})$ the set of rank k, symmetric, d-dimensional matrices.*

Let $\mathcal{F} = \{M \in \mathcal{M}_d(\mathbb{R}) \to 1_{\langle X, M\rangle \geq r}, X \in \mathcal{M}_d^k(\mathbb{R})\}$. Then $\mathrm{VC}(\mathcal{F}) = \mathrm{VC}(\mathcal{M}_d^k(\mathbb{R})) \leq 2(d+1)k\log_2(12e)$.

Proof. Any $X \in \mathcal{M}_d^k(\mathbb{R})$ can be written $X = \sum_{i=1}^k \lambda_i\, x_i x_i^T$, with $(\lambda_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$. Besides, for any M, the function $(\lambda_i, x_i)_{i\leq k} \to \langle X, M\rangle - r$ is a polynomial of degree 3 in $k(d+1)$ variables. Hence, the result follows from Corollary 5.2. ∎

Combining Lemma 5.1 applied with rank-one matrices of the form $xx^T, x \in \mathbb{R}^s \times \{0\}^{d-s}$ and Corollary 5.2 yields the following result.

Proposition 5.2. *Let $\mathcal{F} = \{M \in \mathcal{M}_d(\mathbb{R}) \to 1_{\langle xx^T, M\rangle \geq r}, x \in \mathcal{U}_s\}$. Then $\mathrm{VC}(\mathcal{F}) \leq 16s\log_2(ed/s)$.*

5.2.2 Median-of-mean

This work uses the median-of-means (MOM) approach which was introduced in Nemirovsky and Yudin (1983); Alon et al. (1999); Jerrum et al. (1986) and has received a lot of attention recently in the statistical and machine learning communities Bubeck et al. (2013); Lerasle and Oliveira (2011); Devroye et al. (2016); Minsker and Strawn (2017); Minsker (2015). This approach allows to build estimators that are robust to both outliers and heavy-tail data in various settings Alon et al. (1999); Jerrum et al. (1986); Birgé (1984). It can be defined as follows: we first randomly split the data into K blocks $B_1, \dots, B_K$ of equal-size m (if K does not divide N, we just remove some data). We then compute the empirical mean within each block: for $k = 1, \dots, K$,

$$\bar{X}_k = \frac{1}{m} \sum_{i \in B_k} X_i.$$

In the one-dimensional case, the final estimator is the median of the latter K empirical means. This estimator has subgaussian deviations as shown in Devroye et al. (2016). The extension of this result to higher dimensions is not trivial as there exist several possible generalizations of the one dimensional median, see Minsker (2015).

For any $k \in \{1, \dots, K\}$, let $\mathbf{X}_k := (X_i)_{i \in B_k}$ and $\mathbf{Y}_k := (Y_i)_{i \in B_k}$. We start with a basic observation.

Remark 5.1. *When $K \geq |\mathcal{O}|$, there is at least $K - |\mathcal{O}|$ blocks B_k on which $\boldsymbol{X}_k = \boldsymbol{Y}_k$.*

For instance, if $K \geq 4|\mathcal{O}|$, then, there exist at least three quarters of the blocks B_k where $\mathbf{X}_k = \mathbf{Y}_k$. We can now state the main lemma.

Lemma 5.2. *Let $\mathcal{F}$ be a set of Boolean functions satisfying the following assumptions.*

- *For all $f \in \mathcal{F}$, $\mathbb{P}(f(\boldsymbol{Y}_1) = 0) \geq 15/16$.*
- *$K \geq C(\mathrm{VC}(\mathcal{F}) \vee |\mathcal{O}|)$ where C is a universal constant.*

Then, with probability $\geq 1 - \exp(-K/128)$, for all $f \in \mathcal{F}$, there is at least $3K/4$ blocks B_k on which $f(\boldsymbol{X}_k) = 0$.

In words, if each property f is true for one non corrupted block with constant probability (here 15/16 but it could be any fixed constant $\alpha \geq 1/2$) and K is large enough, then, with very high probability, all properties are "true for most of the blocks". The Boolean functions that we will consider to construct estimators will measure whether the mean of the block is far from the true mean. For instance, for mean estimation, we take the set

$$\mathcal{F} = \{(\mathbf{x}_i)_{i \leq m} \to \mathbf{1}_{\langle \frac{1}{m} \sum_i x_i - \mathbb{E}(Y_1), v \rangle \geq r_K}, v \in V\}.$$

This result is an alternative to (Lugosi and Mendelson, 2018, Theorem 2) where the complexity is measured with VC-dimension instead of the Rademacher complexity. We show below that this difference yields to substantial improvements in some examples such as sparse multivariate mean estimation compared with the bounds in Lugosi and Mendelson (2018). The strength of this result is that it is uniform in $\mathcal{F}$ and gives an exponentially low failure probability, but its proof is quite simple. The proof of this result is given in Section 5.6.2.

Clearly, the fraction 3/4 of the block is arbitrary in Lemma 5.2. In fact, up to some modifications of the constants, the same result holds for any *fixed* fraction $\alpha < 1$.

5.2.3 Mean estimation

Let $\|\cdot\|$ denote a norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ denote its dual norm. Let $\mathcal{B}$ denote the unit ball for the norm $\|\cdot\|$ and $\mathcal{B}^*$ the one for the norm $\|\cdot\|_*$. Let $\mathcal{B}_0^*$ denote the set of extremal vectors of $\mathcal{B}^*$. Let $|||A||| = \sup_{u\in\mathcal{B}^*} \|Au\|_2$ where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^d$. Let

$$\hat{\mu}_K = \operatorname*{argmin}_{a\in\mathbb{R}^d} \max_{u\in\mathcal{B}_0^*} \mathrm{Med}(\langle \bar{X}_k - a, u\rangle).$$

Theorem 5.2. *There exists an universal constant C such that if $K \geq C(\mathrm{VC}(\mathcal{B}_0^*) \vee |\mathcal{O}|)$, then, with probability larger than $1 - \exp(-K/128)$,*

$$\|\hat{\mu}_K - \mu\| \leq 8 \left|\left|\left|\Sigma^{1/2}\right|\right|\right| \sqrt{\frac{K}{N}}.$$

In particular, for any $\delta \in [e^{-cN}, 1/2]$, there exists an estimator μ_δ such that

$$\|\mu - \mu_\delta\| \lesssim \left|\left|\left|\Sigma^{1/2}\right|\right|\right| \left(\sqrt{\frac{\mathrm{VC}(\mathcal{B}_0^*)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon}\right). \tag{5.3}$$

The 'outlier' term $\sqrt{\epsilon}$ can be shown to be optimal in important cases (see the remarks after (Cheng et al., 2019a, Theorem 1.3)). The deviation term ($\left|\left|\left|\Sigma^{1/2}\right|\right|\right| \sqrt{\log(1/\delta)/N}$) is the same as in the Borel-TIS inequality, $\left|\left|\left|\Sigma^{1/2}\right|\right|\right|$ being the weak variance term. It is optimal as shown in Lugosi and Mendelson (2018). The difference with Lugosi and Mendelson (2018) is the complexity term, which is here $\left|\left|\left|\Sigma^{1/2}\right|\right|\right| \sqrt{\mathrm{VC}(\mathcal{B}_0^*)/N}$. Neither Lugosi and Mendelson (2018) nor this work build estimators achieving in every cases the true subgaussian rate, where this complexity is $\mathbb{E}(\|G\|)/\sqrt{N}$, G being a centered Gaussian vector with the same covariance as Y. For now it is not known whether MOM estimators can or cannot achieve this rate in general, for all possible norms. However, as we will show, our rate match the true subgaussian rate in some special cases (so does the one from Lugosi and Mendelson (2018) for some other special cases)

Remark 5.2. *The inequality $\mathrm{VC}(\mathcal{B}_0^*) \leq d + 1$ gives a general bound on the complexity term.*

The complexity term in Lugosi and Mendelson (2018), which can also be found in (Lerasle, 2019, Chapter 4, Lemma 47) is $\mathbb{E}(\|\tilde{Y}\|)/N$ where $\tilde{Y} = \sum \epsilon_i(Y_i - \mu)$, ϵ_i being i.i.d. Rademacher variables. Here it is $\left|\left|\left|\Sigma^{1/2}\right|\right|\right| \sqrt{\mathrm{VC}(\mathcal{B}_0^*)/N}$. Which of them is the best depends on the situation. For instance, when one wishes to estimate with respect to $\|\cdot\|_2$, the Euclidean norm on $\mathbb{R}^d$, $\mathbb{E}(\|\tilde{Y}\|_2)/N \simeq \sqrt{\mathrm{Tr}(\Sigma)/N}$, while $\left|\left|\left|\Sigma^{1/2}\right|\right|\right| \sqrt{\mathrm{VC}(\mathcal{B}_0^*)/N} = \sqrt{\lambda_1 d/N}$, λ_1 being the largest eigenvalue of Σ, so the former is better. In this example, the bound in VC dimension loses the nice dependence in the covariance structure. On the other hand, suppose that we want to estimate μ with respect to the sup norm $\|a\|_\infty = \max\{a_1, ..., a_n\}$ and assume that $\Sigma = \mathrm{Id}$ for simplicity. Then $|||\mathrm{Id}||| = 1$ and $\mathrm{VC}(\mathcal{B}_0^*) \lesssim \log(d)$ so

$$\left|\left|\left|\Sigma^{1/2}\right|\right|\right| \sqrt{\mathrm{VC}(\mathcal{B}_0^*)/N} \simeq \sqrt{\log(d)/N}.$$

On the other hand, if we only have two moments on the coordinates of Y, then the best bound on the Rademacher complexity is $\mathbb{E}(\|\tilde{Y}\|_2)/N$ which is of order $\sqrt{d/N}$ in general (to see that, take for Y_1 a random vector whose coordinates are independent, equal to $\sqrt{dN}$ with probability $1/(dN)$ and 0 otherwise).

Remark 5.3. *The analysis of Section 5.6.2 and in particular Lemma 5.4', shows that the estimator $\hat{\mu}_K$ achieves the bound $\left|\left|\left|\Sigma^{1/2}\right|\right|\right|\sqrt{K/N}$ when $K \geq \mathcal{C} \vee |\mathcal{O}|$, where the complexity $\mathcal{C}$ is the minimum between the VC dimension $\mathrm{VC}(\mathcal{B}_0^*)$ and the Rademacher complexity $\mathbb{E}(\|\tilde{Y}\|)^2/(N\left|\left|\left|\Sigma^{1/2}\right|\right|\right|)$. Therefore both our bounds and the bound of Lugosi and Mendelson (2018) hold simultaneously and we can always keep the "best complexity term" among VC and Rademacher complexity. As the main novelty here is the introduction of the VC-dimension, we do not remind this fact in each application. The interested reader can have in mind that, in most examples, the same result holds and the estimators have risk bounds smaller than both complexities. Our aim is to show that VC type bounds are particularly efficient in structured scenarii, when Rademacher complexity fails to achieve optimal bounds.*

5.3 Sparse setting and other estimation tasks

This section shows that the methodology of Theorem 5.2 also applies to a great variety of estimation tasks. Let us start with the example of sparse mean estimation for the Euclidean norm.

5.3.1 Sparse mean estimation

For any $v_1, ..., v_d \in \mathbb{R}^n$, let $\mathcal{U}_s(v_1, ..., v_d) = \{\sum_i \lambda_i v_i \mid \lambda_i \in \mathbb{R} \ \& \ \sum_i \mathbf{1}_{\lambda_i \neq 0} \leq s\}$ denote the set of s-sparse vectors over the dictionary $\{v_1, ..., v_d\}$. We fix for this part the vectors $v_1, ..., v_d$ and we note $\mathcal{U}_s = \mathcal{U}_s(v_1, ..., v_d)$. We consider Setting 5.1 and assume furthermore that μ belongs to $\mathcal{U}_s$. We note $\mathcal{B}_2$ the unit ball for the canonical Euclidean norm in $\mathbb{R}^n$, and we propose the estimator

$$\hat{\mu}_K = \operatorname*{argmin}_{a \in \mathcal{U}_s} \max_{u \in \mathcal{U}_{2s} \cap \mathcal{B}_2} \mathrm{Med}\, \langle \bar{X}_k - a, u \rangle .$$

Theorem 5.3. *There exists an absolute constant C such that, if $K \geq C(s \log(d/s) \vee |\mathcal{O}|)$, then, with probability larger than $1 - \exp(-K/128)$,*

$$\|\hat{\mu}_K - \mu\|_2 \leq 8\sqrt{\frac{\lambda_1(\Sigma)K}{N}}.$$

Here, $\lambda_1(\Sigma)$ is the largest eigenvalue of Σ.

The conclusion of Theorem 5.3 can be written as follows. For any $\delta \in [e^{-cN}, 1/2]$, there exists an estimator μ_δ such that

$$\|\mu - \mu_\delta\|_2 \lesssim \lambda_1(\Sigma)\left(\sqrt{\frac{s \log(d/s)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon}\right).$$

We see that the complexity ($s \log(d/s)$) is once again decoupled from the deviation ($\log(1/\delta)$), which is not the case in works such as Hsu and Sabato (2016) where those two terms are multiplied together. The complexity term $s \log(d)/N$ is not optimal because it does not depend on the structure of Σ (see Section 5.4 for details). However, our complexity term is interesting for two main reasons:

- This is the first sparsity dependent bound that holds without higher moments conditions than the L_2 ones. By contrast, Lerasle (2019) or Lecué and Lerasle (2020) need to assume the existence $\log(d)$ subgaussian moments in order to make the sparsity appear, and offer no guarantees without that requirement.

- It comes close to the theoretic optimal when $\Sigma \simeq \lambda \mathrm{Id}$.

Remark 5.4. *This theorem can be obtain without resorting to VC-dimensions, simply by using the analysis of Lugosi and Mendelson (2019c) on the $\binom{d}{s}$ subspaces of $\mathcal{U}_s$ and a union bound. In other words, we can get similar rates with the standard median-of-means approach and simple manipulations. However, for other examples where the set considered is not a simple union of subspaces, this kind of trick are no longer possible.*

5.3.2 Regression

In this section, we consider the standard linear regression setting where data are couples $(Y_i, V_i)_i \in \mathbb{R}^d \times \mathbb{R}$ and we look for the best linear combination of the coordinates of Y_i to predict V_i, that is we look for β^* defined as follows: Given $\mathcal{S} \subset \mathbb{R}^d$ (in practice, we will only study $\mathcal{S} = \mathbb{R}^d$ or $\mathcal{S} = \mathcal{U}_s$),

$$\beta^* = \underset{\beta \in \mathcal{S}}{\operatorname{argmin}}\, l(\beta) = \underset{\beta \in \mathcal{S}}{\operatorname{argmin}}\, \mathbb{E}(V_1 - \langle \beta, Y_1 \rangle)^2.$$

As in the previous section, the observed dataset $(X_i, Z_i)_i \in \mathbb{R}^d \times \mathbb{R}$ is a corrupted version of the i.i.d. dataset $\{(Y_i, V_i)_i, i \in \{1, \dots, N\}\}$ in a possibly adversarial way. The assumptions made on good data $(Y_i, V_i)_i$ are gathered in the following setting: (see also Lerasle (2019) or Audibert and Catoni (2011)).

Setting 5.2. *There exists a (possibly random) set $\mathcal{O}$ such that, for any $i \in \mathcal{O}^c$, $(X_i, Z_i) = (Y_i, V_i)$, where (Y_i, V_i) are independent identically distributed observations in $\mathbb{R}^d \times \mathbb{R}$. Let $\xi_i = V_i - \langle \beta^*, Y_i \rangle$, we make four main assumptions:*

- *Y_1 has finite second moment and write its L^2-moments matrix $\Sigma = \mathbb{E}(Y_1 Y_1^T)$. Let also $B_\Sigma = \{x \in \mathcal{S} - \mathcal{S} | \langle x, \Sigma x \rangle \leq 1\}$ be the ellipsoid associated with this L_2 structure.*
- *Let $\sigma^2 := \sup_{u \in B_\Sigma} \mathbb{E}(\xi_1^2 \langle u, Y_1 \rangle^2)$ and assume that $\sigma^2 < \infty$.*
- *There exists an universal constant γ such that, for all $u \in \mathcal{S} - \mathcal{S}$, $\mathbb{E}(|\langle u, X \rangle|) \geq \gamma \sqrt{\mathbb{E}(|\langle u, X \rangle|^2)}$.*
- $\mathbb{E}(\xi_1 Y_1) = 0$

Condition 2 is implied by Assumptions 3.5 and 3.7 in Audibert and Catoni (2011), the same assumption is made in Lerasle (2019).

Condition 3 is called the "small ball hypothesis", it is described in details in Mendelson (2017a) or in Lecué and Mendelson (2016) for instance. It is implied by Condition 3.5 in Audibert and Catoni (2011), it is stated similarly in Lerasle (2019).

Condition 4 is always true in the non sparse-case. In the sparse case, it is true in a number of applications, for instance, the very important when the noise ξ and Y are independent.

The two last conditions may seem exotic, we refer to (Audibert and Catoni, 2011, Section 3) for detailed discussions and examples where these are satisfied. For the moment, we may emphase that they involve only first and second moment conditions on ξ_1 and $\langle u, Y_1 \rangle$

Our estimator is the following: Let $\hat{B}_\Sigma = \{u \in \mathcal{S} - \mathcal{S} \mid \mathcal{Q}^k_{1/4} \frac{1}{m} \sum_{i \in B_k} |\langle u, X_i \rangle|^2 \leq 1\}$, where $\mathcal{Q}^k_{1/4}$ is the first quartile over $k \leq K$: for any sequence $x_1, ..., x_n \in \mathbb{R}$, if we note $x_1^*, ..., x_n^*$ the corresponding increasingly ordered sequence, $\mathcal{Q}^k_{1/4} x_k = x^*_{\lfloor n/4 \rfloor}$. Then,

$$\hat{\beta} = \underset{a \in \mathcal{S}}{\operatorname{argmin}} \max_{u \in \hat{B}_\Sigma} \underset{k}{\operatorname{Med}} \sum_{i \in B_k} (Z_i - \langle a, X_i \rangle) \langle u, X_i \rangle .$$

This new estimator satisfies the following result.

Theorem 5.4. *There exists an absolute constant C such that the following holds. Let $\mathcal{S} = \mathbb{R}^d$ or $\mathcal{U}_s$ and let $N\gamma^2/64 \geq K \geq C(\mathrm{VC}(\mathcal{S}-\mathcal{S}) \vee |\mathcal{O}|)$. Then, with probability $\geq 1 - \exp(-K/128)$,*

$$\langle \hat{\beta} - \beta^*, \Sigma(\hat{\beta} - \beta^*) \rangle \leq 128\ \sigma^2 \frac{K}{N\gamma^4}.$$

For all β,

$$l(\beta) = l(\beta^*) + 2\mathbb{E}(\xi_1 \langle \beta - \beta^*, Y_1 \rangle) + (\beta - \beta^*)\Sigma(\beta - \beta^*) \leq l(\beta^*) + (\beta - \beta^*)\Sigma(\beta - \beta^*).$$

So, if $r = \sqrt{128}\ \sigma\sqrt{\frac{K}{N}}$, then

$$l(\hat{\beta}) - l(\beta^*) \leq r^2/\gamma^4.$$

The conclusion of Theorem 5.4 can be written as follows: for any $\delta \in [e^{-cN}, 1/2]$, there exists an estimator μ_δ such that

$$\langle \hat{\beta} - \beta^*, \Sigma(\hat{\beta} - \beta^*) \rangle \lesssim \sigma(\sqrt{\frac{\mathrm{VC}(\mathcal{S})}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon}).$$

Once again we notice the nice decoupling between complexity and deviation. This result is interesting for a several reasons, the main one being that this work is the first that gives a bound holding with exponential probability, that holds without assuming more than 2 moments on the design Y_1, even in the sparse setting. By comparison, Lecué and Lerasle (2020) or Lugosi and Mendelson (2017) for instance, assume that at least $\log(d)$ subgaussian moments exist to achieve this kind of rate and offers no guarantees without that requirement and are the best to the date.

5.3.3 Low rank matrix estimation

We now turn to the problem of matrix estimation, presented for instance in Zinodiny et al. (2017, 2018); Tsukuma (2010). We have observations in $\mathcal{M}_d(\mathbb{R})$ (square matrices of size d) and we try to recover their mean, assuming a kind low-ranked structure. The setting is the following: we have, as in setting 5.1, N (corrupted) observations $(X_i)_i \in \mathcal{M}_d(\mathbb{R})$ of original (Y_i) satisfying $\mathbb{E}(Y_i) = B$ and we try to recover the (non necessarily low rank) mean B. We will assume for simplicity that $\mathbb{E}((Y^{ij} - B^{ij})(Y^{kl} - B^{kl})) = \sigma^2\delta_{(i,j)=(k,l)}$. We try to estimate B with respect to the following norm :

$$||A||_r = \sup_{U \in \mathcal{M}_d^r(\mathbb{R}),\ \|U\|_F = 1} \langle U, A \rangle_F ,$$

where we recall that $\mathcal{M}_d^k(\mathbb{R})$ is the set of rank k, symmetric, d-dimensional matrices.We will try to show that this structure can not be recovered through the analysis based on Rademacher complexity: we give this example to illustrate the benefit of our approach.

$$\hat{B}_K = \operatorname*{argmin}_{M \in \mathcal{M}_d(\mathbb{R})} \sup_{U \in \mathcal{M}_d^{2r}(\mathbb{R}),\ \|U\|_F = 1} \operatorname*{Med}_k \frac{1}{m} \sum \langle U, \bar{X}_k - M \rangle_F ,$$

Theorem 5.5. *There exists an absolute constant C such that, if $K \geq C(kd \vee |\mathcal{O}|)$, then, with probability larger than $1 - \exp(-K/128)$,*

$$\left\| \hat{B}_K - B \right\|_r \leq 8\sqrt{\frac{\sigma K}{N}}.$$

The conclusion of Theorem 5.5 can be writen as follows. For any $\delta \in [e^{-cN}, 1/2]$, there exists an estimator $\hat{B}_\delta$ such that

$$\left\|\hat{B}_\delta - B\right\|_r \lesssim \sigma(\sqrt{\frac{kd}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\epsilon}).$$

While in X_i to be i.i.d. gaussian in Zinodiny et al. (2017), we only need second-order moments.

We want to show at this point that those results could not be obtained using standard analysis with Rademacher complexity (Lugosi and Mendelson (2018); Lerasle (2019)). Indeed this analysis would give a bound of order $\mathbb{E}(\max_{U\in\mathcal{M}_d^{2r}(\mathbb{R}),\ \|U\|_F=1} \langle \tilde{Y}, U\rangle)/N$ (with $\tilde{Y} = \sum \epsilon_i(Y_i - B)$, ϵ_i being i.i.d. Rademacher variables) instead of $\sigma\sqrt{\frac{kd}{N}}$, as mentionned in section 5.2.3. Let us show a case where those two quantities have different behaviours.

We take for instance $N = 1$ and, independent identically distributed $(Y^{kl})_{1\le k,l\le d}$ so that

$$Y^{kl} = \begin{cases} +\sigma d & \text{with probability } 1/(2d^2) \\ -\sigma d & \text{with probability } 1/(2d^2) \\ 0 & \text{with probability } 1 - 1/d^2, \end{cases}$$

If one of the Y^{kl} is non zero, then $\max_{U\in\mathcal{M}_d^{2r}(\mathbb{R}),\ \|U\|_F=1} \langle Y, U\rangle)/N \ge \sigma d$. Given that $\mathbb{P}(\forall(k,l), Y^{kl} = 0) = (1 - 1/d^2)^{d^2} < e^{-1}$, we get that

$$\mathbb{E}(\max_{U\in\mathcal{M}_d^{2r}(\mathbb{R}),\ \|U\|_F=1} \langle \tilde{Y}, U\rangle) \ge \sigma d(1 - e^{-1}).$$

In this case, the quantity we get from the Rademacher analysis scales as d whereas our bound scales as $\sqrt{kd}$! Moreover, use of union bounds (as in Section 5.3.1, Remark 4 for the sparse case) is not possible here because $\mathcal{M}_d^{2r}$ is not an union of linear subspaces.

5.3.4 Covariance estimation

This section studies the problem of robust covariance estimation. Consider Setting 5.1, and assume that μ is known, fixed to 0 without loss of generality. We want to estimate Σ. This problem has a number of applications: the bounds we present can for instance easily be transposed (with the Davis–Kahan theorem) to the problem of robust PCA. It has already been studied in Wei and Minsker (2017), or Hsu and Sabato (2016), but these estimators do not exhibit any decoupling between complexity and deviation. In Mendelson and Zhivotovskiy (2018), the authors propose a robust estimator for covariance using the MOM method, and get the optimal complexity-deviation decoupling. They also give interesting comments and insights about this estimation problem. However, they do not study the problem of low rank estimation that we present here.

For any matrix A, define its spectral norm by

$$|||A||| = \sup_x \frac{\|Ax\|_2}{\|x\|_2}.$$

Let $\mathrm{Sym}(d)$ denote the set of d dimensional symmetric positive matrices. Assume that

$$\sigma^2 = \sup_{u\in\mathcal{B}_2} \mathbb{E}\left(\langle u, (\Sigma - Y_1Y_1^T)u\rangle^2\right) < \infty.$$

This quantity is sometime refered to as *weak variance* of a random matrix Mendelson and Zhivotovskiy (2018). Our estimator is defined as follows

$$\hat{\Sigma} = \operatorname*{argmin}_{M\in\mathrm{Sym}_d} \sup_{\|u\|_2=1} \operatorname*{Med}_k \langle u, (\frac{1}{m}\sum X_iX_i^T - M)u\rangle\,.$$

It satisfies the following bound.

Theorem 5.6. *There exists an absolute constant C such that, if $K \geq C(d \vee |\mathcal{O}|)$, then, with probability larger than $1 - \exp(-K/128)$,*

$$\left|\left|\left|\hat{\Sigma}_K - \Sigma\right|\right|\right| \leq 8\ \sigma\sqrt{\frac{K}{N}}.$$

Corollary 5.3. *Assume that $R = \sup_u \frac{\sqrt{\mathbb{E}\langle u,Y\rangle)^4}}{\mathbb{E}\langle u,Y\rangle^2} < \infty$, then, for $K \geq C(d \vee |\mathcal{O}|)$*

$$\left|\left|\left|\hat{\Sigma}_K - \Sigma\right|\right|\right| \leq 8\ R|||\Sigma|||\sqrt{\frac{K}{N}}.$$

The "bounded kurtosis assumption" $R < \infty$ appears similarly in Hsu and Sabato (2016). In Hsu and Sabato (2016), the estimator achieves a bound of order $r(\Sigma)|||\Sigma|||\sqrt{K/N}$ where $r(\Sigma)$ is the effective rank of the covariance matrix: once again, the complexity $r(\Sigma)|||\Sigma|||$ is multiplied by the deviation term $K \propto \log(1/\delta)$ in this case, while here they are decoupled: the dimension does not multiply K in our bound. In Mendelson and Zhivotovskiy (2018), authors give a better rate (because they only need K to be larger than $r(\Sigma)$ instead of d for the bound to hold) but they use the $L_4 - L_2$ norm equivalence, which is an hypothesis we do not need here.

5.4 An algorithm to improve risk bounds

The different applications of Lemma 5.2 show that, in general, the complexity term derived from this result is not optimal. For example, for mean estimation in Euclidean norm, the complexity term reached by our estimator is proportional to $\sqrt{\lambda_1(\Sigma)d/N}$, where the best rate would be $\sqrt{\mathrm{Tr}(\Sigma)/N}$.

In this section, we provide an algorithm that leads to better and in some cases optimal complexity rates. The price to pay is that these new estimators require some knowledge on the covariance matrix Σ. We will consider the example of sparse mean estimation for the sake of clarity, but we argue that it also holds for sparse (and non-sparse) regression. We therefore use the setting 5.1 and assume furthermore that the mean μ belongs to $\mathcal{U}_s(v_1, ..., v_d)$, where the set of vectors $(v_1, ..., v_d) \in \mathbb{R}^d$ is fixed and known.

Let $\lambda_1, \lambda_2, ..., \lambda_d$ denote the eigenvalues of Σ in decreasing order, and let $e_1, ..., e_d$ denote a set of normalized corresponding eigenvectors. For any $1 \leq n < \lfloor \log_2(d) \rfloor := n_l$, let s_n denote the largest index such that $\lambda_{s_n} \geq \lambda_1/2^n$. In particular, $\lambda_1 \geq ... \geq \lambda_{s_1} \geq \lambda_1/2 > \lambda_{s_1+1}....$ By convention, let $s_0 = 0$. Finally, we note $E_n = \mathrm{Vect}\{e_{s_{n-1}+1}, e_{s_{n-1}+2}, ..., e_{s_n}\}$, with convention $E_{n_l} = \mathrm{Vect}(e_{s_{n_l-1}+1}, e_{s_{n_l-1}+2}, ..., e_d)$. If we know the matrix Σ, we can identify the eigenspaces E_n and thus compute the orthogonal projections of the data on these subspaces: $X_k^i := \mathbf{proj}_{E_i}(X_k)$, for $i \in \{1, ..., n_l\}$ and $k \in \{1, ..., K\}$.

In Section 5.3.1, we described a procedure that takes as input an integer $K \geq c_0(\tilde{s}\log(\tilde{d}/\tilde{s}) \vee |\mathcal{O}|)$ and a (possibly corrupted) dataset $Z_1, ...Z_N$ having common mean $\tilde{\mu}$ which is $\tilde{s}$-sparse relatively to a set of vectors $(u_1, u_2, ..., u_{\tilde{d}})$ and common covariance matrix $\tilde{\Sigma}$. The procedure returns $\hat{\mu}_K$ satisfying, with probability at least $1 - \exp(-c_1 K)$

$$\|\hat{\mu}_K - \mu\|_2 \leq 8\sqrt{\frac{\left|\left|\left|\tilde{\Sigma}\right|\right|\right| K}{N}}.$$

Let $\texttt{proc}(Z_1, ..., Z_N, K, u_1, u_2, ..., u_{\tilde{d}}, \tilde{s})$ denote the output of this procedure. The idea of the algorithm is to project on the subspaces E_i and apply this preliminary procedure on those subspaces. Let d_i the dimension of E_i. The algorithm is formally defined as follows:

input : $X_1, \dots, X_N$ and $K \geq |\mathcal{O}|$.
output : A robust subgaussian estimator $\hat{\mu}_\delta$.

$i \leftarrow 1$.
while $i \leq n_l$ **do**
  Compute $X_1^i, ..., X_n^i$.
  Compute $u_1^i, u_2^i, ..., u_d^i$ the orthogonal projections of $v_1, v_2, ...v_d$ onto E_i.
  $K_i \leftarrow K \times 2^{i-1}$.
  if $d_i < s\log(d/s)$ **then**
    $\mu_i \leftarrow \texttt{proc}(X_1^i, ..., X_n^i, K_i, e_{s_{i-1}+1}, ..., e_{s_i}, d_i)$.
  end
  else
    $\mu_i \leftarrow \texttt{proc}(X_1^i, ..., X_n^i, K_i, u_1^i, u_2^i, ..., u_d^i, s)$.
  end
  $i \leftarrow i + 1$.
end
$\hat{\mu}_K \leftarrow \sum_{j \leq i} \mu_i$.
Return $\hat{\mu}_K$.

Algorithm 12: Pseudo-code of the robust sub-gaussian estimator of μ

The algorithm produces an estimator $\hat{\mu}_K$ satisfying the following result.

Proposition 5.3. *Assume setup 5.1. There exists an absolute constant C such that, if $K \geq C[\sum_i 2^{-i}(d_i \wedge s\log(d/s)) \vee |\mathcal{O}|]$, the output $\hat{\mu}_K$ of Algorithm 12, satisfies, with probability $\geq 1 - 2\exp(-K/128)$,*

$$\|\hat{\mu}_K - \mu\|_2 \leq 8\log(d)\sqrt{\frac{\lambda_1 K}{N}}.$$

The complexity term in Proposition 5.3 is better than in Theorem 5.3, because $\sum_i 2^{-i}(d_i \wedge s\log(d/s)) \leq s\log(d/s)$. More importantly, this complexity term depends on the the covariance structure of the data through the d_i. In the case of sparse mean estimation, we can deduce a precise estimate of this complexity term: for any $\delta \in [e^{-cN}, 1/2]$, there exists an estimator $\hat{\mu}_\delta$ such that:

$$\|\hat{\mu}_\delta - \mu\|_2 \leq C\log(d)\left(\sqrt{\frac{\log(d/s)\sum_{i=1}^s \lambda_i}{N}} + \sqrt{\frac{\lambda_1 \log(1/\delta)}{N}} + \sqrt{\lambda_1 \epsilon}\right).$$

This estimate comes from the bounds $\sum_{i=1}^s \lambda_i \geq \sum_{i \leq j+1} \lambda_1 2^{-i-1} d_i \wedge s$, where j is such that $\sum_{i \leq j} d_i \leq s \leq \sum_{i \leq j+1} d_i$. We then write $\sum_i 2^{-i}(d_i \wedge s\log(d/s)) \leq \log(d/s)\sum_{i \leq j+1} 2^{-i} d_i \wedge s + 2^{-j} s\log(d/s) \leq 4\log(d/s)\sum_{i=1}^s \lambda_i$.

This result is proved in Section 5.6.5. We argue that the very same proof can be replicated for the regression problem, if $\mathbb{E}(\xi_1^2 Y_1 Y_1^T)$ has the same eigenspaces as $\mathbb{E}(Y_1 Y_1^T)$. This happens for instance when ξ is independent of Y. We end up with the bound of Theorem 5.4, holding whenever $K \geq C[\sum_i 2^{-i}(d_i \wedge \mathrm{VC}(\mathcal{S}) \vee |\mathcal{O}|]$ instead of $K \geq C[\mathrm{VC}(\mathcal{S} \vee |\mathcal{O}|]$

5.5 Conclusion: concurrent work and discussion

This work is not the first to deal with robust estimation: a lot of results and algorithms have been developed over the past few years for sparse estimation in presence of outliers (see for instance Li (2017)) but most of these works assume that non-corrupted data are Gaussian. For instance Chen et al. (2018) already deal with mean and covariance estimation using extensions of the Tukey-depth (and using VC-dimension), but their methods rely on informative data being Gaussian.

Robustness to heavy-tailed data has also been studied in various works, see Lugosi and Mendelson (2019a) for a survey of recent developments. We already mentioned two articles that this work tries to complete and improve: Lugosi and Mendelson (2018) for mean estimation under any norm and Lecué and Lerasle (2020) for sparse regression. Though the techniques involved are close, this work illustrates that using VC-dimension can drastically improve risk bounds in various applications, in particular in the sparse setting.

Concurrent work: After the initial submission of this manuscript, we became aware of two concurrent works Prasad et al. (2019) and Prasad et al. (2020). Authors use an approach based on the 1/2-cover of the unit sphere to deal with mean and sparse-mean estimation for the euclidean norm (Prasad et al. (2019)), and with covariance estimation for spectral norms (Prasad et al. (2020)). They get close-to-optimal bounds for those two problems, with a remaining extra-logarithmic term. They do not tackle mean estimation in any norms, regression or low-rank covariance estimation.

There are still many exciting open questions. The quantity that is crucial in all the studies is

$$\mathbb{E}\left(\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) - K\mathbb{E}(f(\mathbf{Y}_k))\right)$$

where f are boolean functions. In mean estimation for instance,

$$\mathbb{E}\left(\sup_{v\in V} \sum_{k=0}^{K} \mathbf{1}_{\langle \bar{Y}_k - \mu, v\rangle \geq r} - K\mathbb{E}(\mathbf{1}_{\langle \bar{Y}_k - \mu, v\rangle \geq r})\right)$$

is the important quantity. Bounding this quantity using the VC-dimension of V yields a bound independent of the covariance of Y. On the other hand, bounding that quantity by the Rademacher complexity of the Y_i (like in Lecué and Lerasle (2020), Lugosi and Mendelson (2018) or here in Part 5.6.2) stating that

$$\mathbb{E}(\sup_{v\in V} \mathbf{1}_{\langle \bar{Y}_k - \mu, v\rangle \geq r} - K\mathbb{E}(\mathbf{1}_{\langle \bar{Y}_k - \mu, v\rangle})) < K\frac{\sqrt{\mathbb{E}(\|\tilde{Y}\|)}}{r\sqrt{N}}$$

does not exploit the boundedness of the indicator function and necessitates unnecessary stronger assumptions on data. The ideal would be to conciliate both ideas, and to find a nice in-between that would take into account both the boundedness and the dependency in the covariance structure.

The last point we make is about computational issues. The estimators presented can not be implemented as is. Nevertheless, encouraging recent works have shown that "relaxed", computable estimators can be derived from this kind of work. For instance the pioneer work of Hopkins (2018), followed by Depersin and Lecué (2019) and Cherapanamjeri et al. (2019) for instance,

derived tractable estimators, in polynomial times, from the work of Lugosi and Mendelson (2019c). Even more recently, some new tractable estimators for regression and covariance estimation with heavy-tailed data have emerged in Cherapanamjeri et al. (2020a). We can hope for this work to be made tractable as well, which seems to be quite a challenge.

5.6 Main Proofs

5.6.1 A fact about VC-dimension

For any Euclidean space E, and any $C \in E$

Lemma 5.3. $\mathrm{VC}(\{x \in E \to \mathbf{1}_{\langle x,v\rangle \geq r}, v \in C\}) \leq \mathrm{VC}(C-C) \leq c_0 \mathrm{VC}(C)$ *where* c_0 *is universal constant.*

Proof. Assume that a set $x_1, ..., x_d \in E$ is shattered by $\mathcal{F} = \{x \in E \to \mathbf{1}_{\langle x,v\rangle \geq r}, v \in C\}$. Then, for any $I \subset \{1, 2, ..., d\}$, there is a vector v_1 so that $\langle v_1, x_i\rangle \geq r$ if and only if $i \in I$. There is a vector v_2 so that $\langle v_2, x_i\rangle < r$ if and only if $i \in I$. Then we have $\langle v_1 - v_2, x_i\rangle \geq 0$ if and only if $i \in I$, so $\{x \in E \to \mathbf{1}_{\langle x,v\rangle \geq 0}, v \in C - C\}$ shatters $x_1, ..., x_d$, and $\mathrm{VC}(\{x \in E \to \mathbf{1}_{\langle x,v\rangle \geq r}, v \in C\}) \leq \mathrm{VC}(C - C)$.

Now we see that $\mathrm{VC}(\{(x, y) \in E^2 \to \mathbf{1}_{\langle x,v\rangle + \langle y,w\rangle \geq 0} | (v, w) \in C \times C\}) \geq \mathrm{VC}(C - C)$ because if $x_1, ..., x_d \in E$ is shattered by $\{x \in E \to \mathbf{1}_{\langle x,v\rangle \geq 0}, v \in C - C\}$ then $((x_1, -x_1), ..., (x_d, -x_d)) \in E \times E$ is shattered by $\{(x, y) \in E^2 \to \mathbf{1}_{\langle x,v\rangle + \langle y,w\rangle \geq 0} | (v, w) \in C \times C\}$. Theorem 1.1 in van der Vaart and Wellner (2009) states that $\mathrm{VC}(C \times C) \leq c_0 \mathrm{VC}(C)$ for some constant c_0, and that concludes the proof. ∎

5.6.2 General methodology

We begin by proving the main lemma 5.2 of Part 5.2.3

Proof. We want to prove that, with probability $\geq 1 - \exp(-K/128)$,

$$\sup_f \sum_{k=0}^{K} f(\mathbf{X}_k) \leq K/4.$$

If $C \geq 16$, $K \geq 16|\mathcal{O}|$ and it is sufficient to show that $\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) \leq 3K/16$ by Remark 5.1. Now we write

$$\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) \leq \underbrace{\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) - \mathbb{E}\left(\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k)\right)}_{Deviation=D} + \underbrace{\mathbb{E}\left(\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k)\right)}_{Magnitude=M}.$$

By the bounded difference inequality (Boucheron et al., 2013, Theorem 6.2), with probability $\geq 1 - \exp(K/128)$, $D \leq K/16$.

For the magnitude term, we write

$$M \leq \mathbb{E}\left(\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) - K\mathbb{E}(f(\mathbf{Y}_k))\right) + \sup_f K\mathbb{E}(f(\mathbf{Y}_k)).$$

By hypothesis, $\sup_f K\mathbb{E}(f(\mathbf{Y}_k)) \leq K/16$. Then, we just have to use a classical result of Vapnik-Chervonenkis theory, either in the version of (Vershynin, 2018, Theorem 8.3.23), or of (van Handel, 2016, Corollary 7.18). There exists a universal constant C' such that

$$\mathbb{E}\left(\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) - K\mathbb{E}(f(\mathbf{Y}_k))\right) \leq C' K \sqrt{\frac{\mathrm{VC}(\mathcal{F})}{K}}.$$

Hence, if $K \geq 256\ C'^2\,\mathrm{VC}(\mathcal{F})$,

$$\mathbb{E}\left(\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) - K\mathbb{E}(f(\mathbf{Y}_k))\right) \leq \frac{K}{16}.$$

Putting everything together, we have the following. If $C \geq 256\ C'^2$, with probability $\geq 1 - \exp(K/128)$, $\sup_f \sum_{k=0}^{K} f(\mathbf{Y}_k) \leq K/16 + K/16 + K/16$. Therefore, by Remark 5.1, for all $f \in F$

$$\sum_{k=0}^{K} f(\mathbf{X}_k) \leq K/4.$$

■

We state a technical lemma that appears in most proofs. Let g be any measurable function $\mathbb{R}^d \to E$ so that $\mathbb{E}(g(Y_1))$ exists. We take

$$\hat{a} = \operatorname*{argmin}_{a \in U} \max_{v \in V} \mathrm{Med}\,\langle \frac{1}{m} \sum_{i \in B_k} g(X_i) - a, v\rangle$$

where U, V are any sets of E. We have:

Lemma 5.4. *If $K \geq C(\mathrm{VC}(V) \vee |\mathcal{O}|)$ and if $\mathbb{E}(g(Y_1)) \in U$, then, with probability $\geq 1 - \exp(-K/126)$,*

$$\max_{v \in V} \langle \mathbb{E}(g(Y_1)) - \hat{a}, v\rangle \leq 8 \sup_{u \in V} \mathbb{E}\left(\langle g(Y_1) - \mathbb{E}(g(Y_1)), u\rangle^2\right)^{1/2} \sqrt{\frac{K}{N}}$$

where C is a universal constant

$\sup_{u \in V} \mathbb{E}\left(\langle g(Y_1) - \mathbb{E}(g(Y_1)), u\rangle^2\right)^{1/2}$ is the "weak variance" of the problem.

Proof. Let $K \geq C(\mathrm{VC}(\mathcal{F}) \vee |\mathcal{O}|)$ with C the universal constant from Lemma 5.2, let $\bar{g} = \mathbb{E}(g(Y_1))$ and let

$$r_K = 4 \sup_{u \in V} \mathbb{E}\left(\langle g(Y_1) - \bar{g}, u\rangle^2\right)^{1/2} \sqrt{\frac{K}{N}}.$$

Let $\mathcal{F} = \{(\mathbf{x}_i)_{i \leq m} \to \mathbf{1}_{\langle \frac{1}{m}\sum_i g(x_i) - \mathbb{E}(g(Y_1)), v\rangle \geq r_K},\ v \in V\}$. The function $f \in \mathcal{F}$ are compositions of the function $\mathbf{x} \to \frac{1}{m}\sum_i g(x_i) - \mathbb{E}(g(Y_1))$ and of the functions $x \to \mathbf{1}_{\langle x, v\rangle \geq r_K}$ for $v \in V$. The VC-dimension of the set of these compositions is smaller than the VC-dimension of the set of indicator functions indexed by V, as recalled in the basic fact 2 at the beginning of Section 5.2.1. We just use fact 3 to remove the r_K and we get $\mathrm{VC}(\mathcal{F}) \leq c_0\,\mathrm{VC}(V)$ for some constant c_0.

By Markov's inequality, for any $v \in V$,

$$\mathbb{P}(|\,\langle \frac{1}{m} \sum_{i \in B_1} g(Y_i) - \bar{g}, v\rangle\,| \geq r_K) \leq \frac{\mathbb{E}\left(\sum_{i \in B_1} \langle g(Y_i) - \bar{g}, u\rangle^2\right)}{m^2 r_K^2} \leq \frac{1}{16}.$$

By Lemma 5.2, applied with $\mathcal{F}$, the following event $\mathcal{E}$ has probability $\mathbb{P}(\mathcal{E}) \geq 1 - \exp(-K/128)$.

$$\sup_{v\in V} \text{Med}\,|\,\langle \frac{1}{m}\sum_{i\in B_k} g(X_i) - \bar{g}, v\rangle\,| \leq r_K$$

For any $a \in U$ if there exists $v^* \in V$ such that $\langle \bar{g} - a, v^*\rangle > 2r_k$, then, on $\mathcal{E}$

$$\text{Med}\,\langle \frac{1}{m}\sum_{i\in B_k} g(X_i) - a, v^*\rangle = \langle \bar{g} - a, v^*\rangle + \text{Med}\,\langle \frac{1}{m}\sum_{i\in B_k} g(X_i) - \bar{g}, v^*\rangle$$
$$> r_K \geq \max_{v\in V} \text{Med}\,\langle \frac{1}{m}\sum_{i\in B_k} g(X_i) - \bar{g}, v\rangle\,.$$

Therefore $a \neq \hat{a}$. As this holds for any $a \in U$ such that $\sup_{v\in V}\langle \bar{g} - a, v\rangle > 2r_k$, it follows that, on $\mathcal{E}$,

$$\sup_{v\in V}\langle \bar{g} - \hat{a}, v\rangle \leq 2\ r_K.$$

■

We can give a somewhat improved version of that lemma: let us note,

$$\mathcal{R}(g, V) = \frac{1}{\sqrt{N}}\mathbb{E}(\sup_{v\in V}\langle \sum_i \epsilon_i g(Y_i), v\rangle)\ ,\ \sigma^2 = \sup_{u\in V}\mathbb{E}\left(\langle g(Y_1) - \mathbb{E}(g(Y_1)), u\rangle^2\right)$$

$\mathcal{R}$ is the Rademacher complexity associated to a given problem. The following lemma shows that we can take the best term between the one given by a rescaled Rademacher complexity and the one given by VC-dimension.

Lemma 5.5. *general If $K \gtrsim C((\text{VC}(V) \wedge (\mathcal{R}(g, V)/\sigma)^2) \vee |\mathcal{O}|)$ and if $\mathbb{E}(g(Y_1)) \in U$, then, with probability $\geq 1 - \exp(-K/126)$,*

$$\max_{v\in V}\langle \mathbb{E}(g(Y_1)) - \hat{a}, v\rangle \leq 16\sigma\sqrt{\frac{K}{N}}$$

where C is a universal constant

Proof. We know that this holds when $K \geq C(\text{VC}(V) \vee |\mathcal{O}|)$.

Now if $K \geq C(\mathcal{R}(g, V)/\sigma)^2 \vee |\mathcal{O}|$, we only need to prove that, for $r_K = 8\sigma\sqrt{K/N}$

$$\sup_{v\in V}\sum_k \mathbf{1}_{\langle \frac{1}{m}\sum_{i\in B_k} g(X_i) - \mathbb{E}(g(Y_1)), v\rangle \geq r_K} \leq K/2$$

and then we follow the path of the previous proof.

We do this in the classic way, that can be found, for instance in Depersin and Lecué (2019) or the supplementary material of M. Lerasle and Lecué (2017)

As $K \geq 4|\mathcal{O}|$, we only need to show that

$$\sup_{v\in V}\sum_k \mathbf{1}_{\langle \frac{1}{m}\sum_{i\in B_k} g(Y_i) - \mathbb{E}(g(Y_1)), v\rangle \geq r_K} \leq K/4$$

We define $\phi(t) = 0$ if $t \leq 1/2$, $\phi(t) = 2(t - 1/2)$ if $1/2 \leq t \leq 1$ and $\phi(t) = 1$ if $t \geq 1$. We have $I(t \geq 1) \leq \phi(t) \leq I(t \geq 1/2)$ for all $t \in \mathbb{R}$ and so for $v \in V$

$$\begin{aligned}
&\sum_k I(|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r_K) \\
&\leq \sum_k I(|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r_K) - \mathbb{P}[|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r_K/2] \\
&+ \mathbb{P}[|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r_K/2] \\
&\leq \sum_k \phi\left(\frac{|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle|}{r_K}\right) - \mathbb{E}\phi\left(\frac{|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle|}{r_K}\right) \\
&+ \mathbb{P}[|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r_K/2]
\end{aligned}$$

For all $v \in V$, we have

$$\mathbb{P}[|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r/2] \leq \frac{\mathbb{E}\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle^2}{(r_K/2)^2} \leq \frac{1}{16}$$

Next, using the bounded difference inequality (Theorem 6.2 in Boucheron et al. (2013)), the symmetrization argument and the contraction principle (Chapter 4 in Ledoux and Talagrand (2011)) – we refer to the supplementary material of M. Lerasle and Lecué (2017) for more details – with probability at least $1 - \exp(-K/128)$,

$$\begin{aligned}
&\sup_{v \in V} \left(\sum_k \phi\left(\frac{|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle|}{r_K}\right) - \mathbb{E}\phi\left(\frac{|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle|}{r_K}\right)\right) \\
&\leq \mathbb{E} \sup_{v \in V} \left(\sum_k \phi\left(\frac{|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle|}{r_K}\right) - \mathbb{E}\phi\left(\frac{|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle|}{r_K}\right)\right) + \frac{K}{16} \\
&\leq \frac{4K}{N r_K} \mathbb{E} \sup_{v \in V} \langle v, \sum_{i \in \cup_k B_k} \epsilon_i (g(Y_i) - \bar{g})\rangle + \frac{K}{16} \\
&= \frac{\sqrt{K}}{2\sigma} \mathbb{E} \langle \frac{1}{\sqrt{N}} \sum_{i \in \cup_k B_k} \epsilon_i (g(Y_i) - \bar{g}), v\rangle + \frac{K}{16} \leq \frac{K}{8}
\end{aligned}$$

when $\sqrt{K} \geq 8\mathcal{R}(g, V)/\sigma$ or $K \geq 64(\mathcal{R}(g, V)/\sigma)^2$

As a consequence, when $K \geq 64(\mathcal{R}(g, V)/\sigma)^2$, with probability at least $1 - \exp(-K/126)$, for all $v \in V$,

$$\sum_{k \in [K]} I(|\langle \frac{1}{m} \sum_{i \in B_k} g(Y_i) - \bar{g}, v\rangle| > r_K) \leq \frac{K}{8} + \frac{K}{16} \leq \frac{K}{4}.$$

■

5.6.3 Proof of Theorem 5.2, 5.3, 5.5, 5.6

This proofs are very similar: we just apply lemma 5.4 with the right g, U and V. We begin with Theorem 5.2 for estimating the mean with respect to a general norm.

Proof of Theorem 5.2. We just use lemma 5.4 with $g : x \to x$, $U = \mathbb{R}^d$ and $V = \mathcal{B}_0^*$. We have

$$\sup_{u \in V} \mathbb{E}\left(\langle g(Y_1) - \mathbb{E}(g(Y_1)), u\rangle^2\right) = \sup_{u \in \mathcal{B}_0^*} \|\Sigma u\|_2 = |||\Sigma|||$$

and, for any $a \in \mathbb{R}^d$

$$\sup_{v \in \mathcal{B}_0^*} \langle \mathbb{E}(Y_1) - a, v\rangle = \|\mu - a\|$$

so by Lemma 5.4 , we get that if $K \geq C(\mathrm{VC}(V) \vee |\mathcal{O}|)$, then, with probability $\geq 1 - \exp(-K/126)$

$$\|\hat{\mu} - \mu\| \leq 8|||\Sigma|||\sqrt{\frac{K}{N}}$$

■

We continue with the proof of Theorem 5.3 for estimating sparse means.

Proof of Theorem 5.3. We just use lemma 5.4, this time with $g : x \to x$, $U = \mathcal{U}_s$ and $V = \mathcal{U}_{2s} \cap \mathcal{B}_2$
We have

$$\sup_{u \in \mathcal{U}_{2s} \cap \mathcal{B}_2} \mathbb{E}\left(\langle g(Y_1) - \mathbb{E}(g(Y_1)), u\rangle^2\right) = \sup_{u \in \mathcal{U}_{2s} \cap \mathcal{B}_2} \left\|\Sigma^{1/2} u\right\|_2^2 = \lambda_1(\Sigma)$$

and, for any $a \in \mathcal{U}_s$ (so a fortiori for $\hat{\mu} \in \mathcal{U}_s$)

$$\sup_{v \in \mathcal{U}_{2s} \cap \mathcal{B}_2} \langle \mathbb{E}(Y_1) - a, v\rangle = \|\mu - a\|_2$$

because we assumed that $\mu \in \mathcal{U}_s$. So by Lemma 5.4, as $\mu \in \mathcal{U}_s$, we get that if $K \geq C(\mathrm{VC}(\mathcal{U}_{2s}) \vee |\mathcal{O}|)$, then, with probability $\geq 1 - \exp(-K/126)$

$$\|\hat{\mu} - \mu\|_2 \leq 8\lambda_1(\Sigma)\sqrt{\frac{K}{N}}$$

We recalled in part 5.2.1 that $\mathrm{VC}(\mathcal{U}_{2s}) \leq 2s\log(d/s)$, which concludes the proof.

■

Proof of Theorem 5.5. Let $V = \{U \in \mathcal{M}_d^{2r}(\mathbb{R}),\ \|U\|_F = 1\}$. This is just Theorem 5.2 , because we recalled in part 5.2.1 (Proposition 5.1) that $\mathrm{VC}(V) \leq c_0 kd$, for some universal constant c_0, which concludes the proof. ■

We move to the proof of Theorem 5.6, for estimating covariance with respect to the canonical euclidean operator norm.

Proof of Theorem 5.6. This time, we take $g : x \to xx^T$, $U = \mathrm{Sym}(d)$, and $V = \{uu^T | u \in \mathcal{B}_2(\mathbb{R}^d)\}$. We notice that $\mathbb{E}(g(Y_1)) = \Sigma$

We have

$$\sup_{M \in V} \mathbb{E}\left(\langle g(Y_1) - \mathbb{E}(g(Y_1)), M\rangle^2\right) = \sigma^2$$

by definition of σ^2, and for any $A \in \mathrm{Sym}(d)$ (so a fortiori for $\hat{\Sigma} \in \mathrm{Sym}(d)$)

$$\sup_{M \in V} \langle \Sigma - A, M\rangle = |||\Sigma - A|||$$

So by Lemma 5.4, as $\Sigma \in \mathrm{Sym}(d)$, we get that if $K \geq C(\mathrm{VC}(V) \vee |\mathcal{O}|)$, then, with probability $\geq 1 - \exp(-K/126)$

$$\left|\left|\left|\hat{\Sigma} - \Sigma\right|\right|\right| \leq 8\beta\sqrt{\frac{K}{N}}$$

We recalled in part 5.2.1 (Proposition 5.1) that $\mathrm{VC}(V) \leq c_0 d$, for some universal constant c_0, which concludes the proof. ∎

Appendix

5.6.4 Proof of Theorem 5.4

This proof is a bit different from the rest because we will have to control two different events.

Proof. Let $\mathcal{F} = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{(d+1)\times m} \to \mathbf{1}_{\langle u, \sum_i (y_i - \langle \beta^*, x_i\rangle)x_i\rangle^2 \geq m^2 r^2}, u \in B_\Sigma\}$. This is not a set of indicators of half-spaces, but $\mathcal{F}$ is the composition of $g : (\mathbf{x}, \mathbf{y}) \in R^{(d+1)\times m} \to (u \to \langle u, \sum_i (y_i - \langle \beta^*, x_i\rangle)x_i\rangle^2 - m^2 r^2) \in \mathbb{R}_2^d[X]$ and of $\{P \in \mathbb{R}_2^d[X] \to 1_{P(u)\geq 0}, u \in \mathcal{S} - \mathcal{S}\}$. By Lemma 5.3 there exists an absolute constant c such that $\mathrm{VC}(\mathcal{F}) \leq \mathrm{VC}(\mathcal{S} - \mathcal{S})$.

Let $\mathcal{G} = \{(\mathbf{x}_i) \to \mathbf{1}_{\sum \langle x_i, u\rangle^2 \geq \tilde{r}}, u \in \mathcal{S} - \mathcal{S}\}$. The same way, by Lemma 5.3, there exists an absolute constant c such that $\mathrm{VC}(\mathcal{G}) \leq c\,\mathrm{VC}(\mathcal{S})$. Assume that $K \geq C(\mathrm{VC}(\mathcal{F}) \vee \mathrm{VC}(\mathcal{G}) \vee |\mathcal{O}|)$ where C is the universal constant introduced in Lemma 5.2.

Multiplier process: Let

$$r = 4\sqrt{\frac{\sup_{u\in B_\Sigma} \mathbb{E}(\xi_1^2 \langle u, Y_1\rangle^2)K}{N}}.$$

For all $u \in B_\Sigma$,

$$\mathbb{P}\left(\frac{1}{m}\left|\sum_{i\in B_1} (V_i - \langle \beta^*, Y_i\rangle)\langle u, Y_i\rangle\right| \geq r\right) \leq \frac{\mathbb{E}(\xi_1^2 \langle u, Y_1\rangle^2)}{mr^2} \leq \frac{1}{16}.$$

By Lemma 5.2 applied with $\mathcal{F}$, it follows that the following event $\mathcal{E}$ has probability $\geq 1 - \exp(-K/128)$: for all $u \in B_\Sigma$, there exist more than $3/4K$ blocks k where

$$\left|\sum_{i\in B_k} (Z_i - \langle a, X_i\rangle)\langle u, X_i\rangle\right| \leq mr.$$

Quadratic process: From Chebyshev's inequality, for any $u \in \mathcal{S} - \mathcal{S}$,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i\in B_1} |\langle u, Y_i\rangle| \leq \mathbb{E}|\langle u, Y_1\rangle| - 4\sqrt{\frac{\mathbb{E}\langle u, Y_1\rangle^2}{m}}\right) \leq \frac{1}{16}.$$

So, when $K \leq \gamma^2 N/64$, by the small ball hypothesis,

$$\mathbb{P}\left(\frac{1}{m}\sum_{i\in B_1} |\langle u, Y_i\rangle| \leq \gamma/2\sqrt{\mathbb{E}\langle u, Y_1\rangle^2}\right) \leq \frac{1}{16}.$$

As $\frac{1}{m}\sum_{i\in B_k} |\langle u, X_i\rangle| \leq \sqrt{\frac{1}{m}\sum_{i\in B_k} |\langle u, X_i\rangle|^2}$, by Lemma 5.2 applied with $\mathcal{G}$ and $\tilde{r} = m\gamma^2/4\mathbb{E}\langle u, Y_1\rangle^2$, the following event $\mathcal{A}$ has probability probability $\geq 1 - \exp(-K/128)$: for all $u \in \mathcal{S} - \mathcal{S}$, there exists more than $3/4K$ blocks k where

$$\frac{1}{m}\sum_{i\in B_k} |\langle u, X_i\rangle|^2 \geq \gamma^2/4\,\langle u, \Sigma u\rangle\,.$$

So we have $\mathcal{Q}^k_{1/4}\frac{1}{m}\sum_{i\in B_k} |\langle u, X_i\rangle|^2 \geq \gamma^2/4 \times \langle u, \Sigma u\rangle$.

Conclusion of the proof. The event $\mathcal{E}\cap\mathcal{A}$ has probability at least $1-2\exp(-K/128)$. On $\mathcal{A}$, if $u\in\hat{B}_\Sigma$, then $\langle u,\Sigma u\rangle \leq 4/\gamma^2$, so, on $\mathcal{A}\cap\mathcal{E}$,

$$\max_{u\in\hat{B}_\Sigma} \operatorname*{Med}_k \sum_{i\in B_k} (Z_i - \langle\beta^*, X_i\rangle)\langle u, X_i\rangle \leq 2/\gamma \max_{u\in B_\Sigma} \operatorname*{Med}_k \sum_{i\in B_k} (Z_i - \langle\beta^*, X_i\rangle)\langle u, X_i\rangle \leq 2mr/\gamma.$$

For any $\beta\in\mathcal{S}$ such that $\Sigma(\beta-\beta^*)\neq 0$, let

$$u^* = \frac{\beta-\beta^*}{\sqrt{\mathcal{Q}^k_{1/4}\frac{1}{m}\sum_{i\in B_k}|\langle\beta-\beta^*, X_i\rangle|^2}}.$$

By construction $u^*\in\hat{B}_\Sigma$, so for 3/4 of the blocks, on $\mathcal{E}\cap\mathcal{A}$,

$$\Big|\sum_{i\in B_k} (Z_i - \langle\beta^*, X_i\rangle)\langle u^*, X_i\rangle\Big| \leq 2mr/\gamma.$$

On the other hand, by definition, for 3/4 of the blocks,

$$\frac{1}{m}\sum_{i\in B_k}|\langle\beta-\beta^*, X_i\rangle|^2 \geq \mathcal{Q}^{\tilde{k}}_{1/4}\frac{1}{m}\sum_{i\in B_{\tilde{k}}}|\langle\beta-\beta^*, X_i\rangle|^2.$$

Therefore, for at least half the blocks, both inequalities hold, so, on $\mathcal{E}\cap\mathcal{A}$,

$$\begin{aligned}\sum_{i\in B_k} (Z_i - \langle\beta-\beta^*+\beta^*, X_i\rangle)\langle u^*, X_i\rangle &\geq -2mr/\gamma + m\sqrt{\mathcal{Q}^{\tilde{k}}_{1/4}\frac{1}{m}\sum_{i\in B_{\tilde{k}}}|\langle\beta-\beta^*, X_i\rangle|^2}\\ &\geq -2mr/\gamma + m\gamma/2\sqrt{(\beta-\beta^*)\Sigma(\beta-\beta^*)}.\end{aligned}$$

It follows that, on $\mathcal{E}\cap\mathcal{A}$, if $\sqrt{(\beta-\beta^*)\Sigma(\beta-\beta^*)}\geq 8r/\gamma^2$, then $\sum_{i\in B_k}(Z_i-\langle\beta, X_i\rangle) \geq -2mr/\gamma + 4mr/\gamma \geq \sum_{i\in B_k}(Z_i-\langle\beta^*, X_i\rangle)$ and β can not be the chosen estimator. This concludes the proof. ∎

5.6.5 Proof of Proposition 5.3

Proof. We study separately what happens on each subspace E_i. The dimension of E_i is d_i and the orthogonal projection μ_i of μ is s-sparse on the set of vectors $u^i_1, u^i_2, ..., u^i_d$. μ_i is also generated by $e_{s_{i-1}+1}, ..., e_{s_i}$ which is a base of E_i. We choose which representation of μ_i leads to the best bound: if $d_i \geq s\log(d/s)$, we choose the first, else we choose the second. The preliminary bound holds if K_i is larger than either d_i or $s\log(d/s)$. Let $\tilde{\mu}_i$ denote our estimation on E_i:

$$\begin{aligned}\tilde{\mu}_i = &\texttt{proc}(X^i_1, ..., X^i_n, K_i, e_{s_{i-1}+1}, ..., e_{s_i}, d_i)\mathbf{1}_{d_i < s\log(d/s)}\\ &+\texttt{proc}(X^i_1, ..., X^i_n, K_i, u^i_1, u^i_2, ..., u^i_d, s)\mathbf{1}_{d_i\geq s\log(d/s)}.\end{aligned}$$

If $K_i \geq (Cd_i \wedge s\log(d/s)) \vee |\mathcal{O}|$, on an event $\mathcal{E}_i$ of probability $\geq 1-\exp(-K_i/128)$,

$$\|\tilde{\mu}_i - \mu_i\|_2 \leq 8\sqrt{\frac{\left\|\tilde{\Sigma}_i\right\| K_i}{N}} = \sqrt{\frac{\lambda_1 K}{N}}.$$

Let $\mathcal{E} = \cap \mathcal{E}_i$, so $\mathbb{P}(\mathcal{E}) \geq 1 - \sum \exp(-2^i K/128) \geq 1 - 2\exp(-K/128)$ if both $K \geq 128$ and $K \geq 2^{-i} C d_i \wedge s \log(d/s)$ for all i. As the subspaces E_i are orthogonal to each other (as eigenspaces of a symmetric matrix), by Pythagoras theorem,

$$\|\tilde{\mu} - \mu\|_2^2 = \sum_i \|\tilde{\mu}_i - \mu_i\|_2^2 \leq \log(d) \sqrt{\frac{\lambda_1 K}{N}}.$$

■

CHAPTER 6

On the robustness to adversarial corruption and to heavy-tailed data of the Stahel-Donoho median of means

Contents

6.1 Introduction

As mentioned in the general introduction, robust estimation of a mean vector has witnessed an important renewal during the last decade. The aim here is to construct an estimator achieving statistical bounds with the same confidence as if all the data were i.i.d. Gaussian even though the data at hand are only assumed to have a second moment. For the mean estimation problem in $\mathbb{R}^d$, most of the results have been given w.r.t. the Euclidean ℓ_2^d distance. There is however no statistical justification for this choice but that the ℓ_2^d metric is simply the most natural Hilbert metric in $\mathbb{R}^d$ and so it seems natural to use it as a way to measure the statistical performance of an estimator of a d-dimensional vector. The resulting confidence sets have therefore the form $\hat{\mu} + r^*_{N,\delta} B_2^d$ where $\hat{\mu}$ is an estimator, $B_2^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is the unit Euclidean ball and

$r^*_{N,\delta}$ is the rate of convergence w.r.t. ℓ_2^d achieved by $\hat{\mu}$ with confidence $1-\delta$. When estimating w.r.t. the ℓ_2^d metric, confidence sets are therefore ℓ_2^d-balls. One may wonder if these confidence sets are the best from a statistical point of view, for instance, the one with smallest volume for a fixed confidence $1-\delta$. To answer this type of question, we usually go back to the ideal i.i.d. Gaussian case, and use results obtained in that framework as benchmark results. We may also consider this model to design optimal benchmark confidence sets, that could be used to define more appealing estimation metric of a mean vector in $\mathbb{R}^d$.

Let us now see what are the "best" (in some sense given later) confidence sets in the i.i.d. Gaussian case: let $X_1, \dots, X_N$ be i.i.d. distributed like $\mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^d$ is the mean and Σ is a symmetric definite positive matrix (we assume here that Σ is invertible). The MLE is the empirical mean $\bar{X}_N$ and $\sqrt{N}(\bar{X}_N - \mu) \sim \mathcal{N}(0, \Sigma)$. The latter result holds asymptotically if the data are only assumed to be in L_2 thanks to the CLT. The key observation here is that Σ is the inverse of the Fisher information in this model and thus there are no regular asymptotically normal M-estimator that can estimate the mean with an asymptotic covariance matrix better than Σ. Moreover, level sets of the standard Gaussian density function are Euclidean B_2^d balls centered at zero. As a consequence, the best confidence sets for μ with confidence $1-\delta$ are ellipsoids $\Sigma^{1/2}B_2^d$ with radius given by the quantile of order $1-\delta$ of a chi-square variable with parameter d centered at the estimator. This type of confidence region are equivalently written as estimation results of μ with respect to the norm $x \in \mathbb{R}^d \to \left\|\Sigma^{-1/2}x\right\|_2$. It follows that the best metric – that is the one leading to minimal volume confidence sets for a given confidence in the benchmark i.i.d. Gaussian case – is the norm $\left\|\Sigma^{-1/2}\cdot\right\|_2$ whose unit ball is the ellipsoid $\Sigma^{1/2}B_2^d$.

Regarding our robust mean estimation problem, the two next natural questions are the following: is it possible to construct robust mean estimators w.r.t. the $\left\|\Sigma^{-1/2}\cdot\right\|_2$ metric? and what is the best convergence rate one can hope for? In the literature, see Lugosi and Mendelson (2019b); Depersin and Lecué (2020), one may find estimators which can estimate in a robust way a mean vector w.r.t. any metric of the type $u \in \mathbb{R}^d \to \|u\|_S = \sup_{v\in S}\langle v, u\rangle$ where $S \subset \mathbb{R}^d$. In particular, for $S = \Sigma^{-1/2}B_2^d$, this metric coincides with the one we want to use, i.e. $\left\|\Sigma^{-1/2}\cdot\right\|_2$. It has also been proved that the optimal deviation minimax rate (the one obtained in the benchmark i.i.d. Gaussian case) is for the mean estimation problem with respect to $\|\cdot\|_S$ given by (see Depersin and Lecué (2020))

$$\sqrt{\frac{\ell^*(\Sigma^{1/2}S)}{N}} + \sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\frac{\log(1/\delta)}{N}}. \tag{6.1}$$

For instance, for $S = B_2^d$ that is for $\|\cdot\|_S = \|\cdot\|_2$, the later rate is the classical $\sqrt{\mathrm{Tr}(\Sigma)/N} + \sqrt{\|\Sigma\|_{op}\log(1/\delta)/N}$ rate. The case that is interesting to us is when $\|\cdot\|_S = \left\|\Sigma^{-1/2}\cdot\right\|_2$, that is for $S = \Sigma^{-1/2}B_2^d$. In that case, the subgaussian rate is

$$\sqrt{\frac{d}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}. \tag{6.2}$$

This is the rate we will try to reach from an adversarial corrupted and heavy-tailed dataset. We will also have to take into account the price for corruption. There are indeed known information theoretic lower bounds showing that there are no statistics that can do better than $(|\mathcal{O}|/N)^\alpha$ where $\alpha \in [1/2, 1]$ is some exponent depending on properties of the 'good' inliers data. For instance, $\alpha = 1$ for Gaussian variables and $\alpha = 1/2$ for some L_2 variables. However, we will see that the best possible cost $|\mathcal{O}|/N$ (i.e. for $\alpha = 1$) can be achieved even for variables which do

not have a first moment as long as the cdfs of all one-dimensional projections of the centered and normalized data are regular enough.

Unfortunately, all estimators known to achieve the subgaussian rate in (6.2) (the Le Cam test estimator in Lugosi and Mendelson (2019b), the minmax MOM estimator with loss function $\ell(x,u) = \left\|\Sigma^{-1/2}(u-v)\right\|$ from Lerasle et al. (2019) or the Fenchel-Legendre estimators from Depersin and Lecué (2020)) are using the set S in their construction. This is something we cannot do here because $S = \Sigma^{-1/2}B_2^d$ depends on Σ which is unknown in general. One therefore has to consider other types of estimators than the ones cited above. In this work, we will do it thanks to a notion of depth/outlyingness introduced at the beginning of the 80's which, unlike the last cited estimators, uses a normalization by a robust estimation of the scale.

There are several ways to measure how 'deep' is a vector with respect to a cloud of points, see for instance the half-space depth of Tukey (1975); Nagy et al. (2019), the simplicial depth in Liu (1990); Liu and Singh (1992), Mahalanobis depth or the projection depth Liu (1992). Taking a point with maximal depth is usually seen as a way to define a median in $\mathbb{R}^d$ (see Radon points inBárány and Mustafa (2020) or Fermat Points in Haldane (1948)). There are therefore several ways to define a median of a cloud of points in $\mathbb{R}^d$. One depth has received a particular attention both in theory and in practice and is known as the Stahel-Donoho outlyingness (SDO), see Stahel (1981); Donoho (1982b). It can be used to construct estimators of multivariate location and scatter known as the Stahel-Donoho estimators (SDE) which were the first equivariant estimators with a high breakdown point. The aim of this work is to show that this notion of depth can be used to construct estimator of a mean vector in $\mathbb{R}^d$ which is robust to adversarial contamination and to heavy-tailed data with respect to $\left\|\Sigma^{-1/2}.\right\|_2$. Let us now define this notion of depth[1] and recall some of its properties.

There is a common approach to many notion of depths for a general d-dimensional set of vectors: first, a definition of depth in $\mathbb{R}$ is given and second, this notion is extended to $\mathbb{R}^d$ simply by applying this one-dimensional definition to the set of one-dimensional projections of the data in all directions $v \in \mathbb{R}^d$ (or all $v \in S$ for some subset $S \subset \mathbb{R}^d$) and then by taking the supremum over all $v \in \mathbb{R}^d$ (or $v \in S$). This approach is based on the idea that if a point in $\mathbb{R}^d$ is an outlier then there must be some direction v such that it is an (univariate) outlier when projected into that direction.

The SDO of $z \in \mathbb{R}$ with respect to a dataset $\{a_1, \dots, a_K\}$ in $\mathbb{R}$ is defined as

$$SDO(z; \{a_1, \dots, a_K\}) = \frac{|z - \mathrm{Med}(a_k)|}{\mathrm{Med}(|a_k - \mathrm{Med}(a_k)|)} \tag{6.3}$$

and a natural extension to $\mathbb{R}^d$ is using the previous one for all one-dimensional projections of the data and by taking the supremum over all directions: for any $\nu \in \mathbb{R}^d$ and a dataset $\{Z_1, \dots, Z_K\}$ in $\mathbb{R}^d$, we set

$$\begin{aligned} SDO(\nu, \{Z_1, \dots, Z_K\}) &= \sup_{v \in \mathbb{R}^d} SDO(\langle \nu, v\rangle; \{\langle Z_1, v\rangle, \dots, \langle Z_K, v\rangle\}) \\ &= \sup_{v \in \mathbb{R}^d} \frac{|\langle \nu, v\rangle - \mathrm{Med}(\langle Z_k, v\rangle)|}{\mathrm{Med}(|\langle Z_k, v\rangle - \mathrm{Med}(\langle Z_k, v\rangle)|)}. \end{aligned} \tag{6.4}$$

A natural way to define a median of the Z_k's is obtained by taking a point with minimal outlyingness (i.e. maximal depth):

$$\hat{\mu}^{SDO} \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} SDO(\mu, \{Z_1, \dots, Z_K\}).$$

[1] The concepts of depth and outlyningness are expressing the same notion but in reverse order.

We note that $\hat{\mu}^{SDO}$ is not the only possible choice to estimate some location of the Z_k's. The Stahel-Donoho location estimator, for instance, is rather defined as a convex sums of the data:

$$\hat{\mu}_K^{SDE} = \frac{\sum_{k=1}^K w_k Z_k}{\sum_{k=1}^K w_k} \tag{6.5}$$

where the weights are some function of the outlyingness of the data, i.e. $w_k = w(SDO(Z_k))$ for some (decreasing) weight function $w : \mathbb{R}^+ \to \mathbb{R}^+$. The weights can also be used to estimate the scatter of the set of points $\{Z_1, \ldots, Z_K\}$ by

$$\hat{\Sigma}_K^{SDE} = \frac{\sum_k w_k (Z_k - \hat{\mu}^{SDE})(Z_k - \hat{\mu}^{SDE})^\top}{\sum_k w_k}. \tag{6.6}$$

Note that there is a more general definition of SDO than the one considered in (6.3) with general (one dimensional) definitions of location and scale statistics; in (6.3), we used the median $\text{Med}(a_k)$ and Median Absolute Deviation (MAD) $\text{Med}(|a_k - \text{Med}(a_k)|)$ for these statistics, see Hampel (1974) for more details.

As mentioned previously several results on the Stahel-Donoho Estimator (SDE) have been established during the last forty years. They are *affine equivariant* meaning that for any affine transformation $x \in \mathbb{R}^d \to Ax + b$ of the dataset by a nonsingular matrix $A \in \mathbb{R}^{d\times d}$ and a vector $b \in \mathbb{R}^d$ the location estimator $\hat{\mu}_K^{SDE}$ is following the same transformation and the scatter estimator $\hat{\Sigma}_K^{SDE}$ is transformed via $M \in \mathbb{R}^{d\times d} \to AMA^\top$. SDE have been proved to have a *finite-sample breakdown point* Donoho and Huber (1983) which is the "smallest amount of contamination necessary to upset an estimator entirely" from Donoho and Gasko (1992) in Donoho (1982a). In Tyler (1994), it is proved that the SDE with MAD replaced by the average of the k_1th and k_2th smallest absolute deviations about the median $\text{Med}(a_k)$ for $k_1 = d - 1 + [(K+1)/2]$ and $k_2 = d - 1 + [(K+2)/2]$ achieves the best finite-sample replacement breakdown point among all affine equivariant estimators obtained in Davies (1987) which is $[(K - d + 1)/2]/K$ (this result holds when the weight function w is continuous and there is an absolute constant c_0 such that $w(r) \leq c_0$, $w(r) \leq c_0/r^2$ for all $r \geq 0$). This result was later extended in Theorem 3.2 from Zuo et al. (2004a). The influence function and the maximum bias of SDE and SD median have been obtained in Zuo et al. (2004b), they can be used to prove robustness properties in Huber's contamination model but not in the adversarial contamination model considered here. These are to our knowledge the only established non-asymptotic properties of Stahel-Donoho estimators.

There are however several asymptotic results for SDE such as a $\sqrt{n}$-consistency in Maronna and Yohai (1995): if the Z_k's are i.i.d. then $\sqrt{K}\left((\hat{\mu}_K^{SDE}, \hat{\Sigma}_K^{SDE}) - (\mathbf{t}, \mathbf{V})\right)$ tends to 0 in probability when $K \to +\infty$ where $\mathbf{t}$ and $\mathbf{V}$ are some location and scatter parameters of the distribution of Z_1. This result holds when the weight function w is such as $|w(r) - w(r')| \leq \gamma \min(1, 1/\min(r, r')^3)|r - r'|$ for all $r, r' \in \mathbb{R}$ and when for all $v \in \mathbb{R}^d$ the cumulative distribution function (cdf) of $\langle Z_1, v\rangle$ denoted by F_v satisfies the following assumption: there exists some absolute constants $c_0 > 0$ and $c_1 > 0$ such that for all $|\epsilon| \leq c_0$

$$|F_v(\text{Med}(F_v) + \epsilon) - F_v(\text{Med}(F_v))| \geq c_1|\epsilon| \text{ and } |F_v(\text{Med}(F_v) \pm \sigma_v + \epsilon) - F_v(\text{Med}(F_v \pm \sigma_v))| \geq c_1|\epsilon| \tag{6.7}$$

where $\text{Med}(F_v) = \inf(x \in \mathbb{R} : F_v(x) \geq 1/2)$ is the median of F_v and $\sigma_v = \text{Med}(G_v)$ where G_v is the cumulative distribution of the random variable $MAD(\langle Z_1, v\rangle) := \text{Med}(|\langle Z_1, v\rangle - \text{Med}(\langle Z_1, v\rangle)|)$. A typical situation mentioned in Maronna and Yohai (1995) where (6.7) holds is when the cdf $F : \mathbb{R}^d \to [0, 1]$ of Z is such that $F = (1 - \eta)F_0 + \eta F^*$ where $\eta < 1$ and F^* is any cdf and F_0 is such that there exists $c_0 > 0$ and $c_1 > 0$ such that for all $v \in \mathbb{R}^d$, $\langle Z_1, v\rangle$ has a density denoted by f_v satisfying $f_v(t) \geq c_1$ for all $t \in [\text{Med}(F_v) \pm c_0] \cup [\text{Med}(F_v) - \sigma_v \pm c_0] \cup [\text{Med}(F_v) + \sigma_v \pm c_0]$.

According to Maronna and Yohai (1995), the later holds when F is spherical with positive density in a neighborhood of 0 and $\sigma_{e_1} e_1$ where $e_1 = (1, 0, \cdots, 0) \in \mathbb{R}^d$. We will come back later on these conditions since we will encounter similar assumptions for our analysis. Finally, asymptotic normality of SDE location estimators have been obtained in Zuo et al. (2004a) under great generality for the location and scatter estimators as well as for the weight function including the median and MAD estimators as in (6.3) and the projection depth obtained for the weight function $w : r \in \mathbb{R}^+ \to 1/(1+r)$. From a stochastic point of view, asymptotic results for $\hat{\mu}_K^{SDE}$ hold when the cdf F is *elliptically symmetric around* μ which means that there exists a symmetric definite positive matrix Σ such that for all $v \in \mathcal{S}_2^{d-1} := \{v \in \mathbb{R}^d : \|v\|_2 = 1\}$, $\langle \Sigma^{-1/2}(Z_1 - \mu), v \rangle$ has the same distribution as $\langle \Sigma^{-1/2}(Z_1 - \mu), e_1 \rangle$ which is a univariate symmetric variable with density function f. In that case, asymptotic normality was obtained when $f(0)f(\sigma) > 0$ where $\sigma = MAD(\langle \Sigma^{-1/2}(Z_1 - \mu), e_1 \rangle)$. Again we will meet this type of condition in our analysis.

On the practical side, SDEs have been used a lot in practice and implementation on various languages such as R exists; and that is one reason why the study of the SDO may be useful, maybe more than some other notions of depth. In the original paper of Stahel (1981), the author proposes a random algorithm where the supremum over all directions $v \in \mathbb{R}^d$ is approximated by subsampling orthogonal directions to $d-1$ hyperplanes generated by d randomly chosen points in the dataset. Other strategies mixing random and deterministic directions have been proposed for instance in Peña and Prieto (2007). Several adaptations and extensions of this algorithm may be found in Debruyne (2009) for an extension to an arbitrary kernel space or in Van Aelst et al. (2011); Van Aelst (2016) for a "cell-wise weights" extension of the SDO where each coordinate of each data receives its own weight. However, only very little is known on the theoretical computational side. In Section 5 of Donoho and Gasko (1992), an algorithm running in time $\mathcal{O}(K^{d+1} \log K)$ is mentioned but its time complexity is making this approach impractical for dimensions larger than 5. There are to our knowledge no theoretical results of any kind on the convergence of some approximate algorithm for the computation of the SDO of a point in $\mathbb{R}^d$ that could be used in practice. As mentioned already in Donoho and Gasko (1992), "some sort of computational breakthrough is necessary to make the estimators, as defined here, really practical". This looks to be still the case. We will however not discuss about this issue in the present work and leave this question still opened.

The aim of this work is to construct mean vector estimators robust to adversarial outliers and heavy-tailed data achieving the deviation-minimax subgaussian rate from (6.2) with respect to the metric $\left\|\Sigma^{-1/2}.\right\|_2$. On our way to our goal, we complement the results on the $\sqrt{n}$-consistency and the asymptotic normality of SDE, by deriving the first non-asymptotic convergence rate for the original SDO median (as well as its median of means version). We also show that the robustness properties of the original SD median and its MOM version goes beyond the Huber's contamination model and that they still persist in the following adversarial corruption model.

Assumption 6.1. *[Adversarial contamination and L_2 inliers] There exist N random vectors $(\tilde{X}_i)_{i=1}^N$ in $\mathbb{R}^d$ which are independent with mean μ and covariance matrix Σ. The N random vectors $(\tilde{X}_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives back the modified dataset $(X_i)_{i=1}^N$ to the statistician. Hence, the statistician receives an "adversarially" contaminated dataset of N vectors in $\mathbb{R}^d$ which can be partitioned into two groups: the modified data $(X_i)_{i\in\mathcal{O}}$, which can be seen as outliers and the "good data" or inliers $(X_i)_{i\in\mathcal{I}}$ such that $\forall i \in \mathcal{I}, X_i = \tilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O} \cup \mathcal{I} = \{1, \ldots, N\}$ is unknown to the statistician.*

In the setup defined by Assumption 6.1, we will use the SDO as one of our building block to

achieve our goal as well as the Median-of-means principle Nemirovsky and Yudin (1983); Alon et al. (1999); Jerrum et al. (1986). This principle has been extensively used during the last decades in particular for the problem of robust mean estimation Lerasle and Oliveira (2011); Devroye et al. (2016); Minsker (2015); Lugosi and Mendelson (2019b,a); M. Lerasle and Lecué (2017); Depersin and Lecué (2020); Hopkins (2018); Cherapanamjeri et al. (2019). The starting point of MOM estimator is to choose an integer $K \in [N]$, split the dataset into K equal size blocks $B_1 \sqcup \cdots \sqcup B_K = [N]$ (w.l.o.g. we assume that N can be divided by K) and construct K empirical means $\bar{X}_k = |B_k|^{-1} \sum_{i \in B_k} X_i$, one over each block. The Stahel-Donoho Median-of-Means that will be used to achieve the subgaussian rate (6.2) with respect to $\left\|\Sigma^{-1/2}.\right\|_2$ in the adversarial and heavy-tailed setup from Assumption 6.1 is

$$\hat{\mu}^{SDO}_{MOM,K} \in \operatorname*{argmin}_{\mu \in \mathbb{R}^d} \sup_{\|v\|_2=1} \frac{|\langle \mu, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|}{\mathrm{Med}(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|)}.$$

It is a min-max MOM estimator but it differs for the min-max MOM estimators introduced in Lecué and Lerasle (2020) because of the renomalization MAD term.

Indeed, unlike recently introduced robust mean estimators, $\hat{\mu}^{SDO}_{MOM,K}$ is using a robust scatter estimator for normalization. Here it is a MOM version of MAD which is used to construct $\hat{\mu}^{SDO}_{MOM,K}$, i.e. $v \to \mathrm{Med}(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|)$. We will show that this normalization plays a central role in the analysis when one wants results w.r.t. the $\left\|\Sigma^{-1/2}.\right\|_2$-norm. But beyond this observation, we will show that MAD and its MOM version satisfy isomorphic and almost-isometric properties that can be used for other tasks such as to construct estimators of the covariance matrix under the existence of only a second moment or for the estimation of a scale matrix when no moment exist but a regularity assumption holds (see Section 6.4 below).

The chapter is organized as follows. In the next section, we consider the case where the good data have a Gaussian distribution and the dataset has been adversarially corrupted. In that case, no need to construct bucketed means and the original Stahel-Donoho median is proved to achieve the subgaussian rate (6.2). The Section 6.3 considers the general adversarial corrupted and heavy-tailed framework from Assumption 6.1 where the MOM version of the SDO is proved to achieve the subgaussian rate. We also exhibit in this section a family of cdfs denoted here by $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$ which plays a key role in our analysis. In particular, when the behavior of these functions around 0 is similar to the one described above in (6.7) then the same result as in the Gaussian case can be obtained and that may hold without the existence of any moment (see Section 6.3.3). In Section 6.4, we show how to use the MOM version of MAD to construct an estimator of the scale matrix under a regularity assumption. In Section 6.5, we explore the properties of the family of functions $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$. A conclusion and open questions are provided in Section 6.6 that are followed by the proofs of all the results in Section 6.7.

Notations. We denote by $x \in \mathbb{R}^d \to \|x\|_2 = \left(\sum_j x_j^2\right)^{1/2}$ the Euclidean norm with associated unit sphere $\mathcal{S}_2^{d-1}$ and unit ball B_2^d. We also denote by $g \sim \mathcal{N}(0,1)$ a standard one-dimensional Gaussian variable and its associated standard Gaussian cdf by $\Phi : t \in \mathbb{R} \to \mathbb{P}[g \le t] = \int_{-\infty}^t \phi(u)du$ where $\phi : u \in \mathbb{R} \to (2\pi)^{-1/2} \exp(-u^2/2)$ is the one dimensional Gaussian density function. We also set $H_G : t \to 1 - \Phi(t)$ and $W_G : p \in (0,1) \to H_G^{(-1)}(p)$ the inverse function of H_G so that $W(p) = \Phi^{-1}(1-p)$.

6.2 The Gaussian case

In this section, we prove that the original SDO median achieves the (non-asymptotic) subgaussian rate (6.2) when the dataset may have been corrupted by an adversary and when the good data have a Gaussian distribution; our main model assumption is the following.

Assumption 6.2. *[Adversarial contamination and Gaussian inliers] There exists N i.i.d. Gaussian vectors $(G_i)_{i=1}^N$ in $\mathbb{R}^d$ with mean μ and (unknown) covariance matrix Σ. We assume that Σ is invertible. The N random vectors $(G_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives the modified dataset $(X_i)_{i=1}^N$ to the statistician.*

We will use the Gaussian case studied in the section as a benchmark case for the later more involved heavy-tailed situations considered after – in these cases, we will bucket the data and make some assumptions on the distribution of the good data. When the "good" data are Gaussian there is no need to bucket the data and the elliptically symmetric property of the Gaussian variables is simplifying the analysis. The mean estimator we use in this section is therefore the median $\hat{\mu}^{SDO} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} SDO(\mu)$ of the original Stahel-Donoho outlyingness function

$$SDO : \mu \in \mathbb{R}^d \to \sup_{v\in\mathbb{R}^d} \frac{|\langle \mu, v\rangle - \mathrm{Med}(\langle X_i, v\rangle)|}{\mathrm{Med}(|\langle X_i, v\rangle - \mathrm{Med}(\langle X_i, v\rangle)|)}. \tag{6.8}$$

Our main result in the adversarial corruption setup with Gaussian inliers is the following:

Theorem 6.1. *We assume that the adversarial contamination with Gaussian inliers model Assumption 6.2 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Let κ_0 be the absolute constant defined in Section 6.5. We assume that $|\mathcal{O}| \le \kappa_0 N$ and $N \ge 4C_0^2\kappa_0^{-2}(d+1)$ (where C_0 is the absolute constants defined in (6.29)). For all $0 < u < \kappa_0^2 N/8$ such that $C_0\sqrt{(d+1)/N} + \sqrt{2u/N} + |\mathcal{O}|/N \le 1/\phi(1)$, with probability at least $1 - 2\exp(-u)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO} - \mu)\right\|_2 \le \frac{6}{\phi(1)}\left(C_0\sqrt{\frac{d+1}{N}} + \sqrt{\frac{2u}{N}} + \frac{|\mathcal{O}|}{N}\right).$$

Let us first remark that if $N < d$ then the N data $X_1, \dots, X_N$ cannot span the entire $\mathbb{R}^d$ space and so there exists a non zero vector $v \in \mathbb{R}^d$ which is orthogonal to all the data points. Hence, $MAD(v) := \mathrm{Med}(|\langle X_i, v\rangle - \mathrm{Med}(\langle X_i, v\rangle)|) = 0$ a.s. and so $SDO(\mu) = +\infty$ for all $\mu \in \mathbb{R}^d$. Therefore, assuming that $N \ge d$ is a minimal assumption when we work with the SDO function.

Theorem 6.1 shows that the SD median $\hat{\mu}^{SDO}$ is robust to adversarial contamination up to a universal constant proportion κ_0 of N and that the rate achieved remains the same as if there was no contamination when $|\mathcal{O}| \lesssim \sqrt{N}\max(\sqrt{u}, \sqrt{d})$. If we put this result with regard to the finite-sample replacement breakdown point (RBP) achieved by the SDE (with a slight modification of MAD at the denominator as recalled in the Introduction), we see that the order of magnitude are the same: SDE and $\hat{\mu}^{SDO}$ can both handle a constant proportion of N adversarial outliers.

It is important to note that the RBP (which is close to $(1/2)N$ when $N >> d$) has the same order of magnitude but a better constant than the one obtained in Theorem 6.1 ($1/2$ versus κ_0 defined in Section 6.5). We want to point out two observations: the result in Theorem 6.1 shows that the estimator still achieves the deviation minimax subgaussian rate (6.2) even up to $\kappa_0 N$ outliers whereas RBP only insures that the estimator does not go to infinity: the two results (RBP and Theorem 6.1) do not quantify the same property. In other words, RBP does not insure

any statistical convergence rate after data corruption whereas Theorem 6.1 does: Theorem 6.1 does not only guarantee that the estimator does not go to infinity, it insures that it stays in a statistically optimal confidence sets (an $\Sigma^{1/2}B_2^d$ ellipsoid with a minimax optimal radius) around the mean. As a consequence, Theorem 6.1 is a stronger statement than a RBP and it shows that the price to pay for this stronger guarantee is just at the level of absolute constants. Secondly, as pointed out in Chen et al. (2018), Remark 2.1, one can in fact ensure some statistical convergence up to a bigger constant fraction of corruption (in their case, 1/3). We point out that, following a large line of works in robust estimation, especially when it comes to achieving (non-asymptotic) sub-gaussian rates, we did not try to optimize the constants in this work, which explain the looseness of this constant. While getting tight constants is interesting and non-trivial, we leave this question opened for future work.

The rate of convergence obtained in Theorem 6.1 has been obtained by several other procedures. For instance, it has been proved that the Tukey median achieves this rate in Chen et al. (2018) when the covariance is proportional to the identity and for the Huber-contamination setup. The same bound was also obtained by a polynomial time algorithm in Dalalyan and Thompson (2019) when the covariance matrix Σ is known.

The proof of Theorem 6.1 (which may be found in Section 6.7) is based on two isomorphic principles of the MAD and SDO functions. We will extend these two properties to the MOM versions of MAD and SDO in the next section. For the moment, let us recall their definitions and write these two properties that are interesting beyond the proof of Theorem 6.1.

The normalization factor in the SDO function (6.8) is called the MAD (median absolute deviation), see Hampel (1973)

$$MAD : v \in \mathbb{R}^d \to \mathrm{Med}(|\langle X_i, v\rangle - \mathrm{Med}(\langle X_i, v\rangle)|).$$

It plays a key role to get an estimation result w.r.t. the $\left\|\Sigma^{-1/2}\cdot\right\|_2$ norm when Σ is unknown. However, this normalization factor requires some more work than for the analysis of classical robust estimators that are only focused on the estimation of the mean. Indeed, $MAD(v)$ is actually a robust estimator of the scatter of $\langle g, v\rangle$ which is $\Phi^{-1}(3/4)\left\|\Sigma^{1/2}v\right\|_2$ (note that if $g \sim \mathcal{N}(0,1)$ then $MAD(g) = \mathrm{Med}(|g - \mathrm{Med}(g)|) = \Phi^{-1}(3/4)$). It is therefore a 'second order' robust estimator but since it appears in the denominator of the SDO function, we cannot only prove an upper estimate for this quantity and we need an isomorphic result – that is upper and lower matching (up to constants) bounds – on the MAD. This result is of independent interest and we are therefore stating it here. The proof is given in Section 6.7.1. We also state a similar isomophic result for SDO which can be use to prove Theorem 6.1. We will see later in Section 6.5 that these metric properties of SDO and MAD can be extended to cases where the mean does not even exist (in that case μ is a *location* parameter) showing that these properties have actually more to do with elliptical symmetry of the underlying data distribution than they have to do with concentration or moment assumption.

Proposition 6.1. *Let $0 < \epsilon < \kappa_0$ (where κ_0 is an absolute constant defined in Section 6.5). We assume that the adversarial contamination model with Gaussian inliers Assumption 6.2 holds with a number of adversarial outliers $|\mathcal{O}| \leq \epsilon N$. We assume that $N \geq 4C_0\epsilon^{-2}(d+1)$ (where C_0 is the absolute constant defined in (6.29)). With probability at least $1 - \exp(-\epsilon^2 N/8)$, for all $v \in \mathbb{R}^d$,*

$$(\Phi^{-1}(3/4) - 2c_0'\epsilon)\left\|\Sigma^{1/2}v\right\|_2 \leq MAD(v) \leq (\Phi^{-1}(3/4) + 2c_0'\epsilon)\left\|\Sigma^{1/2}v\right\|_2$$

where c_0' is the absolute constant defined in Section 6.5.

Moreover, for all $0 < u < \epsilon^2 N/8$, *with probability at least* $1 - 2\exp(-u)$, *for all* $a \in \mathbb{R}^d$, *if* $\left\|\Sigma^{-1/2}(a-\mu)\right\|_2 \geq 2r^*$ *then*

$$\frac{\left\|\Sigma^{-1/2}(a-\mu)\right\|_2}{2(\Phi^{-1}(3/4)+2c_0'\epsilon)} \leq SDO(a) \leq \frac{3\left\|\Sigma^{-1/2}(a-\mu)\right\|_2}{2(\Phi^{-1}(3/4)-2c_0'\epsilon)}$$

and if $\left\|\Sigma^{-1/2}(a-\mu)\right\|_2 \leq 2r^*$ *then* $SDO(a) \leq 3r^*(\Phi^{-1}(3/4)-2c_0'\epsilon)^{-1}$ *where* r^* *is the subgaussian rate from* (6.2) *with the additive adversarial contamination term* $|\mathcal{O}|/N$ *given by*

$$r^* = \frac{1}{\phi(1)}\left(C_0\sqrt{\frac{d+1}{N}} + \sqrt{\frac{2u}{N}} + \frac{|\mathcal{O}|}{N}\right) \tag{6.9}$$

as long as the right-hand side term in (6.9) *is smaller than* 1.

The isomorphic properties of the MAD and SDO functions uniformly over $\mathbb{R}^d$ imply the robustness and subgaussian properties of the SDO median in Theorem 6.1. Similar results for other depths may be found in the literature on robust mean estimation such as the isomorphic property of the Tukey depth proved in Chen et al. (2018).

6.3 The L_2 case and beyond

In this section, we do not anymore assume that the good data follow a Gaussian distribution but we only assume that they have a second moment or that a location and scale parameter exists and some regularity assumption holds (and that the dataset may still be contaminated by an adversary). Nevertheless, even though we are in the heavy tailed setup with adversarially corrupted data we still want to achieve the subgaussian rate for the $\left\|\Sigma^{-1/2}\cdot\right\|_2$-norm. To achieve such a result the median-of-means principle has been proved to perform well. We will therefore use this principle together with the Stahel-Donoho concept of outlyingness. We introduce now an estimator constructed according to these two principles.

Let $K \in [N]$ be the number of blocks and let $\bar{X}_k = (1/|B_k|)\sum_{i\in B_k} X_i, k \in [K]$ be the bucketed means. Outlyingness / depth of a point $\mu \in \mathbb{R}^d$ is measured with respect to the bucketed means:

$$SDO_K(\mu) = \sup_{v\in\mathbb{R}^d} \frac{|\langle \mu, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|}{\mathrm{Med}(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|)}$$

and the Stahel-Donoho Median of means is defined as

$$\hat{\mu}^{SDO}_{MOM,K} \in \operatorname*{argmin}_{\mu\in\mathbb{R}^d} SDO_K(\mu).$$

As for the Gaussian case, the isomorphic and nearly-isometric properties of SDO_K and its denominator, called $MOMAD_K$, play a key role in our analysis. The $MOMAD_K$ is a Median of means version of the Median Absolute Deviation function. We denote it as MOMAD for Median Of Means Absolute Deviation:

$$MOMAD_K : v \in \mathbb{R}^d \to \mathrm{Med}\left(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|\right). \tag{6.10}$$

In the next section, we study metric properties of $MOMAD_K$ and SDO_K that will be useful for our analysis of $\hat{\mu}^{SDO}_{MOM,K}$. Then, we will turn to the statistical bounds obtained for the median $\hat{\mu}^{SDO}_{MOM,K}$ in the general heavy-tailed L_2 setup in Section 6.3.2 and finally we will study some extra regularity assumption of the cdfs $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$ at 0 that allows to get better rates in Section 6.3.3.

6.3.1 Some isomorphic and almost isometric properties of $MOMAD_K$ and SDO_K

In this section, we show that the MOM versions of the SDO and MAD operators (called SDO_K and $MOMAD_K$) satisfy isomorphic and almost-isometry properties under a L_2 moment assumption.

We introduce two families of functions which play a central role in our analysis. They involve the non-corrupted random variables $\tilde{X}_i, i \in [N]$ (and not the corrupted data $X_i, i \in [N]$).

Definition 6.1. *For all* $v \in \mathcal{S}_2^{d-1}$,

$$H_v := H_{N,K,v} : r \in \mathbb{R} \to \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle \geq r\right] \tag{6.11}$$

and

$$W_v := W_{N,K,v} : p \in (0,1) \to H_v^{(-1)}(p),$$

where $H_v^{(-1)}(p) = \max(r \in \mathbb{R} : H_v(r) \geq p)$ *is the generalized inverse of* H_v.

As already observed in the proof of the $\sqrt{n}$-consistency of SDE from Maronna and Yohai (1995) as well as its asymptotic normality in Zuo et al. (2004a), the behavior of the one-dimensional projection cdfs at the median and the two 1/4 and 3/4 quartiles play a central role in the analysis of SDO based estimators. This will also be the case for the MOM version of the SD median. It will appear in Section 6.5 that taking bucketed mean may force toward the Gaussian case for which all these conditions are naturally satisfied because of the elliptical symmetry of Gaussian variables. Let us now state our main assumption on the behavior of the one-dimensional quantile functions $W_v : v \in \mathcal{S}_2^{d-1}$.

Assumption 6.3. *There exists some* $0 < \epsilon < 1/8$ *and some absolute constants* $0 < \varphi_l(\epsilon) < \varphi_u(\epsilon)$ *such that for all* $v \in \mathcal{S}_2^{d-1}$,

$$\max\left(W_v\left(\frac{1}{4} - 2\epsilon\right) - W_v\left(\frac{1}{2} + 2\epsilon\right), W_v\left(\frac{1}{2} - 2\epsilon\right) - W_v\left(\frac{3}{4} + 2\epsilon\right)\right) \leq \varphi_u(\epsilon)$$

and

$$\min\left(W_v\left(\frac{1}{4} + 2\epsilon\right) - W_v\left(\frac{1}{2} - 2\epsilon\right), W_v\left(\frac{1}{2} + 2\epsilon\right) - W_v\left(\frac{3}{4} - 2\epsilon\right)\right) \geq \varphi_l(\epsilon).$$

Assumption 6.3 is a pretty weak assumption since, intuitively, it requires that the distribution of the centered and variance one real-valued random variables $\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle$ have their 1/4-quartiles and medians constant far away as well as for their 3/4-quartiles and medians, and this has to hold uniformly in all directions $v \in \mathcal{S}_2^{d-1}$. For instance, in the Gaussian case, Assumption 6.3 holds for $\varphi_u(\epsilon) = \Phi^{-1}(3/4) + c_0\epsilon$ and $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - c_0\epsilon$ for some absolute constant c_0 and for all $0 < \epsilon < 1/12$ (where we recall that $\Phi : t \to \mathbb{P}[g \leq t]$ where $g \sim \mathcal{N}(0,1)$). Assumption 6.3 appears in our analysis because of the renormalization $MOMAD_K$ term which should not vanish. To understand why the interquartiles range (IQR) appear in this assumption, one may observe that if U is a real-valued random variable and $MAD(U) = \text{Med}(|U - \text{Med}(U)|)$ then

$$\begin{aligned}\min\left(W_U(1/4) - W_U(1/2), W_U(1/2) - W_U(1/4)\right) &\leq MAD(U) \\ &\leq \max\left(W_U(1/4) - W_U(1/2), W_U(1/2) - W_U(1/4)\right)\end{aligned}$$

where W_U is the generalized inverse function of $r \to \mathbb{P}[U \geq r]$. As a consequence, the (IQR) of the projections of the scaled and centered bucketed means should be controlled in all directions v and since we are concerned with non-asymptotic results we allows for small perturbations ϵ around the quartiles: this gives Assumption 6.3.

Proposition 6.2. *We assume that Assumption 6.3 holds for some $0 < \epsilon < 1/8$ and constants $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$. We assume that the adversarial contamination with L_2 inliers model from Assumption 6.1 holds with a number of adversarial outliers $|\mathcal{O}| \leq \epsilon K$. We assume that $K \geq 4C_0^2\epsilon^{-2}(d+1)$ where C_0 is the absolute constant from (6.29). With probability at least $1 - \exp(-\epsilon^2 K/8)$, for all $v \in \mathbb{R}^d$,*

$$\varphi_l(\epsilon)\sqrt{\frac{K}{N}}\left\|\Sigma^{1/2}v\right\|_2 \leq MOMAD_K(v) \leq \varphi_u(\epsilon)\sqrt{\frac{K}{N}}\left\|\Sigma^{1/2}v\right\|_2.$$

Proposition 6.2 shows that $MOMAD_K$ is equivalent to $v \to \sqrt{K/N}\left\|\Sigma^{1/2}v\right\|_2$ up to the two constants $\varphi_u(\epsilon)$ and $\varphi_l(\epsilon)$. We will be interested in two situations regarding these constants. The first one is when their ratio is upper bounded by some absolute constant: there exists an absolute constant $c_1 > 0$ such that for some $0 < \epsilon < 1/8$

$$\frac{\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \leq c_1. \tag{6.12}$$

This condition will be enough to obtain robust optimal subgaussian bounds for $\hat{\mu}^{SDO}_{MOM,K}$ in the two following theorems. If condition (6.12) holds we say that $MOMAD_K$ is isomorphic to $v \to \sqrt{K/N}\left\|\Sigma^{1/2}v\right\|_2$. The second condition, that will be of interest to us is when we will estimate Σ using $MOMAD_K$ in Section 6.4, is when the two constants $\varphi_u(\epsilon)$ and $\varphi_l(\epsilon)$ can be made arbitrarily close to the same constant by taking ϵ small enough, that is when there exists some absolute constants ϕ_0 and $c_0, c_1 > 0$ such that for all $0 < \epsilon < c_0$,

$$\varphi_l(\epsilon) = \phi_0 - c_1\epsilon \text{ and } \varphi_u(\epsilon) = \phi_0 + c_1\epsilon. \tag{6.13}$$

In that case, we speak about an *almost-isometric property* of $MOMAD_K$. The latter condition is stronger than an isomorphic property but it allows to solve a higher order moment estimation problem, the one of estimating Σ. In Section 6.5, we provide several examples where these conditions hold (as well as other properties of the family of cdfs $(H_v : v \in \mathcal{S}_2^{d-1})$) even when there is not even a first moment.

We finish this section with an isomorphic result for SDO_K. The rate of convergence appears in this result: it is the level r^* above which SDO_K is isomorphic to $\nu \in \mathbb{R}^d \to \left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 / \sqrt{K/N}$. One can define it as a solution to

$$C_0\sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + \sup_{\|v\|_2=1} H_{N,K,v}(r^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{6.14}$$

where u is a confidence parameter and C_0 is the absolute constant appearing in (6.29).

Proposition 6.3. *We assume that Assumption 6.3 holds for some $0 < \epsilon < 1/8$ and constants $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$. We assume that the adversarial contamination with L_2 inliers model from Assumption 6.1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. We assume that $|\mathcal{O}| \leq \epsilon K$ and $K \geq 4C_0^2\epsilon^{-2}(d+1)$. Let $u > 0$ and r^* be such that (6.14) holds. Then, with probability at least $1 - \exp(-u) - \exp(-\epsilon^2 K/8)$, for all $\nu \in \mathbb{R}^d$, if $\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 \geq 2\sqrt{K/N}r^*$ then*

$$\frac{\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2}{2\varphi_u(\epsilon)\sqrt{K/N}} \leq SDO_K(\nu) \leq \frac{3\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2}{2\varphi_l(\epsilon)\sqrt{K/N}}$$

and if $\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 \leq 2\sqrt{K/N}r^*$ then $SDO_K(\nu) \leq (3/\varphi_l(\epsilon))r^*$.

Proposition 6.3 may be seen as a MOM version holding in the heavy-tailed case of the Proposition 6.1 obtained in the Gaussian case. Such an extension from the Gaussian case to the L_2 heavy-tail case is made possible thanks to the median-of-means principle and the use of the bucketed means instead of the data themselves. However, we will identify situations where condition (6.12) and (6.14) with an optimal choice of rate r^* (that is for the subgaussian rate (6.2)) hold for $K = N$ even when a first moment does not exist. In that case, one can get a contamination price down to $|\mathcal{O}|/N$ instead of the information theoretic lower bound in the general L_2 case given by $\sqrt{|\mathcal{O}|/N}$ (see Section 6.3.3). We start with the general L_2 case and then we will consider an extra assumption that allows for such better bounds.

6.3.2 The L_2 case

Unlike in Section 6.2 or Section 6.3.3 below where we demand that for all $v \in \mathcal{S}_2^{d-1}$ and all $0 < r < c_0$ the deviation function $H_{N,K,v}(r)$ is less than $1/2 - c_1 r$ here (in this section) we simply use Markov inequality to control the functions $H_{N,K,v}$ around 0. The price we pay by using this approach is that we will not prove anymore estimation results for the SDO MOM over K blocks which hold for all deviation parameters u up to K but only for $u \sim K$. The other price we pay here is for the adversarial contamination cost that will be of the order of $\sqrt{|\mathcal{O}|/N}$ whereas (as proved in Theorem 6.4 below) it can be better up to $|\mathcal{O}|/N$ (as in the Gaussian case from Theorem 6.1). We will be able to achieve this result thanks to a regularity assumption of the cdfs H_v of all one-dimensional projections around 0 (see Assumption 6.4 below). But, for the moment, we do not grant this type of assumption in this section and obtain a general result under the existence of a second moment as well as Assumption 6.3. Subgaussian rates can be derived out of this result when condition (6.12) holds (we refer to Section 6.5 where this condition is studied).

In this section, the bound we use is simply the one deduced from Markov's inequality that is for all $r > 0$ and $K \in [N]$:

$$H_{N,K,v}(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle \geq r\right] \leq \frac{1}{1+r^2}. \tag{6.15}$$

(Note that we used a slightly modification of Markov's inequality: if Z is a centered variance one real-valued random variable then $\mathbb{P}[Z \geq r] = \min_{a\in\mathbb{R}} \mathbb{P}[Z + a \geq r + a] \leq (1 + r^2)^{-1}$). Our main result in the general L_2 setup will follow from this bound and a general result stated in Section 6.7. It is now stated in the following theorem.

Theorem 6.2. *We assume that Assumption 6.3 holds for some $0 < \epsilon < 1/8$. We assume that the adversarial contamination with L_2 inliers model from Assumption 6.1 holds with a number of adversarial outliers $|\mathcal{O}| \leq \epsilon K$. We assume that $K \geq 100\epsilon^{-2}d$. With probability at least $1 - 2\exp(-\epsilon^2 K/15)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO}_{MOM,K} - \mu)\right\|_2 \leq \frac{6\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{K}{N}}.$$

The rate of convergence in Theorem 6.2 can be written like the one in Theorem 6.1 and Theorem 6.4 below where the three terms: complexity, deviation and price for adversarial corruption appear. Indeed, one should notice here that the deviation probability in Theorem 6.2 is fixed equal to $1 - 2\exp(-c_0\epsilon^2 K)$ because we had to take the deviation parameter u equal to K because of the approach based on Markov's inequality (6.15). It is however, equivalent to replace $\sqrt{K/N}$ by $\sqrt{d/(\epsilon^2 N)} + \sqrt{u/\epsilon^2 N} + \sqrt{|\mathcal{O}|/(\epsilon N)}$ for $u = K$ since the two quantities are equivalent under the assumptions of Theorem 6.2. In that case, one may recognize the complexity term $\sqrt{d/N}$, the deviation term $\sqrt{u/N}$ as well as the price for adversarial corruption $\sqrt{|\mathcal{O}|/N}$. In

particular, we see that the price we pay for the corruption is of the order of $\sqrt{|\mathcal{O}|/N}$ which is larger than the $|\mathcal{O}|/N$ term in the Gaussian case from Theorem 6.1 and it is the worst case of Theorem 6.4 below. Indeed, in Theorem 6.2 we did not exploit any other property than the existence of a second moment whereas the other two Theorems 6.1 and Theorem 6.4 exploit some regularity assumption around 0 of the family of functions $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$.

Adaptation to K via Lepski's method. It follows from Theorem 6.2 that $\hat{\mu}^{SDO}_{MOM,K}$ is an estimator which depends on the deviation parameter K : we need to specify a value of K, which is used to build the estimator, so we need, for instance, some prior knowledge on the number of outliers. However, even if we do not have such prior knowledge, it is possible to overcome this difficulty by constructing an adaptive to K version of this estimator to disentangle the estimator from the parameter K. The classical way to do it is via Lepski's method Lepskiĭ (1990, 1991). Usually, the price we pay to make this approach work is some extra knowledge on Σ such as its trace and operator norm (see for instance Depersin and Lecué (2019), Section 6). An interesting feature of SDO type estimators is that we do not need no such information on Σ to build this adaptation scheme: we only need knowledge on $\varphi_u(\epsilon)$ and $\varphi_l(\epsilon)$. Let us now construct this adaptive scheme: the number of blocks is chosen adaptively via

$$\hat{K} = \min\left(K \in [N] : SDO_k(\hat{\mu}^{SDO}_{MOM,K} - \hat{\mu}^{SDO}_{MOM,k}) \leq \max\left(\frac{9}{\varphi_l(\epsilon)}, \frac{6\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)}\left(1+\sqrt{\frac{K}{k}}\right)\right), \forall k = N, \dots, K\right). \tag{6.16}$$

Theorem 6.3. *There are absolute constants c_0 and c_1 such that the following holds. We assume that Assumption 6.3 holds for some $0 < \epsilon < 1/8$ and all $K \in [N]$. We assume that the adversarial contamination with L_2 inliers model from Assumption 6.1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Then, for all $K \geq \max(c_0\epsilon^{-2}d, c_0|\mathcal{O}|)$ with probability at least $1 - 2\exp(-c_1\epsilon^2 K)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO}_{MOM,\hat{K}} - \mu)\right\|_2 \leq \frac{28\varphi_u^2(\epsilon)}{\varphi_l^2(\epsilon)}\sqrt{\frac{K}{N}}.$$

where $\hat{K}$ is the adaptive choice of number of blocks from (6.16).

6.3.3 Beyond the L_2 case and a regularity condition around 0 of the H_v's

In this section, we obtain an estimation bound for the MOM version of the SDO median in the adversarial corruption model under an extra assumption on the regularity at 0 of the family of functions $H_v, v \in \mathcal{S}_2^{d-1}$ that is stated now.

Assumption 6.4. *There exists a location parameter $\mu \in \mathbb{R}^d$, a scale matrix $\Sigma \succeq 0$ and some absolute constants $c_0, c_1 > 0$ and $c_2 > 0$ and such that for all $v \in \mathcal{S}_2^{d-1}$ and all $(2C_0/c_1)\sqrt{(d+1)/K} \leq r \leq c_0$ (where C_0 is the absolute constant from (6.29))*

$$H_{N,K,v}(r) = H_v(r) := \mathbb{P}\left[\frac{1}{\sqrt{N/K}}\sum_{i=1}^{N/K}\langle\Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle \geq r\right] \leq \frac{1}{2} - c_2 r.$$

This assumption is about the behavior around the origin of the cdf of all one-dimensional projections of the random vectors $(N/K)^{-1/2}\sum_{i=1}^{N/K}\Sigma^{-1/2}(\tilde{X}_i - \mu)$ where the $\tilde{X}_i$ are the non-corrupted data. The term $\frac{1}{2} - c_2 r$ in the bound above is the behavior of regular in 0 cdfs such as in the Gaussian case (see Section 6.5 for more details and more examples).

Our main result in the adversarial corruption model under Assumption 6.4 is the following theorem. The proof may be found in Section 6.7.

Theorem 6.4. *We assume that Assumption 6.3 holds for some* $0 < \epsilon < 1/8$ *and that Assumption 6.4 holds as well with constants* c_0, c_1 *and* c_2*. We assume that the adversarial contamination model holds with a number of adversarial outliers* $|\mathcal{O}| \leq \epsilon K$*. We assume that* $K \geq 4C_0^2\epsilon^{-2}(d+1)$*. For all* $0 < u \leq \epsilon^2 K/8$ *such that* $C_0\sqrt{(d+1)/K} + \sqrt{2u/K} + |\mathcal{O}|/K \leq c_0c_2/2$*, with probability at least* $1 - 2\exp(-u)$*,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO}_{MOM,K} - \mu)\right\|_2 \leq \frac{4\varphi_u(\epsilon)}{c_2\varphi_l(\epsilon)}\left(C_0\sqrt{\frac{d+1}{N}} + \sqrt{\frac{2u}{N}} + \frac{|\mathcal{O}|}{\sqrt{NK}}\right). \tag{6.17}$$

We recover the optimal subgaussian rate (6.2) in Theorem 6.4 when for some $0 < \epsilon < 1/8$, condition (6.12) holds and $|\mathcal{O}| \lesssim \sqrt{Kd}$. The term $|\mathcal{O}|/\sqrt{KN}$ appearing in the convergence rate of Theorem 6.4 is the price we pay for the adversarial contamination. It is between $\sqrt{|\mathcal{O}|/N}$ when $K \sim |\mathcal{O}|$ and $|\mathcal{O}|/N$ when $K \sim N$. We note that the rate gets better when K comes closer to N: however we note that we cannot always choose K as we please. Indeed both Assumptions 6.3 and 6.4 are assumptions on the functions $H_{N,K,v}$, and they might be true for some K and not for others. Usually when the inliers are in L_2, and without further assumptions, the information theoretic lower bound is known to be of the order of $\sqrt{|\mathcal{O}|/N}$ and not of order $|\mathcal{O}|/N$. We get a better rate in Theorem 6.4 thanks to Assumption 6.4 which is using in some more efficient way the regularity of the H_v functions at 0.

Remark 6.1. *Assumption 6.3 and Assumption 6.4 do not need the data to have a first moment: both assumptions may hold* without the existence of any moment*. In these assumptions,* μ *or* Σ *are not used in the role of mean and variance matrix but can be thought of as location and scatter parameters. Such parameters may exist even in situations where there is not even a first moment. We note that in Theorem 6.2, on the contrary, we use Markov's inequality instead of Assumption 6.4 and so we need* μ *to be the mean and* Σ *to be the covariance matrix and not just a scatter matrix – hence, we need the existence of two moments in Theorem 6.2 but not of any moment in Theorem 6.4.*

Unlike typical results in the MOM literature except for the one obtained in Minsker and Strawn (2017), the deviation rate in Theorem 6.4 is $1 - 2\exp(-u)$ for all $u \lesssim K$, in particular it does not have to depend on parameter K. As a consequence, the estimator $\hat{\mu}^{SDO}_{MOM,K}$ does not depend on the deviation parameter. Usually, results for MOM estimators constructed on K blocks are given with probability at least $1 - \exp(-c_0K)$ and then a Lepski's method is used to construct an adaptive to K procedure (as we did in the previous section). This is not the case here nor it is for the Gaussian case in Section 6.2. This is again because Assumption 6.4 is efficiently using the behavior of $H_{N,K,v}$ around 0.

6.4 Estimation of Σ using MOMAD

In this section, we show that it is possible to estimate a scale or covariance matrix Σ using the MOMAD estimator. In particular, given that the isomorphic property of MOMAD hold under Assumption 6.3 (which does not grants the existence of a second moment), we show that it is possible to estimate a scale matrix under only this assumptions. This differs from approaches based on the empirical covariance matrix where at best a $L_{2+\delta}$-moment assumption for some positive δ is granted for the estimation of the covariance matrix, see Lounici (2014); Cai et al. (2016); Lu et al. (2020). In this section, we construct two estimators of Σ.

We show that for the estimation of Σ via the MOMAD, the properties of $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ introduced in Assumption 6.3 play a key role. Let us first have a look at these quantities in the Gaussian case. In that case, there are some absolute constants ϕ_0 and $c_0, c_1 > 0$ such that for all $0 < \epsilon < c_0$,

$$\varphi_l(\epsilon) = \phi_0 - c_1\epsilon \text{ and } \varphi_u(\epsilon) = \phi_0 + c_1\epsilon \tag{6.18}$$

where $\phi_0 = \Phi^{-1}(3/4)$ (see Section 6.5 or the proof of Proposition 6.1 for more details). This latter result holds in the Gaussian case first because the two interquartile intervals have the same length: $\Phi^{-1}(0) - \Phi^{-1}(1/4) = \Phi^{-1}(3/4) - \Phi(0) = \phi_0$ and, second, because the Gaussian density function is uniformly lower bounded by an absolute positive constant locally around the two $1/4$ and $3/4$ quartiles $\Phi^{-1}(1/4)$ and $\Phi^{-1}(3/4)$ as well as around the median $\Phi^{-1}(1/2) = 0$. If this last condition were not true at some $q \in \{W(1/4), W(1/2), W(3/4)\}$ where $W = W_{N,K,v}$ for some direction $v \in \mathcal{S}_2^{d-1}$ then there will be some plateau of the cdf $r \in \mathbb{R} \to 1 - H_{N,K,v}(r)$ starting at q and thus there would be a constant factor gap between $W(\ell/4 - 2\epsilon)$ and $W(\ell/4 + 2\epsilon)$ for some $\ell \in \{1, 2, 3\}$. In that case, there would be some absolute constants $c_0 > 0$ and $c_1 > 0$ such that $|\varphi_l(\epsilon) - \varphi_l(\epsilon)| \geq c_0$ for all $0 < \epsilon \leq c_1$. In particular, we would only have an isomorphic property for the MOMAD and thus it is not clear how to estimate Σ using MOMAD at a better rate than a constant rate. Typical values of ϕ_0 in (6.18) will be $\phi_0 = W(1/4) - W(1/2) = W(1/2) - W(3/4)$. In particular, the interquartile interval lengths have to be equal in all directions $v \in \mathcal{S}_2^{d-1}$; this will hold, in particular, under a spherical symmetry assumption of the $\Sigma^{-1/2}(\tilde{X}_i - \mu)$ (see Section 6.5 for a more formal statement). That is a reason why we will use an isometric property of MOMAD (and not just an isomorphic property) and that to insure this property we consider the following assumption.

Assumption 6.5. *For the same choice of K as in Assumption 6.3 where $\epsilon > 0 \to \varphi_l(\epsilon), \varphi_u(\epsilon)$ are defined, there are absolute constants ϕ_0, $c_0, c_1 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $0 < \epsilon < c_0$, $\varphi_l(\epsilon) = \phi_0 - c_1\epsilon$ and $\varphi_u(\epsilon) = \phi_0 + c_1\epsilon$.*

Let us now turn to the construction of two estimators of the covariance matrix Σ using MOMAD under Assumption 6.5 (as well as Assumption 6.3). Because of the constant factor ϕ_0 in Assumption 6.5 we will provide an estimator of the *scatter matrix* $\phi_0^2\Sigma$ (according to Maronna et al. (2006b), a scatter matrix is any matrix proportional to the covariance matrix – this type of matrix gives in particular information on the relative uncertainty in all directions).

It follows from Proposition 6.2 that $MOMAD_K$ is isomorphic to $v \in \mathbb{R}^d \to \phi_0\sqrt{K/N}\left\|\Sigma^{1/2}v\right\|_2$ and that under Assumption 6.5 it becomes an almost isometry, that is, with probability at least $1 - \exp(-\epsilon^2 K/8)$, for all $v \in \mathbb{R}^d$,

$$\left|MOMAD_K(v) - \phi_0\sqrt{\frac{K}{N}}\left\|\Sigma^{1/2}v\right\|_2\right| \leq c_1\epsilon\sqrt{\frac{K}{N}}\left\|\Sigma^{1/2}v\right\|_2 \tag{6.19}$$

as long as $|\mathcal{O}| \leq \epsilon K$, $K \geq 4C_0^2\epsilon^{-2}(d+1)$ and $0 < \epsilon < c_0$. In the Gaussian case and other spherical cases studied in Section 6.5, this almost isometric property holds for $K = N$ (and $MOMAD_N = MAD$) and any $0 < \epsilon < 1/12$: it follows from Proposition 6.1 that with probability at least $1 - \exp(-\epsilon^2 N/8)$, for all $v \in \mathbb{R}^d$,

$$\left|MAD(v) - \Phi^{-1}(3/4)\left\|\Sigma^{1/2}v\right\|_2\right| \leq c_1\epsilon\left\|\Sigma^{1/2}v\right\|_2. \tag{6.20}$$

We then may use two distinct ideas to build an estimator from (6.19) and (6.20). The first one is to consider the matrix $\check{\Sigma}$ defined by

$$\check{\Sigma} \in \frac{N}{K}\operatorname*{argmin}_{A \succeq 0} \max_{\|A^{1/2}v\|_2 = 1} |MOMAD_K(v) - 1|$$

where the minimum is taken over the cone of semi-definite positive matrices. The following estimation error bound follows from (6.19) and basic algebra (see the proof in Section 6.7).

Proposition 6.4. *Assume that Assumption 6.1 holds. Let $K \in [N]$, φ_l and φ_u be such that Assumption 6.3 and Assumption 6.5 hold with constants ϕ_0, c_0 and c_1, and that $4c_1\epsilon < \phi_0$. Then, for all $0 < \epsilon < c_0$ such that $|\mathcal{O}| \leq \epsilon K$ and $K \geq 4C_0^2\epsilon^{-2}(d+1)$, with probability at least $1 - \exp(-\epsilon^2 K/8)$,*

$$\left\|\Sigma^{-1/2}\check{\Sigma}\Sigma^{-1/2} - \phi_0^2 \,\mathrm{Id}\right\|_{op} \leq 12\phi_0 c_1 \epsilon$$

Comparing the rate obtained in Proposition 6.4 with the ones from the literature, we notice that this estimator achieves a rate of the order of the contamination rate $\epsilon = |\mathcal{O}|/N$ when one can choose $K \sim N$ in the assumptions of Proposition 6.4. This is the typical rate when the data are Gaussian, while the typical and information-theoretically optimal rate for L_2 inliers is like $\sqrt{\epsilon}$ (see for instance Kothari and Steurer (2017), discussion after Theorem 1.2). Once again, we get a better rate in Proposition 6.4 thanks to Assumption 6.5 which is using in some efficient way the behavior of the two φ_u, φ_l functions around 0 and so the isometric property of MOMAD, which explains this gap.

We then present a second way to use (6.19) to estimate directly the entries of Σ following an idea from Gnanadesikan and Kettenring (1972). This way will lead to a somehow worse rate, but it provides an estimator that is very easy to compute, and that is tractable in time $\mathcal{O}(N\log(N)d)$, that is in linear time.

Let $(e_j)_{j=1}^d$ denote the canonical basis of $\mathbb{R}^d$. We have, for all $i, j \in [d]$,

$$4\Sigma_{ij} = 4\langle e_i, \Sigma e_j\rangle = \left\|\Sigma^{1/2}(e_i + e_j)\right\|_2^2 - \left\|\Sigma^{1/2}(e_i - e_j)\right\|_2^2.$$

As a consequence, a natural estimator of $\phi_0^2\Sigma$ based on $MOMAD_K$ is the matrix $\hat{\Sigma}$ whose entries are defined for all $i, j \in [d]$ by

$$\hat{\Sigma}_{ij} = \frac{N}{4K}\left(MOMAD_K^2(e_i + e_j) - MOMAD_K^2(e_i - e_j)\right).$$

Note that $\hat{\Sigma}$ is symmetric but it may not be positive semi-definite (PSD). To overcome this issue, a projection method has been introduced in Lu et al. (2020) which may also be used as well for $\hat{\Sigma}$. Our main statistical bound for $\hat{\Sigma}$ is the following.

Proposition 6.5. *Assume that Assumption 6.1 holds. Let $K \in [N]$, φ_l and φ_u be such that Assumption 6.3 and Assumption 6.5 hold with constants ϕ_0, c_0 and c_1. Then, for all $0 < \epsilon < c_0$ such that $|\mathcal{O}| \leq \epsilon K$ and $K \geq 4C_0^2\epsilon^{-2}(d+1)$, with probability at least $1 - \exp(-\epsilon^2 K/8)$,*

$$\max_{i,j\in[d]}\left|\frac{\phi_0^2\Sigma_{ij} - \hat{\Sigma}_{ij}}{\Sigma_{ii} + \Sigma_{jj}}\right| \leq \sup_{\|u\|_1 = \|v\|_1 = 1}\left|\frac{\langle u, (\phi_0^2\Sigma - \hat{\Sigma})v\rangle}{\sum_i(|u_i| + |v_i|)\Sigma_{ii}}\right| \leq c_1\epsilon(c_1\epsilon + \phi_0)/2.$$

In particular, if one can choose $K = N$ so that Assumption 6.3 and Assumption 6.5 hold – for instance, in the Gaussian case or for other spherical variables as in Section 6.5 – then the $MOMAD_N$ estimator becomes the classical MAD one and for $\epsilon^2 = c_2 d/N$ we have that with probability at least $1 - \exp(-c_4 d)$,

$$\max_{i,j\in[d]}\left|\frac{\phi_0^2\Sigma_{ij} - \hat{\Sigma}_{ij}}{\Sigma_{ii} + \Sigma_{jj}}\right| \leq \sup_{\|u\|_1 = \|v\|_1 = 1}\left|\frac{\langle u, (\phi_0^2\Sigma - \hat{\Sigma})v\rangle}{\sum_i(|u_i| + |v_i|)\Sigma_{ii}}\right| \leq c_5\sqrt{\frac{d}{N}}.$$

as long as $|\mathcal{O}| \leq c_6 d$.

6.5 Study of the $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ functions

The functions $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ play a key role in our analysis. Their behavior in a neighborhood of their 1/4 and 3/4 quartiles and medians should be controlled so that Assumption 6.3 may hold: they are driving the isomoprhic properties and almost isometric properties of the $MOMAD_K$ and SDO_K functions and so of the statistical performance of the Stahel Donoho Median and its MOM version. Their behavior around 0 also drives the improved rates obtained under Assumption 6.4. From our perspective, it is of the utmost importance to understand the behavior of these functions at these particular points.

Let us first settle down the properties of the $H_{N,K,v}$ functions desirable for our analysis. We set $Z_i = \Sigma^{-1/2}(\tilde{X}_i - \mu)$ for all $i \in [N]$ so that the Z_i's are independent centered isotropic vectors in $\mathbb{R}^d$ and $n = N/K$. We want to identify conditions on the distributions of the Z_i's such that

- for Assumption 6.4: there exists some absolute constants $c_0, c_1 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $0 < r < c_0$,

$$H_{n,v}(r) := \mathbb{P}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\langle Z_i, v\rangle \geq r\right] \leq \frac{1}{2} - c_1 r. \tag{6.21}$$

- for Assumption 6.3 and the two following conditions (6.12) and (6.18): there exists an absolute constant $c_1 > 0$ and $0 < \epsilon < 1/8$ such that $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ exist and are such that

$$\frac{\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \leq c_0 \tag{6.22}$$

  or there are absolute constants ϕ_0 and $c_0, c_1 > 0$ such that for all $0 < \epsilon < c_0$,

$$\varphi_l(\epsilon) = \phi_0 - c_1\epsilon \text{ and } \varphi_u(\epsilon) = \phi_0 + c_1\epsilon \tag{6.23}$$

  which are respectively Condition (6.12) (insuring an isomorphic property of $MOMAD_K$ and SDO_K as well as optimal subgaussian rates for SD median and median of means) and Condition (6.18) (insuring almost isometric property of $MOMAD_K$ as well as estimation properties for $\hat{\Sigma}$ in Section 6.4).

Let us first study the Gaussian case which is our benchmark situation. We will then study other cases where the family of functions $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ satisfies these conditions.

The Gaussian case. We recall that $\Phi : t \in \mathbb{R} \to \mathbb{P}[g \leq t] = \int_{-\infty}^{t}\phi(u)du$ where $\phi : u \in \mathbb{R} \to (2\pi)^{-1/2}\exp(-u^2/2)$ is the Gaussian density function. We also denote $H_G : t \to 1 - \Phi(t)$ and $W_G : p \in (0,1) \to H_G^{(-1)}(p)$ the inverse function of H_G so that $W_G(p) = \Phi^{-1}(1-p)$. It follows from the mean value theorem that for all $t, \epsilon \in \mathbb{R}_+$, $|H_G(t+\epsilon) - H_G(t)| \geq \phi(t+\epsilon)\epsilon$ so that around 0 we have for all $c_0 > 0$ and $0 < r < c_0$, $H_G(r) \leq 1/2 - \phi(c_0)r$. As a consequence, (6.21) holds in the Gaussian case, for instance, with $c_0 = 1$ and $c_1 = \phi(1)$. Let us now look at the two other conditions in the Gaussian case. Using Taylor formulae and that for all $p \in (0,1)$, $W_G'(p) = [\phi(W_G(p))]^{-1}$ and $W_G''(p) = -W_G(p)/[\phi(W_G(p))]^2$, we show that one can take φ_u and φ_l defined for all $0 \leq \epsilon \leq 1/12$ by

$$\varphi_u(\epsilon) = \Phi^{-1}(3/4) + c_0'\epsilon + c_1'\epsilon^2 \text{ and } \varphi_l(\epsilon) = \Phi^{-1}(3/4) - c_0'\epsilon - c_1'\epsilon^2$$

where $c_0' := 2(\phi(\Phi^{-1}(3/4))^{-1} + \phi(0)^{-1})$ and $c_1' := 4\Phi^{-1}(11/12)/[\phi(\Phi^{-1}(11/12))]^2$. In particular, using that $(1-t)^{-1} \leq 1 + 2t$ for all $0 \leq t \leq 1/2$, we get that one can choose $\varphi_u(\epsilon) =$

$\Phi^{-1}(3/4) + 2c_0'\epsilon$ and $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - 2c_0'\epsilon$ and $\varphi_u(\epsilon)/\varphi_l(\epsilon) \leq 3$ for all $0 \leq \epsilon \leq \kappa_0$ where $\kappa_0 := \min(1/[8c_0'], 1/\sqrt{2c_1'}, c_0'/c_1')$.

So that both conditions (6.21) and (6.23) hold with $\phi_0 = \Phi^{-1}(3/4)$. In particular, we also recover that the values of the density function ϕ at the 1/4 and 3/4 quartiles (here we used that $\phi(\Phi^{-1}(3/4)) = \phi(\Phi^{-1}(1/4))$) and at the median $\phi(0)$ (here we used that $\Phi^{-1}(1/2) = 0$) play a key role for the study SDO estimators as it was previously observed in several works on SDE.

In the following, we identify situations where the $H_{N,K,v}, v \in \mathcal{S}_2^{d-1}$ functions and their pseudo inverses mimic the H_G and W_G functions from the Gaussian case. There are at least two reasons for that to happen: the first one is that we are projecting random vectors leaving in $\mathbb{R}^d$ onto one dimensional subspaces; the second reason is that we are averaging random variables having a second moment. We will explore these two observations in the two following paragraphs.

One dimensional projections and elliptically contoured distributions. The fact that the H_v functions deal only with one-dimensional marginals is making these functions likely to behave as in the Gaussian case since one-dimensional projections of sufficiently spherically symmetric random vectors in $\mathbb{R}^d$ are expected to behave like one-dimensional Gaussian variables and this phenomenon is even more accentuated when d is large (this is one particular situation where large dimension d may help in Statistics). Indeed, one may have in mind an observation – sometimes attributed to H. Poincaré – that the density function of the one-dimensional projection $\langle \sqrt{d}U, e_1 \rangle$ – where $\sqrt{d}U$ is uniformly distributed over $\sqrt{d}\mathcal{S}_2^{d-1}$ and $(e_j)_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$ – converges to the density of a $\mathcal{N}(0,1)$ when $d \to \infty$ (see page 16 in Ledoux and Talagrand (2011) or Chapter 4 in Bryc (1995)). One may also have in mind that there are directions such as $v = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ which are mixing the coordinates of $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ when projected onto v and therefore may have the tendency to mimic a standard Gaussian variable because of the CLT. Note that all these observations hold for $N = K$ that is even for $n = 1$: because of the one-dimensional projections we may not even have to average the Z_i's to mimic the Gaussian case. Therefore, Theorem 6.5 can be extended beyond the Gaussian case when this phenomenon occurs.

Let us now consider an example of elliptically contoured distributions where this happens to be true. Our aim is to show that Condition (6.12) and Assumption 6.4 (and so Theorem 6.4) may hold for $K = N$ (i.e. $n = 1$) even when the $\tilde{X}_i$'s do not have a first moment.

We assume that the $\tilde{X}_i$'s are i.i.d. and that $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ has a spherically symmetric distribution; in that case, $\tilde{X}_1 - \mu$ is sometimes said to have an *elliptically contoured distribution*. Then, there exists a non-negative random variable R such that $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ is distributed according to RU where U is uniformly distributed on $\mathcal{S}_2^{d-1}$ and is independent of R (see Chapter 4 in Bryc (1995)). In that case, all the $\langle \Sigma^{-1/2}(\tilde{X}_1 - \mu), v \rangle$ for $v \in \mathcal{S}_2^{d-1}$ have the same distribution as $\langle \Sigma^{-1/2}(\tilde{X}_1 - \mu), e_1 \rangle$ (where $(e_j)_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$) which is distributed according to $R\langle U, e_1 \rangle$. Now, using that $\langle U, e_1 \rangle$ is absolutely continuous w.r.t. the Lebesgue measure with density function given by, when $d \geq 2$,

$$t \in \mathbb{R} \to C_d(1-t^2)^{\frac{d-3}{2}} I(|t| \leq 1) \text{ where } C_d = \left(\int_{-1}^{1} (1-t^2)^{\frac{d-3}{2}} dt \right)^{-1} = \frac{2\Gamma(d/2)}{\Gamma((d-1)/2)\sqrt{\pi}}$$

and Γ is the Gamma function, we can deduce that (even for $K = N$), H_v is independent of $v \in \mathcal{S}_2^{d-1}$ and is such that for all $r \geq 0$, $H_v(-r) = 1 - H_v(r)$ and

$$H_v(r) = H(r) := C_d \int_0^1 \mathbb{P}[R \geq r/x] \left(1 - x^2\right)^{\frac{d-3}{2}} dx.$$

In particular, we recover that $H(0) = 1/2$ since $R \geq 0$ a.s.. Let us now consider a simple example for the distribution of R. In that example, R takes values $r_1 < r_2 < \cdots$ such that $\alpha_j = \mathbb{P}[R = r_j]$ for all $j \in \mathbb{N}^*$ so that for all $q > 0$, $\mathbb{E}R^q = \sum_j r_j^q \alpha_j$ which may be infinite even for $q = 1$ (that is when there is not even a first moment). For this example, we have for all $r \geq 0$,

$$H(r) = C_d \sum_{j=1}^{\infty} \alpha_j \int_{r/r_j}^{1} (1 - x^2)^{\frac{d-3}{2}} dx I(r \leq r_j).$$

In particular, H is differentiable and $R\langle U, e_1 \rangle$ is absolutely continuous w.r.t. the Lebesgue measure with a density function given by

$$f : r \in \mathbb{R} \to -H'(r) = C_d \sum_{j=1}^{\infty} \frac{\alpha_j}{r_j} \left[1 - \left(\frac{r}{r_j} \right)^2 \right]^{\frac{d-3}{2}} I(r \leq r_j).$$

In particular, for $r_\infty = \lim_{j \to \infty} r_j$, H is strictly decreasing on $[0, r_\infty)$ from $H(0) = 1/2$ to $H(r_\infty) = 0$ and beyond r_∞ it is constant equal to 0. Therefore, for all $v \in \mathcal{S}_2^{d-1}$, the generalized inverse W_v of H_v is independent of v and it is the inverse of H: for all $p \in (0, 1/2]$ there is a unique element $W(p)(= W_v(p))$ in $[0, r_\infty)$ such that $H(W(p)) = p$ and $W(1 - p) = -W(p)$.

Now, let us choose $r_j = 2^j C_d$ and $\alpha_j = 2^{-j}$ for all $j \in \mathbb{N}$. We also assume $d \geq 4$ to make the presentation simpler (the cases $d = 1, 2, 3$ can be treated separately). In that case, $\mathbb{E}R = +\infty$ and so the mean and covariance matrix do not exist. Nevertheless, one may still assume that there exists $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ definite positive such that $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ is spherically symmetric (without having μ to be a mean vector and Σ to be a covariance matrix). Then, Theorem 6.4 still applies.

Let us first check Condition (6.21). We have $H(0) = 1/2$ and for all $0 \leq r \leq C_d/\sqrt{d-3}$,

$$f(r) \geq \sum_{j=1}^{\infty} \frac{1}{2^{2j}} \left[1 - \frac{d-3}{2C_d^2} \left(\frac{r}{2^j} \right)^2 \right] \geq \frac{1}{3}. \tag{6.24}$$

Moreover, we see that $\sqrt{d} \leq C_d \leq 6\sqrt{d}$, hence, (6.24) holds for all $0 \leq r \leq 1$. Which is according to the mean value theorem enough to show that Condition (6.21) holds (see (6.26) below for more details).

Let us now check conditions (6.22) and (6.23). It follows from Proposition 6.6 below that, it is enough to lower bound the density function f in a neighborhood of p for $p \in \{W(1/4), W(1/2), W(3/4)\}$ and that $W(1/4) - W(3/4)$ is an absolute constant. But, given that $W(1/2) = 0$ and (6.24) holds, that f is symmetric about 0 and that $W(1/4) = -W(3/4)$, we only have to check that $f(q) \geq c_0$ for all $q \in [W(1/4) - 2\epsilon, W(1/4) + 2\epsilon]$ for some $0 < \epsilon < 1/8$ and an absolute constant c_0 and that $W(1/4)$ is an absolute constant. We first have to find $W(1/4)$ which is the unique solution r such that $H(r) = 1/4$. We see that f is symmetric unimodal with maximal value at 0 given by $f(0) = 4/3$ and we showed that $f(r) \geq 1/3$ for all $0 \leq r \leq 1$ in (6.24). Therefore, $H(1/8) \geq 1/3$ and $H(1 - 1/10) \leq 3/15 < 1/4$, hence, $W(1/4) \in [1/8, 1 - 1/10]$. It follows from (6.24) that $f(q) \geq 1/3$ for all $q \in [W(1/4) - 1/10, W(1/4) + 1/10]$. We conclude that both conditions (6.22) and (6.23) hold thanks to Proposition 6.6 below.

For this example, one can take $\varphi_u(\epsilon) = W(1/4) - (4/3)\epsilon$ and $\varphi_l(\epsilon) = W(1/4) + (4/3)\epsilon$ for all $0 < \epsilon < 1/16$. In that case, $MOMAD_K$ is an almost isometry and we can state a result like Theorem 6.1 where $\Phi^{-1}(3/4)$ is replaced by $W(1/4)$ and μ and Σ are not anymore the mean and covariance matrix since they do not exist but 'location' and 'scale' parameters defined such

that $\Sigma^{-1/2}(\tilde{X}_i - \mu)$ are spherically symmetric. As a consequence, the phenomenon underlying the Gaussian case from Section 6.2 has nothing to do with concentration but it is more about elliptical symmetry.

Gaussian approximation. In cases where there is some lack of spherical symmetry of $\Sigma^{-1/2}(\tilde{X}_1 - \mu)$ one may study the H_v functions for a smaller number K of blocks so that $n = N/K$ may be large enough to see some averaging effect. In that case and because Gaussian variables satisfy all the properties we need, it is tempting to use a Gaussian approximation result such as a Berry-Esseen bound (see Petrov (1995); Chen and Shao (2012, 2001)) to approximate the H_v functions by $1 - \Phi$ for $n = N/K$ large enough. This strategy has been used several times in Minsker and co-authors works on Median-of-means and Catoni's type of estimators (see for instance Minsker and Strawn (2017); Minsker (2018b)).

For instance, when for all $v \in \mathcal{S}_2^{d-1}$, $\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle, i \in [n]$ are (independent, centered and variance one) real-valued random variables in $L_{2+\delta}$ such that $\left\|\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle\right\|_{2+\delta} \leq \kappa$ (uniformly in $v \in \mathcal{S}_2^{d-1}$) for some $\delta > 0$ then, it follows from Theorem 5.7 in Petrov (1995) that there is an absolute constant $c_0 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and all $r \in \mathbb{R}$,

$$|H_{N,K,v}(r) - \mathbb{P}[g \geq r]| \leq \frac{c_1 \kappa^{2+\delta}}{n^{\delta/2}} := c_n \tag{6.25}$$

It follows that for all $p \in (0,1)$ and $\epsilon \in \mathbb{R}$ satisfying $p + \epsilon \in (0,1)$ that

$$\Phi^{-1}(1 - p - \epsilon - c_n) \leq W(p + \epsilon) \leq \Phi^{-1}(1 - p - \epsilon + c_n).$$

In particular, for all $0 < \epsilon < 1/16$, if n is large enough so that $c_n \leq \epsilon$ then one can take $\varphi_u(\epsilon) = \Phi^{-1}(3/4) - c_0\epsilon$ and $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - c_0\epsilon$. So that the ratio $\varphi_u(\epsilon)/\varphi_l(\epsilon)$ is constant; in that case, the $MOMAD_K$ and SDO_k are isomorphism (see Proposition 6.2) and we recover a subgaussian rate in Theorem 6.4.

However, a Gaussian approximation result such as the one in (6.25) is not enough for Assumption 6.4. Indeed, it follows from (6.25) that for all $0 \leq r \leq c_0$, $H_v(r) \leq H_G(r) + c_n \leq 1/2 - c_1 r + c_n$ for some absolute constants $c_0 > 0$ and $c_1 > 0$. It appears that our analysis used to prove Theorem 6.4 does not stand this extra error term c_n compare with Assumption 6.4. Gaussian approximation does not help in this case: indeed Assumption 6.4 is more about the existence of a uniform lower bound around 0 of the density functions of the one-dimensional projections $\langle n^{-1/2}\sum_i Z_i, v\rangle$ as we are considering now.

Beyond the Gaussian behavior. In the latter two paragraphs, we identified situations where the $n^{-1/2}\sum_{i=1}^n \langle Z_i, v\rangle$ for $v \in \mathcal{S}_2^{d-1}$ behave like Gaussian variables. We saw that this may be the case because we are considering one-dimensional projections of d-dimensional vectors and/or we are taking empirical means over n variables. But properties we are looking for the $H_{n,v}, v \in \mathcal{S}_2^{d-1}$ functions (see (6.21), (6.22) and (6.23)) are all dealing only with their behavior around 3 (or 4 when the median is not 0) points. So that only the behavior of these functions at these points play a role and there is no need to mimic the Gaussian case for all values of r in $\mathbb{R}$. We now state a general result going in this direction. In particular, we recover the conditions from Maronna and Yohai (1995) and Zuo et al. (2004a) recalled in the Introduction section.

Let us assume that the $n^{-1/2}\sum_{i=1}^n \langle Z_i, v\rangle$ for $v \in \mathcal{S}_2^{d-1}$ are absolutely continuous w.r.t. the Lebesgue measure with a density function denoted by f_v. By the mean value theorem, we have

for all $r \geq 0$, all $p \in (0,1)$ and $\epsilon \geq 0$ such that $p+\epsilon \in (0,1)$,

$$H_v(r) \leq H_v(0) - \min_{0\leq t\leq r} f_v(t)r \text{ and } \frac{\epsilon}{\max_{q\in[p,p+\epsilon]} f_v(W_v(q))} \leq W_v(p) - W_v(p+\epsilon) \leq \frac{\epsilon}{\min_{q\in[p,p+\epsilon]} f_v(W_v(q))}. \tag{6.26}$$

In particular, the values of the density functions $f_v, v \in \mathcal{S}_2^{d-1}$ at $0, W(1/4), W(1/2)$ and $W(3/4)$ drives the quality of inequalities from (6.26) and, as noted in previous works on the Stahel-Donoho outlyingness function, are enough to insure all the conditions we need on H_v and W_v recalled in (6.21), (6.22) and (6.23).

Proposition 6.6. *Let $K \in [N]$ be such that $N/K \in \mathbb{N}$. We assume that the original non-corrupted data $\tilde{X}_i, i \in [N]$ are independent and that there exists $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ definite positive so that for all $v \in \mathcal{S}_2^{d-1}$, $\sqrt{K/N}\sum_{i=1}^{N/K}\langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v\rangle$ are absolutely continuous real valued random variables with a density denoted by f_v.*

If there exists $0 < \epsilon < 1/8$ and $c_0 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$, all $p \in \{W_v(1/4), W_v(1/2), W_v(3/4)\}$ and all $q \in [p-2\epsilon, p+2\epsilon]$, $f_v(q) \geq c_0$ then for $\mathcal{I}_v^{max} = \max\left(W_v(1/4) - W_v(1/2), W_v(1/2) - W_v(3/4)\right)$ and $\mathcal{I}_v^{min} = \min\left(W_v(1/4) - W_v(1/2), W_v(1/2) - W_v(3/4)\right)$ we can take

$$\varphi_u(\epsilon) = \max_{v\in\mathcal{S}_2^{d-1}} \mathcal{I}_v^{max} + 4\epsilon/c_0 \text{ and } \varphi_l(\epsilon) = \min_{v\in\mathcal{S}_2^{d-1}} \mathcal{I}_v^{min} - 4\epsilon/c_0.$$

We also have

$$\frac{\varphi_u(\epsilon)}{\varphi_l(\epsilon)} \leq \frac{\max_{v\in\mathcal{S}_2^{d-1}} \mathcal{I}_v^{max}}{\min_{v\in\mathcal{S}_2^{d-1}} \mathcal{I}_v^{min}} \left(1 + \frac{16\epsilon}{c_0 \min_{v\in\mathcal{S}_2^{d-1}} \mathcal{I}_v^{min}}\right)$$

when $4\epsilon \leq c_0 \min_v \mathcal{I}_v^{min}$. Moreover, if $(c_0/4)\max_v \mathcal{I}_v^{max} < 1/8$ and $\min_v \mathcal{I}_v^{min} \geq c_1$, for some absolute constant $c_1 > 0$, then condition (6.22) holds (and so we recover the optimal subgaussian rates in Theorem 6.2 and Theorem 6.3) and if for all $v \in \mathcal{S}_2^{d-1}$, $\mathcal{I}_v^{max} = \mathcal{I}_v^{min} := \phi_0$ then condition (6.23) holds and so does Proposition 6.5.

If for all $v \in \mathcal{S}_2^{d-1}$, $H_v(0) \leq 1/2$ and there are absolute constants $c_0 > 0$ and $c_1 > 0$ so that for all $0 < r < c_0$, $f_v(v) \geq c_1$ then Assumption 6.4 holds (that is (6.23) holds) and so does Theorem 6.4.

Note that in Proposition 6.6, μ and Σ do not have to be the mean and covariance matrix of the $\tilde{X}_i$'s. In that case, μ and Σ are sometimes called location and scale and so Theorem 6.4 still applies for the robust to adversarial contamination and heavy-tail estimation of location, even in situations where there is not even a first moment.

Proposition 6.6 gives an alternative to Gaussian approximation which does not, in general, allow to check Assumption 6.4 because of the residual terms in Esseen or Berry-Esseen type inequalities. The assumptions in Proposition 6.6 are all granting that the density functions f_v are locally lower bounded around the 'critical' 1/4 and 3/4 quartiles and medians. They are natural assumptions that already appeared in several studies of estimators based on the SDO. In Proposition 6.6 we show that by using the median-of-means principle these assumptions are dealing with the density functions on the bucketed means and not the data themselves. However, Proposition 6.6 may also be applied in the $K = N$ case as for elliptically contoured distributions.

6.6 Conclusion

We showed that it is possible to estimate a mean vector in $\mathbb{R}^d$ w.r.t. the metric $\left\|\Sigma^{-1/2}.\right\|_2$ even though Σ is unknown, the data set is corrupted by an adversary and the data are heavy-tailed.

The rate obtained are the (deviation) minmax one in the ideal i.i.d. Gaussian case. The estimator used to achieve this rate is a deepest point with respect to a median-of-means version of the Stahel-Donoho outlyingness functional. When the data are spherical enough there is no need to bucket the data and then the estimator is using the classical Stahel-Donoho outlyingness. Our analysis shows that the two cases can be handled using the same methodology and that the family of cdfs $(H_{N,K,v} : v \in \mathcal{S}_2^{d-1})$ plays a key role in this analysis, in particular, their behavior around 0, the median and the 1/4 and 3/4 quartiles.

In this work, we have not dealt with several research opportunities opened by the SDO. We now list some of them that may be considered in future works. a) It may look possible to use the isomorphic properties of the $MOMAD_K$ and SDO_K to study the Stahel-Donoho estimator (SDE) or a median-of-means version of the SDE defined as

$$\tilde{\mu}^{SDE}_{MOM,K} = \frac{\sum_{k=1}^K \hat{w}_k \bar{X}_k}{\sum_{k=1}^K \hat{w}_k} \tag{6.27}$$

where $(\hat{w}_k)_{k=1}^K$ are non-negative weights such that $\hat{w}_k$ depends on the outlyingness of the k-th bucketed mean $\bar{X}_k$. For instance,

$$\hat{w}_k = \begin{cases} 1 & \text{if } SDO_K(\bar{X}_k) \leq \hat{\alpha}_K \\ 0 & \text{otherwise.} \end{cases} \quad \text{where } \hat{\alpha}_k = \text{Med}(SDO_K(\bar{X}_k)). \tag{6.28}$$

b) Similarly, the isomorphic or almost-isometry properties of $MOMAD_K$ and SDO_K may also be used to study the properties of a MOM version of the SDE of the covariance matrix:

$$\hat{\Sigma} = \frac{2}{K} \sum_{k=1}^K \hat{w}_k (\bar{X}_k - \tilde{\mu}^{SDE}_{MOM,K})(\bar{X}_k - \tilde{\mu}^{SDE}_{MOM,K})^\top.$$

c) From a computational point of view, it is still an open question to construct an approximate solution to the SDO. The original or MOM version of the Stahel-Donoho median could be approximated via a robust gradient descent algorithm such as the one introduced in Cherapanamjeri et al. (2019); Depersin and Lecué (2019); Lei et al. (2020) with some extra normalization step required by the MAD denominator. We expect this algorithm to be more efficient than the classical weighted SDE because we expect to do only $\log d$ iterations to achieve a subgaussian estimator using a robust gradient descent algorithm whereas the SDE would require to approximate the K depths $SDO_K(\bar{X}_k), k \in [K]$ and should therefore require more computational time (note that, in practice the SDE has been reported to be more efficient than the deepest data that is the data $\bar{X}_k$ with the smallest $SDO_K(\bar{X}_k)$ but the SDE was not compared with an approximate solution of $\hat{\mu}^{SDO}_{MOM,K}$).

6.7 Proofs

In this section, we provide some proofs of all the results from the preceding sections. The only complexity measure we are using in this work is the Vapnik and Chervonenkis (VC) dimension Vapnik and Chervonenkis (2015); Vapnik (2000) of a class $\mathcal{F}$ of Boolean functions, i.e. of functions from $\mathbb{R}^d$ to $\{0,1\}$ in our case, following Chen et al. (2018); Depersin (2020a). We recall that $VC(\mathcal{F})$ is the maximal integer n such that there exists $x_1, \dots, x_n \in \mathbb{R}^d$ for which the set $\{(f(x_1), \cdots, f(x_n)) : f \in \mathcal{F})\}$ is of maximal cardinality that is 2^n. The only VC-dimension we will use is the one of the set of all indicators of half affine spaces in $\mathbb{R}^d$: $VC(\{x \in \mathbb{R}^d \to I(\langle \cdot, v \rangle \geq r) : v \in \mathbb{R}^d, r \in \mathbb{R}\}) = d + 1$ (see Example 2.6.1 in van der Vaart and Wellner (1996)). The main technical tool (see Chapter 3 in Koltchinskii (2011) or Chapters

6.1 and 13.3 in Boucheron et al. (2013)) we will be using is the following one: let $Y_1, \dots, Y_n$ be independent random vectors in $\mathbb{R}^d$, there exists an absolute constant C_0 such that for all $u > 0$, with probability at least $1 - \exp(-u)$,

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(Y_i) - \mathbb{E} f(Y_i) \right) \leq C_0 \sqrt{\frac{V C(\mathcal{F})}{n}} + \sqrt{\frac{2u}{n}}. \tag{6.29}$$

One can for instance take $C_0 = \sqrt{1440\pi/(1 - e^{-1})}$ as in the proof of Lemma 7.3 in Chen et al. (2018).

We recall that for all $v \in \mathcal{S}_2^{d-1}$, $K \in [N]$ and $r > 0$,

$$H_{N,K,v}(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle \geq r \right].$$

The rate of convergence we will obtain is the smallest r^* satisfying

$$C_0 \sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + \sup_{\|v\|_2 = 1} H_{N,K,v}(r^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{6.30}$$

where C_0 is the constant from (6.29) and for some choice of K and u specified in each result depending on the set of assumptions.

6.7.1 Proof of Proposition 6.2 and 6.1 (first part): isomorphic property of MOMAD

We first prove Proposition 6.2 – the proof of Proposition 6.1 is a straighforward application of Proposition 6.2.

Proof of Proposition 6.2. We first observe that by renormalization, it is enough to show that for all $v \in \mathcal{S}_2^{d-1}$,

$$\varphi_l(\epsilon) \leq \mathrm{Med}(|\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle - \mathrm{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v \rangle)|) \leq \varphi_u(\epsilon). \tag{6.31}$$

Moreover, for all $i \in [N], \Sigma^{-1/2}(\tilde{X}_i - \mu)$ has mean zero and covariance I_d. Hence, without loss of generality we assume that $\mu = 0$ and $\Sigma = I_d$.

The strategy we are using to prove (6.31) is the following one. Let K real numbers $a_1, \dots, a_K$ be given and denote by $a_{(1)} \leq \cdots \leq a_{(K)}$ the non-decreasing rearrangement of the $(a_k)_k$ (this is the rearrangement of the a_k's and not of their absolute values). To prove a result like $\varphi_l(\epsilon) \leq \mathrm{Med}(|a_k - \mathrm{Med}(a_k)|) \leq \varphi_u(\epsilon)$, it is enough to show that $\varphi_l(\epsilon) \leq a_{(3(K+1)/4)} - a_{((K+1)/2)} \leq \varphi_u(\epsilon)$ and $\varphi_l(\epsilon) \leq a_{((K+1)/2)} - a_{((K+1)/4)} \leq \varphi_u(\epsilon)$. As a consequence, to prove a result like (6.31), we should study the rearrangement (the two quartiles and the median) of the $\langle \bar{X}_k, v \rangle, k \in [K]$ uniformly over all $v \in \mathcal{S}_2^{d-1}$. But, $|\mathcal{O}|$ elements among the X_i's come from the adversary and we do not have any control on their behavior. We therefore have to consider the worst possible case which is when $|\mathcal{O}|$ bucketed means $\bar{X}_k$ are corrupted by one outlier from $\{X_i : i \in \mathcal{O}\}$. However, one may check that if we change $|\mathcal{O}|$ points in a set $\{a_k : k \in [K]\}$ to get a new set $\{A_k : k \in [K]\}$ then $\varphi_l(\epsilon) \leq a_{(3(K+1)/4)} - a_{((K+1)/2)} \leq \varphi_u(\epsilon)$ will be true if we show that $\varphi_l(\epsilon) \leq A_{(3(K+1)/4 - |\mathcal{O}|)} - A_{((K+1)/2 + |\mathcal{O}|)}$ and $A_{(3(K+1)/4 + |\mathcal{O}|)} - A_{((K+1)/2 - |\mathcal{O}|)} \leq \varphi_u(\epsilon)$ – and a similar observation holds for the other (1/4)-quartile. We will therefore first study the rearrangement of the original (i.e. non corrupted) bucketed means (later denoted by $\overline{\bar{X}}_k, k \in [K]$)

projected on all one dimensional directions uniformly over these directions to deduce the result from (6.31) on the corrupted bucketed means $\bar{X}_k$.

We denote by $\overline{\tilde{X}}_k, k \in [K]$ the bucketed means of the original (non corrupted) dataset, i.e. $\overline{\tilde{X}}_k = (1/|B_k|)\sum_{i\in B_k} \tilde{X}_i$ for $k \in [K]$. To prove (6.31) we first study the rearrangements of vectors $(\langle \overline{\tilde{X}}_k, v\rangle)_{k\in[K]}$ uniformly over all $v \in \mathcal{S}_2^{d-1}$. We will then deal with the adversarial corruption to get (6.31).

We introduce the following supremum of empirical process:

$$Z = \sup_{\ell\in[K-1]} \sup_{\|v\|_2=1} \left| \frac{1}{K}\sum_{k=1}^{K} I\left(\langle \overline{\tilde{X}}_k, v\rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}}\right) - \mathbb{P}\left[\langle \overline{\tilde{X}}_k, v\rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}}\right]\right|$$

where W_v has been defined in Definition 6.1. It follows from (6.29) that for all $u > 0$, with probability at least $1 - \exp(-u)$, $Z \leq C_0\sqrt{(d+1)/K} + \sqrt{2u/K}$ (note that even though the function W_v depends on v, the boolean function $x \to I(\langle x, v\rangle \geq W_v(\ell/N))$ is still the indicator of an affine half-space of $\mathbb{R}^d$ for all $v \in \mathbb{R}^d$ and all $\ell \in [K-1]$ and thus the VC dimension of the set of Boolean functions $\{x \to I(\langle x, v\rangle \geq W_v(\ell/K)) : v \in \mathbb{R}^d, \ell \in [K-1]\}$ is less or equal to $d+1$). As a consequence, for some choice of $0 < \epsilon < 1/8$ such that Assumption 6.3 holds, if $K \geq 4C_0^2(d+1)\epsilon^{-2}$ then with probability at least $1 - \exp(-\epsilon^2 K/8)$, $Z \leq \epsilon$. Let us denote by Ω_ϵ the event onto which $Z \leq \epsilon$; we proved that $\mathbb{P}[\Omega_\epsilon] \geq 1 - \exp(-\epsilon^2 K/8)$.

Let us place ourselves on the event Ω_ϵ up to the end of the proof. Since for all $v \in \mathcal{S}_2^{d-1}$,

$$\mathbb{P}\left[\langle \overline{\tilde{X}}_k, v\rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}}\right] = H_v(W_v(\ell/K)) = \ell/K,$$

(by left continuity of H_v we have $H_v(W_v(p)) = p$ for all $p \in (0,1)$), we have for all $\ell \in [K]$ and $v \in \mathcal{S}_2^{d-1}$, that

$$\left|\left\{k \in [K] : \langle \overline{\tilde{X}}_k, v\rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}}\right\}\right| \in [\ell - \epsilon K, \ell + \epsilon K]\,. \tag{6.32}$$

This last result on the uniform in $v \in \mathcal{S}_2^{d-1}$ rearrangement of $(\langle \overline{\tilde{X}}_k, v\rangle)_k$ will be used to get the desired result on the rearrangement for $(\langle \overline{X}_k, v\rangle)_k$ (uniformly in v). To go from the $\overline{\tilde{X}}_k$'s to the $\overline{X}_k$'s we now have to deal with the adversarial corruption.

Since, there are $|\mathcal{O}|$ original data that may have been modified by the adversary, in the worse case $|\mathcal{O}|$ bucketed means $\overline{\tilde{X}}_k$ may be considered as corrupted and so, from the above cardinality estimation result (6.32), we may only certify (on Ω_ϵ) that

$$\left|\left\{k \in [K] : \langle \bar{X}_k, v\rangle \geq \frac{W_v(\ell/K)}{\sqrt{N/K}}\right\}\right| \in [\ell - \epsilon K - |\mathcal{O}|, \ell + \epsilon K + |\mathcal{O}|] \subset [\ell - 2\epsilon K, \ell + 2\epsilon K]$$

on the K bucketed means $\bar{X}_k$ constructed from the adversarialy corrupted dataset $\{X_i : i \in [N]\}$. We used here the assumption that $|\mathcal{O}| \leq \epsilon K$. If follows from the latter result that if we denote by $q_{K,v}^{1/4}$ the 1/4 quartile of vector $(\langle \bar{X}_k, v\rangle : k \in [K])$, by $q_{K,v}^{1/2}$ its median and by $q_{K,v}^{3/4}$ its 3/4 quartile then,

$$\sqrt{\frac{K}{N}}W_v\left(\frac{3}{4} + 2\epsilon\right) \leq q_{K,v}^{1/4} \leq \sqrt{\frac{K}{N}}W_v\left(\frac{3}{4} - 2\epsilon\right); \sqrt{\frac{K}{N}}W_v\left(\frac{1}{2} + 2\epsilon\right) \leq q_{K,v}^{1/2} \leq \sqrt{\frac{K}{N}}W_v\left(\frac{1}{2} - 2\epsilon\right)$$

and

$$\sqrt{\frac{K}{N}}W_v\left(\frac{1}{4} + 2\epsilon\right) \leq q_{K,v}^{3/4} \leq \sqrt{\frac{K}{N}}W_v\left(\frac{1}{4} - 2\epsilon\right).$$

It follows from these inequalities that on the event Ω_ϵ, we have for all $v \in \mathcal{S}_2^{d_1}$,

$$\mathrm{Med}(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|) \leq \sqrt{\frac{K}{N}} \max\left(W_v\left(\frac{1}{4}-2\epsilon\right) - W_v\left(\frac{1}{2}+2\epsilon\right), W_v\left(\frac{1}{2}-2\epsilon\right) - W_v\left(\frac{3}{4}+2\epsilon\right)\right)$$

and

$$\mathrm{Med}(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|) \geq \sqrt{\frac{K}{N}} \min\left(W_v\left(\frac{1}{4}+2\epsilon\right) - W_v\left(\frac{1}{2}-2\epsilon\right), W_v\left(\frac{1}{2}+2\epsilon\right) - W_v\left(\frac{3}{4}-2\epsilon\right)\right).$$

The result follows from the definition of $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ in Assumption 6.3. ∎

6.7.2 Proof of Proposition 6.3 and 6.1 (second part): isomorphic property of SDO_K.

The proof of Proposition 6.3 and 6.1 (second part) relies on the next result.

Proposition 6.7. *We assume that the adversarial contamination with L_2 inliers model from Assumption 6.1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Let $K \in [N]$, $u > 0$ and r^* be such that (6.30) holds. Then, with probability at least $1 - \exp(-u)$,*

$$\sup_{v \in \mathcal{S}_2^{d-1}} |\mathrm{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)| \leq \sqrt{\frac{K}{N}} r^*.$$

Proof of Proposition 6.7. Denote by $\mathcal{K} = \{k : B_k \cap \mathcal{O} = \emptyset\}$ the set of indices of non-corrupted blocks of data. It follows from (6.29) and the definition of r^* that with probability at least $1 - \exp(-u)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) \\
&= \frac{1}{K}\sum_{k\in\mathcal{K}} I\left(\langle \Sigma^{-1/2}(\overline{\tilde{X}}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) + \frac{1}{K}\sum_{k\in\mathcal{K}^c} I\left(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) \\
&\leq \frac{1}{K}\sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\overline{\tilde{X}}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) + \frac{|\mathcal{O}|}{K} \\
&\leq \sup_{\|v\|_2=1}\left[\frac{1}{K}\sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\overline{\tilde{X}}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) - P\left(\langle \Sigma^{-1/2}(\overline{\tilde{X}}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right)\right] \\
&+ P\left(\langle \Sigma^{-1/2}(\overline{\tilde{X}}_1 - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) + \frac{|\mathcal{O}|}{K} \\
&\leq C_0\sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + H_{N,K,v}(r^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}.
\end{aligned}$$

As a consequence, with probability at least $1 - \exp(-u)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$\sum_{k=1}^K I\left(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle \geq \frac{r^*}{\sqrt{N/K}}\right) < \frac{K}{2}$$

and so

$$\sup_{v \in \mathcal{S}_2^{d-1}} |\mathrm{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)| \leq \sqrt{\frac{K}{N}} r^*. \tag{6.33}$$

Remark 6.2. *It is also possible to consider a "directional version" of Proposition 6.7 if one defines a "directional version" of r^*, that is for all directions $v \in \mathcal{S}_2^{d-1}$, define $r_v^* > 0$ satisfying*

$$C_0\sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + H_{N,K,v}(r_v^*) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}.$$

Then, under the same conditions as in Proposition 6.7, we have with probability at least $1 - \exp(-u)$,

$$\sup_{v\in\mathcal{S}_2^{d-1}} \frac{|\operatorname{Med}(\langle\Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)|}{r_v^*} \leq \sqrt{\frac{K}{N}}.$$

Hence, Proposition 6.7 holds as well for $r^* = \sup_{\|v\|_2=1} r_v^*$. Note that for most of the $v \in \mathcal{S}_2^{d-1}$ the value of r_v^* is expected to be much smaller than r^*. For instance, for vectors v well-spread, we expect them to have a strong "mixing" power (see for instance "super-Gaussian directions" in Klartag (2017) or Klartag and Sodin (2011); Klartag (2009)).

Proof of Proposition 6.3 and 6.1. It follows from Proposition 6.7 and Proposition 6.2 that, with probability at least $1 - \exp(-u) - \exp(-\epsilon^2 K/8)$, for all $v \in \mathcal{S}_2^{d-1}$,

$$|\operatorname{Med}(\langle\Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)| \leq \sqrt{\frac{K}{N}} r^*,$$

where r^* is such that (6.30) holds, and

$$\varphi_l(\epsilon)\sqrt{\frac{K}{N}}\left\|\Sigma^{1/2}v\right\|_2 \leq MOMAD_K(v) \leq \varphi_u(\epsilon)\sqrt{\frac{K}{N}}\left\|\Sigma^{1/2}v\right\|_2.$$

We denote by Ω_0 the event onto which the last two properties hold. On the even Ω_0, for all $\nu \in \mathbb{R}^d$, we have

$$\begin{aligned}
SDO_K(\nu) &= \sup_{v\in\mathbb{R}^d} \frac{|\operatorname{Med}(\langle\bar{X}_k - \nu, v\rangle)|}{MOMAD_K(v)} \leq \sup_{v\in\mathbb{R}^d} \frac{|\operatorname{Med}(\langle\bar{X}_k - \nu, v\rangle)|}{\varphi_l(\epsilon)\sqrt{K/N}\left\|\Sigma^{1/2}v\right\|_2} = \sup_{v\in\mathcal{S}_2^{d-1}} \frac{|\operatorname{Med}(\langle\Sigma^{-1/2}(\bar{X}_k - \nu), v\rangle)|}{\varphi_l(\epsilon)\sqrt{K/N}} \\
&\leq \sup_{v\in\mathbb{R}^d} \frac{|\operatorname{Med}(\langle\Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)| + |\langle\Sigma^{-1/2}(\nu - \mu), v\rangle|}{\varphi_l(\epsilon)\sqrt{K/N}} \leq \sup_{v\in\mathbb{R}^d} \frac{\sqrt{K/N}r^* + |\langle\Sigma^{-1/2}(\nu - \mu), v\rangle|}{\varphi_l(\epsilon)\sqrt{K/N}} \\
&\leq \begin{cases} \frac{3\left\|\Sigma^{-1/2}(\nu-\mu)\right\|_2}{2\varphi_l(\epsilon)\sqrt{K/N}} & \text{if } \left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 \geq 2\sqrt{K/N}r^* \\ 3r^*/\varphi_l(\epsilon) & \text{otherwise} \end{cases}
\end{aligned}$$

and when $\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2 \geq 2\sqrt{K/N}r^*$, we have

$$SDO_K(\nu) \geq \sup_{v\in\mathcal{S}_2^{d-1}} \frac{|\langle\Sigma^{-1/2}(\nu - \mu), v\rangle| - |\operatorname{Med}(\langle\Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)|}{\varphi_u(\epsilon)\sqrt{K/N}} \geq \frac{\left\|\Sigma^{-1/2}(\nu - \mu)\right\|_2}{2\varphi_u(\epsilon)\sqrt{K/N}}.$$

■

6.7.3 Proof of the statistical bounds

Proof of Proposition 6.1. Proposition 6.1 is a corollary of Proposition 6.2 for $K = N$. For this choice of K, there are N blocks, each containing only one data and so $MOMAD_N(v) = MAD(v)$ for all $v \in \mathbb{R}^d$. The only thing that remains to be checked is the validity of Assumption 6.3 in the Gaussian case and the dependency of the $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$ in terms of ϵ.

When the original data $\tilde{X}_i, i \in [N]$ are N i.i.d. Gaussian vectors $G_1, \dots, G_N$ with mean μ and covariance matrix Σ then for all $K \in [N]$, $(1/\sqrt{N/K}) \sum_{i=1}^{N/K} \Sigma^{-1/2}(\tilde{X}_i - \mu)$ is a standard Gaussian vector in $\mathbb{R}^d$. Therefore the $H := H_{N,K,v}$ function from Assumption 6.3 is equal to the function $x \in \mathbb{R} \to 1 - \Phi(x)$ where $\Phi : x \in \mathbb{R} \to \mathbb{P}[g \leq x]$ is the cdf of a standard Gaussian variable $g \sim \mathcal{N}(0,1)$ in $\mathbb{R}$. This holds for all N, K and $v \in \mathcal{S}_2^{d-1}$, that is $H_{N,K,v}$ is independent of N, K and $v \in \mathcal{S}_2^{d-1}$. Since $W := W_{N,K,v}$ is the generalized inverse of H, in the Gaussian case, we obtain that $W(p) = \Phi^{-1}(1-p)$ for all $p \in (0,1)$. It follows from Lemma 5.2 in Petrov (1995) that there exists some absolute constant $C_1 > 0$ such that

$$\min\left(W\left(\frac{1}{4}+2\epsilon\right) - W\left(\frac{1}{2}-2\epsilon\right), W\left(\frac{1}{2}+2\epsilon\right) - W\left(\frac{3}{4}-2\epsilon\right)\right) \geq \Phi^{-1}(3/4) - C_1\epsilon := \varphi_l(\epsilon)$$

and

$$\max\left(W\left(\frac{1}{4}-2\epsilon\right) - W\left(\frac{1}{2}+2\epsilon\right), W\left(\frac{1}{2}-2\epsilon\right) - W\left(\frac{3}{4}+2\epsilon\right)\right) \leq \Phi^{-1}(1/4) \leq \Phi^{-1}(3/4)+C_1\epsilon := \varphi_u(\epsilon).$$

As a consequence, Assumption 6.3 holds in the Gaussian case for all $0 < \epsilon < \Phi^{-1}(3/4)/C_1$ with $\varphi_l(\epsilon) = \Phi^{-1}(3/4) - C_1\epsilon$ and $\varphi_u(\epsilon) = \Phi^{-1}(3/4) + C_1\epsilon$. ∎

Proofs of theorems 6.1, 6.2 and 6.4 Theorems 6.1, 6.2 and 6.4 are corollaries of a general result that we are stating now.

Theorem 6.5. *There are absolute constants c_0, c_1 and c_2 such that the following holds. We assume that Assumption 6.3 holds for some $0 < \epsilon < 1/8$ and constants $\varphi_l(\epsilon)$ and $\varphi_u(\epsilon)$. We assume that the adversarial contamination with L_2 inliers model from Assumption 6.1 holds with a number of adversarial outliers denoted by $|\mathcal{O}|$. Let $K \geq \max(\epsilon^{-1}|\mathcal{O}|, 4C_0\epsilon^{-2}(d+1))$, $0 < u < \epsilon^2 K/8$ and r^* be such that (6.30) holds (where C_0 is the absolute constant defined in (6.29)). Then, with probability at least $1 - 2\exp(-u)$,*

$$\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO}_{MOM,K} - \mu)\right\|_2 \leq \frac{2\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{K}{N}}r^*.$$

Proof of Theorem 6.1. We have for all $0 \leq r \leq 1, \mathbb{P}[g \geq r] \leq 1/2 - \phi(1)r$ where $g \sim \mathcal{N}(0,1)$. Moreover, for all $K \in [N], v \in \mathcal{S}_2^{d-1}$ and $r > 0$, we have $H_{N,K,v}(r) = \mathbb{P}[g \geq r]$. As a consequence, (6.30) holds if one choose r^*, u and K such that

$$\phi(1)r^* = C_0\sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + \frac{|\mathcal{O}|}{K}$$

as long as for such choice $r^* \leq 1$ (which indeed holds under the assumptions of Theorem 6.1). Finally, we apply Theorem 6.5 for $\epsilon = \kappa_0$, $K = N$ and the bound on the ratio $\varphi_u(\epsilon)/\varphi_l(\epsilon)$ in the Gaussian case from Section 6.5. The result follows since $\hat{\mu}^{SDO}_{MOM,N} = \hat{\mu}^{SDO}$. ∎

Proof of Theorem 6.2. It follows from Markov's inequality (6.15) that (6.30) holds when we take u, r^* and K such that

$$C_0\sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + \frac{1}{1+(r^*)^2} + \frac{|\mathcal{O}|}{K} < \frac{1}{2}.$$

The latter holds, for instance, when $r^* = 3$, $K \geq 4|\mathcal{O}|$, $K > 100C_0^2(d+1)$ and $u \leq K/800$. Note however, that because r^* is constant, the convergence rate is proportional to $\sqrt{K/N}$, in particular it does not depend on u. Hence there is no interest to consider values of u smaller than K (up to constant). We therefore apply Theorem 6.5 for this choice of K, $u = \epsilon^2 K/15$ and $r^* = 3$. ∎

Proof of Theorem 6.4. Thanks to Assumption 6.4, there exist absolute constants $c_0, c_1 > 0$ and $c_2 > 0$ such that for all $v \in \mathcal{S}_2^{d-1}$ and $(2C_0/c_1)\sqrt{(d+1)/K} \leq r \leq c_0, H_{N,K,v} \leq 1/2 - c_2 r$. As a consequence, (6.30) holds if one can choose r^*, u and K such that

$$r^* = \frac{2}{c_2}\left(C_0\sqrt{\frac{d+1}{K}} + \sqrt{\frac{2u}{K}} + \frac{|\mathcal{O}|}{K}\right)$$

as long as this latter quantity is less or equal to c_0. Finally, we apply Theorem 6.5 for this choice of K, u and r^*. ∎

Proof of Theorem 6.5. We first note that a proof of Theorem 6.5 may follow from the isomorphic property of SDO_K from Proposition 6.3. However, it is possible to improve constants by using the following strategy.

Let us place ourselves on the intersection of the two events where the results of both Proposition 6.2 and Proposition 6.7 hold. We set $f : v \in \mathbb{R}^d \to \mathrm{Med}(\langle \bar{X}_k, v\rangle)$. Since f is symmetric we have

$$\begin{aligned}
&\left\|\Sigma^{-1/2}(\hat{\mu}^{SDO}_{MOM,K} - \mu)\right\|_2 = \sup_{\|v\|_2=1} \langle \Sigma^{-1/2}(\hat{\mu}^{SDO}_{MOM,K} - \mu), v\rangle = \sup_{v\in\mathbb{R}^d} \langle \hat{\mu}^{SDO}_{MOM,K} - \mu, \frac{v}{\|\Sigma^{1/2}v\|_2}\rangle \\
&= \sup_{v\in\mathbb{R}^d} \frac{\langle \hat{\mu}^{SDO}_{MOM,K}, v\rangle - f(v) + f(v) - \langle \mu, v\rangle}{MOMAD_K(v)} \frac{MOMAD_K(v)}{\|\Sigma^{1/2}v\|_2} \\
&\leq \left(\sup_{v\in\mathbb{R}^d} \frac{\langle \hat{\mu}^{SDO}_{MOM,K}, v\rangle - f(v)}{MOMAD_K(v)} + \sup_{v\in\mathbb{R}^d} \frac{f(v) - \langle \mu, v\rangle}{MOMAD_K(v)}\right) \sup_{v\in\mathbb{R}^d} \frac{MOMAD_K(v)}{\|\Sigma^{1/2}v\|_2} \\
&\leq \left(SDO_K(\hat{\mu}^{SDO}_{MOM,K}) + SDO_K(\mu)\right) \sup_{v\in\mathbb{R}^d} \frac{MOMAD_K(v)}{\|\Sigma^{1/2}v\|_2} \leq 2SDO_K(\mu) \sup_{v\in\mathbb{R}^d} \frac{MOMAD_K(v)}{\|\Sigma^{1/2}v\|_2}.
\end{aligned}$$

where we used that $SDO_K(\hat{\mu}^{SDO}_{MOM,K}) \leq SDO_K(\mu)$ by definition of $\hat{\mu}^{SDO}_{MOM,K}$.

We know how to control $\sup_{v\in\mathbb{R}^d} MOMAD_K(v)/\left\|\Sigma^{1/2}v\right\|_2$ by $\sqrt{K/N}\varphi_u(\epsilon)$ using Proposition 6.2. It remains to control the term $SDO_K(\mu)$. We have

$$\begin{aligned}
SDO_K(\mu) &= \sup_{v\in\mathbb{R}^d} \frac{|\langle \mu, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|}{\mathrm{Med}(|\langle \bar{X}_k, v\rangle - \mathrm{Med}(\langle \bar{X}_k, v\rangle)|)} = \sup_{v\in\mathbb{R}^d} \frac{|\,\mathrm{Med}(\langle \mu - \bar{X}_k, v\rangle)|}{\|\Sigma^{1/2}v\|_2} \frac{\left\|\Sigma^{1/2}v\right\|_2}{MOMAD_K(v)} \\
&\leq \sup_{\|v\|_2=1} |\,\mathrm{Med}(\langle \Sigma^{-1/2}(\bar{X}_k - \mu), v\rangle)| \sup_{v\in\mathbb{R}^d} \frac{\left\|\Sigma^{1/2}v\right\|_2}{MOMAD_K(v)}.
\end{aligned}$$

The term $\sup_{v\in\mathbb{R}^d}\left\|\Sigma^{1/2}v\right\|_2/MOMAD_K(v)$ is smaller than $\sqrt{N/K}/\varphi_l(\epsilon)$ thanks to Proposition 6.2. Finally, to finish the proof, we upper bound the term $\sup_{\|v\|_2=1}|\operatorname{Med}(\langle\Sigma^{-1/2}(\bar{X}_k-\mu),v\rangle)|$ by $\sqrt{K/N}r^*$ thanks to Proposition 6.7. ∎

Proof of Theorem 6.3 For all $k\in[N]$, we set $\hat{\mu}_k=\hat{\mu}^{SDO}_{MOM,k}$ we denote by Ω_k the event onto which

$$\left\|\Sigma^{-1/2}(\hat{\mu}_k-\mu)\right\|_2\leq\frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{k}{N}}$$

and, for all $\nu\in\mathbb{R}^d$, if $\left\|\Sigma^{-1/2}(\nu-\mu)\right\|_2\geq 6\sqrt{k/N}$ then

$$\frac{\left\|\Sigma^{-1/2}(\nu-\mu)\right\|_2}{2\varphi_u(\epsilon)\sqrt{k/N}}\leq SDO_k(\nu)\leq\frac{3\left\|\Sigma^{-1/2}(\nu-\mu)\right\|_2}{2\varphi_l(\epsilon)\sqrt{k/N}}$$

and if $\left\|\Sigma^{-1/2}(\nu-\mu)\right\|_2\leq 6\sqrt{k/N}$ then

$$SDO_k(\nu)\leq\frac{9}{\varphi_l(\epsilon)}.$$

It follows from Proposition 6.3 for $r^*=3$ and $u=K/(16C_0^2)$ and Theorem 6.2 that $\mathbb{P}[\Omega_k]\geq 1-3\exp(-c_1\epsilon^2k)$ when $k\geq\max(|\mathcal{O}|/\epsilon,c_0d/\epsilon^2)$.

Let $K\geq\max(|\mathcal{O}|/\epsilon,c_0d/\epsilon^2)$. On the event $\cap_{k=K}^N\Omega_k$, we have for all $K\leq k\leq N$,

$$SDO_k(\hat{\mu}_K-\hat{\mu}_k)\leq\max\left(\frac{9}{\varphi_l(\epsilon)},\frac{3\left\|\Sigma^{-1/2}(\hat{\mu}_K-\hat{\mu}_k)\right\|_2}{2\varphi_l(\epsilon)\sqrt{k/N}}\right)\leq\max\left(\frac{9}{\varphi_l(\epsilon)},\frac{6\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)}\left(1+\sqrt{\frac{K}{k}}\right)\right)$$

and so, by definition of $\hat{K}$, we have $\hat{K}\leq K$. We also have by definition of $\hat{K}$ and because $\hat{K}\leq K$ that

$$SDO_K(\hat{\mu}_{\hat{K}}-\hat{\mu}_K)\leq\max\left(\frac{9}{\varphi_l(\epsilon)},\frac{6\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)}\left(1+\sqrt{\frac{\hat{K}}{K}}\right)\right)\leq\frac{12\varphi_u(\epsilon)}{\varphi_l^2(\epsilon)}.$$

We conclude that either $\left\|\Sigma^{-1/2}(\hat{\mu}_{\hat{K}}-\hat{\mu}_K)\right\|_2\leq 6\sqrt{K/N}$ and so

$$\left\|\Sigma^{-1/2}(\hat{\mu}_{\hat{K}}-\mu)\right\|_2\leq\left\|\Sigma^{-1/2}(\hat{\mu}_{\hat{K}}-\hat{\mu}_K)\right\|_2+\left\|\Sigma^{-1/2}(\mu-\hat{\mu}_K)\right\|_2\leq\left(6+\frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\right)\sqrt{\frac{K}{N}}$$

or $\left\|\Sigma^{-1/2}(\hat{\mu}_{\hat{K}}-\hat{\mu}_K)\right\|_2\geq 6\sqrt{K/N}$ and so

$$\begin{aligned}\left\|\Sigma^{-1/2}(\hat{\mu}_{\hat{K}}-\mu)\right\|_2&\leq\left\|\Sigma^{-1/2}(\hat{\mu}_{\hat{K}}-\hat{\mu}_K)\right\|_2+\left\|\Sigma^{-1/2}(\hat{\mu}_K-\mu)\right\|_2\\&\leq SDO_K(\hat{\mu}_{\hat{K}}-\hat{\mu}_K)2\varphi_u(\epsilon)\sqrt{\frac{K}{N}}+\frac{4\varphi_u(\epsilon)}{\varphi_l(\epsilon)}\sqrt{\frac{K}{N}}\leq\frac{28\varphi_u^2(\epsilon)}{\varphi_l^2(\epsilon)}\sqrt{\frac{K}{N}}.\end{aligned}$$

∎

Proof of Proposition 6.4. Let us place ourselves on the event where (6.19) holds. We therefore have for all $v \in \mathbb{R}^d$,

$$\left| MOMAD_K(v) - \phi_0 \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \right| \leq c_1 \epsilon \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 ,$$

and so, by definition of $\check{\Sigma}$, we have for all $v \in \mathbb{R}^d$,

$$\left| MOMAD_K(v) - \sqrt{\frac{K}{N}} \left\| \check{\Sigma}^{1/2} v \right\|_2 \right| \leq \frac{c_1 \epsilon}{\phi_0} \sqrt{\frac{K}{N}} \left\| \check{\Sigma}^{1/2} v \right\|_2 .$$

In particular, it follows from the two isomorphic results above that for all $v \in \mathbb{R}^d$,

$$\left(1 - \frac{c_1 \epsilon}{\phi_0}\right) \sqrt{\frac{K}{N}} \left\| \check{\Sigma}^{1/2} v \right\|_2 \leq MOMAD_K(v) \leq (\phi_0 + c_1 \epsilon) \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \tag{6.34}$$

It follows from the two isomorphic results above and (6.34) that for all $v \in \mathbb{R}^d$,

$$\left| \left\| \check{\Sigma}^{1/2} v \right\|_2 - \phi_0 \left\| \Sigma^{1/2} v \right\|_2 \right| = \left| \left\| \check{\Sigma}^{1/2} v \right\|_2 - \sqrt{\frac{N}{K}} MOMAD_K(v) + \sqrt{\frac{N}{K}} MOMAD_K(v) - \phi_0 \left\| \Sigma^{1/2} v \right\|_2 \right|$$
$$\leq \frac{c_1 \epsilon}{\phi_0} \left\| \check{\Sigma}^{1/2} v \right\|_2 + c_1 \epsilon \left\| \Sigma^{1/2} v \right\|_2 \leq \kappa_1 \left\| \Sigma^{1/2} v \right\|_2$$

as long as $c_1 \epsilon < \phi_0$ and for $\kappa_1 := (2 c_1 \epsilon \phi_0)/(\phi_0 - c_1 \epsilon)$.

We deduce that for all $v \in \mathbb{R}^d$, $\left\| \check{\Sigma}^{1/2} \Sigma^{-1/2} v \right\|_2$ lies in $[\phi_0 - \kappa_1, \phi_0 + \kappa_1] \|v\|_2$, so that all the eigenvalues of $\Sigma^{-1/2} \check{\Sigma} \Sigma^{-1/2}$ are in $[(\phi_0 - \kappa_1)^2, (\phi_0 + \kappa_1)^2]$ as long as $\kappa_1 < \phi_0$. Finally, as long as $4 c_1 \epsilon < \phi_0$, we get

$$\left\| \Sigma^{-1/2} \check{\Sigma} \Sigma^{-1/2} - \phi_0^2 \, \mathrm{Id} \right\|_{op} < 3 \phi_0 \kappa_1 < 12 \phi_0 c_1 \epsilon .$$

Proof of Proposition 6.5. We have for all $i, j \in [d]$, $\left| \phi_0^2 \Sigma_{ij} - \hat{\Sigma}_{ij} \right| \leq c_1 \epsilon (c_1 \epsilon + \phi_0) (\Sigma_{ii} + \Sigma_{ij})$ because, it follows from (6.19) that for all $v \in \mathbb{R}^d$,

$$\left| MOMAD_K^2(v) - \phi_0^2 \frac{K}{N} \left\| \Sigma^{1/2} v \right\|_2^2 \right| = \left| MOMAD_K(v) - \phi_0 \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \right| \left(MOMAD_K(v) + \phi_0 \sqrt{\frac{K}{N}} \left\| \Sigma^{1/2} v \right\|_2 \right)$$
$$\leq c_1 \epsilon \frac{K}{N} \left\| \Sigma^{1/2} v \right\|_2^2 (c_1 \epsilon + \phi_0).$$

Next, we have for all $u, v \in \mathbb{R}^d$ such that $\|u\|_1 = \|v\|_1 = 1$

$$|\langle u, (\phi_0^2 \Sigma - \hat{\Sigma}) v \rangle|$$
$$= \frac{N}{4K} \left| \sum_{i,j} u_i v_j \left(\phi_0^2 \frac{K}{N} \left\| \Sigma^{1/2}(e_i + e_j) \right\|_2^2 - MOMAD_K^2(e_i + e_j) + \phi_0^2 \frac{K}{N} \left\| \Sigma^{1/2}(e_i - e_j) \right\|_2^2 - MOMAD_K^2(e_i - e_j) \right) \right|$$
$$\leq \frac{c_1 \epsilon (c_1 \epsilon + \phi_0)}{4} \sum_{i,j} |u_i| |v_j| \left(\left\| \Sigma^{1/2}(e_i + e_j) \right\|_2^2 + \left\| \Sigma^{1/2}(e_i - e_j) \right\|_2^2 \right) = \frac{c_1 \epsilon (c_1 \epsilon + \phi_0)}{2} \sum_{i,j} |u_i| |v_j| \left(\Sigma_{ii} + \Sigma_{jj} \right).$$

■

CHAPTER 7

Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms

Contents

7.1 Introduction

We consider the problem of robust (to adversarial corruption and heavy-tailed data) multivariate mean and location estimation with respect to any pseudo-norm $\nu \in \mathbb{R}^d \to \|\nu\|_S = \sup_{\mu \in S} \langle \mu, \nu \rangle$ where S is any symmetric subset of $\mathbb{R}^d$ (i.e. if $x \in S$ then $-x \in S$). Only little is known for general symmetric sets S and we will mainly refer to Lugosi and Mendelson (2019b) where this problem has been handled for S which is the unit dual ball B° of a norm $\|\cdot\|$ (so that $\|\cdot\|_S = \|\cdot\|$).

In Lugosi and Mendelson (2019b), the authors introduced the problem of robust to heavy-tailed data estimation of a mean vector w.r.t. any norm. The problem can be stated as follow: given N i.i.d. random vectors $X_1, \ldots, X_N$ in $\mathbb{R}^d$ with mean μ^* and covariance matrix Σ, a norm $\|\cdot\|$ on $\mathbb{R}^d$ and a confidence parameter $\delta \in (0,1)$ find an estimator $\tilde{\mu}_N(\delta)$ and the best possible accuracy $r^*(N,\delta)$ such that with probability at least $1-\delta$, $\|\tilde{\mu}_N(\delta) - \mu^*\| \leq r^*(N,\delta)$. In Lugosi and Mendelson (2019b), the authors use the median-of-means principle Nemirovsky and Yudin (1983); Jerrum et al. (1986); Alon et al. (1999) to construct an estimator satisfying the following result.

Theorem 7.1. *[Theorem 2 in Lugosi and Mendelson (2019b)] There exist an absolute constant c such that the following holds. Given a norm $\|\cdot\|$ on $\mathbb{R}^d$ and a confidence $\delta \in (0,1)$, one can*

construct $\tilde{\mu}_N(\delta)$ *such that with probability at least* $1-\delta$

$$\|\tilde{\mu}_N(\delta) - \mu^*\| \leq \frac{c}{\sqrt{N}}\left(\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N \epsilon_i(X_i-\mu^*)\right\| + \mathbb{E}\left\|\Sigma^{1/2}G\right\| + \sup_{v\in B^\circ}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}\right)$$

where B° *is the unit dual ball associated with* $\|\cdot\|$, (ϵ_i) *are i.i.d. Rademacher variables independent of the* X_i*'s and* $G \sim \mathcal{N}(0, I_d)$.

The construction of $\tilde{\mu}_N(\delta)$ is pretty involved and it seems hard to design an algorithm out of this procedure. In particular, $\tilde{\mu}_N(\delta)$ has not been proved to be solution to a convex optimization problem. Theorem 7.1's main interest is thus from a theoretical point of view, while robust multivariate mean estimation can also be interesting from a practical point of view Diakonikolas et al. (2017).

The rate obtained in Theorem 7.1 can be decomposed into two terms: a deviation term

$$\sup_{v\in B^\circ}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}$$

where $\sup_{v\in B^\circ}\left\|\Sigma^{1/2}v\right\|_2$ is a weak variance term and a complexity term which is the sum of a Rademacher complexity $\mathbb{E}\left\|N^{-1/2}\sum_{i=1}^N \epsilon_i(X_i-\mu^*)\right\|$ and a Gaussian mean width $\mathbb{E}\left\|\Sigma^{1/2}G\right\|$. The intuition behind this rate is explained in Lugosi and Mendelson (2019b), in particular, in Question 1. We will however show that this rate is not the right one and that the Gaussian mean width term is actually not necessary. Moreover, we will show that the improved rate can be achieved by an estimator solution to a convex optimization problem in Section 7.3 and that this holds even in the adversarial corruption model (see Assumption 7.1 in Section 7.3 below for a formal definition) and even in some situations where there is not even a first moment; in that case, μ^* is a *location* parameter and Σ a *scatter* parameter.

The optimality of the rate in Theorem 7.1 has been raised in Lugosi and Mendelson (2019b). The classical approach to answer this type of question is to consider the Gaussian case that is when the data $X_i, i \in [N]$ are i.i.d. $\mathcal{N}(\mu^*, \Sigma)$. This is also the strategy used in Lugosi and Mendelson (2019b) to obtain the following deviation-minimax lower bound result[1].

Theorem 7.2. *[Theorem 3 and first paragraph in p.962 in Lugosi and Mendelson (2019b)] There exists an absolute constant* $c > 0$ *such that the following holds. If* $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ *is an estimator such that for all* $\mu^* \in \mathbb{R}^d$ *and all* $\delta \in (0, 1/4)$,

$$\mathbb{P}_{\mu^*}^N\left[\|\hat{\mu} - \mu^*\| \leq r^*\right] \geq 1-\delta$$

where $\mathbb{P}_{\mu^*}^N$ *is the probability distribution of* $(X_i)_{i\in[N]}$ *when the* X_i *are i.i.d.* $\mathcal{N}(\mu^*, \Sigma)$ *then*

$$r^* \geq \frac{c}{\sqrt{N}}\left(\sup_{\eta>0}\eta\sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)} + \sup_{v\in B^\circ}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}\right)$$

where $N(\Sigma^{1/2}B^\circ, \eta B_2^d)$ *is the minimal number of translated of* ηB_2^d *needed to cover* $\Sigma^{1/2}B^\circ$.

The term $\sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}$ in the lower bound from Theorem 7.2 is obtained in Lugosi and Mendelson (2019b) from Proposition 6.1 in Catoni (2012) which is a deviation-

[1] the result from Lugosi and Mendelson (2019b) is proved for $\Sigma = I_d$, it is however straightforward to extend it to the general case.

minimax lower bound result holding in the one dimensional case which relies on the fact that the empirical mean is a sufficient statistics in the Gaussian shift theorem[2].

The complexity term $\sup_{\eta>0} \eta\sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)}$ obtained in Theorem 7.2 follows from the duality theorem of metric entropy from Artstein et al. (2004) and a volumetric argument in the Gauss space similar to the one used to prove dual Sudakov's inequality in p.82-83 in Ledoux and Talagrand (2011) which has also been used to obtain minimax lower bounds based on the entropy in Lecué and Mendelson (2013) and Mendelson (2017b).

In general, there is a gap between the upper bound from Theorem 7.1 and the lower bound from Theorem 7.2 even in the Gaussian case. This gap is characterized by Sudakov's inequality (see Theorem 3.18 in Ledoux and Talagrand (2011) or Theorem 5.6 in Pisier (1989)):

$$\sup_{\eta>0} \eta\sqrt{\log N(\Sigma^{1/2}B^\circ, \eta B_2^d)} \leq c\mathbb{E}\left\|\Sigma^{1/2}G\right\| \tag{7.1}$$

where $G \sim \mathcal{N}(0, I_d)$. Indeed, in the Gaussian case the complexity term of the rate obtained in Theorem 7.1 is the Gaussian mean width, that is the right-hand term from (7.1) whereas the complexity term from Theorem 7.2 is the entropy, that is the left-hand term in (7.1).

As mentioned in Remark 3 from Lugosi and Mendelson (2019b), when Sudakov's inequality (7.1) is sharp then upper and lower bounds from Theorem 7.1 and 7.2 match in the Gaussian case (in that case the Rademacher complexity is equal to the Gaussian mean width in Theorem 7.1). Sharpness in Sudakov's inequality is however not a typical situation. In particular, for ellipsoids, Sudakov's bound (7.1) is not sharp in general and therefore the lower bound from Theorem 7.2 fails to recover the classical subgaussian rate for the standard Euclidean norm case (that is for $S = B_2^d$) which is given in Lugosi and Mendelson (2019c) by

$$\sqrt{\frac{\mathrm{Tr}\,(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op}\log(1/\delta)}{N}}. \tag{7.2}$$

Indeed, when $\|\cdot\|$ is the ℓ_2^d Euclidean norm then $\mathbb{E}\left\|\Sigma^{1/2}G\right\| = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \sim \sqrt{\mathrm{Tr}(\Sigma)}$ (see, for instance, Proposition 2.5.1 in Talagrand (2014)). Whereas, for the entropy of $\Sigma^{1/2}B^\circ = \Sigma^{1/2}B_2^d$ w.r.t. ηB_2^d, it follows from equation (5.45) in Pisier (1989) that

$$\sup_{\eta>0} \eta\sqrt{\log_2 N(\Sigma^{1/2}B_2^d, \eta B_2^d)} = \sup_{n\geq 1} e_{n+1}(\Sigma^{1/2})\sqrt{n+1} \sim \sqrt{\sup_{k\in[d]} k\left|\prod_{j=1}^k \lambda_j\right|^{1/k}} \tag{7.3}$$

where $(e_{n+1}(\Sigma^{1/2}))_n$ are the entropy numbers of $\Sigma^{1/2} : \ell_2^d \to \ell_2^d$ (see page 62 in Pisier (1989) for a definition) and $\lambda_1 \geq \ldots \geq \lambda_d$ are the singular values of Σ. In particular, when $\lambda_j = 1/j$, the entropy bound (7.3) is of the order of a constant whereas the Gaussian mean width is of the order of $\sqrt{\log d}$. We will fill this gap in Section 7.2 by showing a lower bound where the entropy is replaced by the (larger) Gaussian mean width. We will therefore obtain matching upper and

[2]The argument used in Lugosi and Mendelson (2019b) goes from the one dimensional case studied in Catoni (2012) to the d-dimensional case. It is given in a none formal way and may require some extra argument to hold. Indeed the estimator $x^*(\hat{\Psi}_N)$ in Lugosi and Mendelson (2019b) is constructed using the d-dimensional data $X_1, \ldots, X_N$ and not one-dimensional data such as $x^*(X_1), \ldots, x^*(X_N)$. However, the result from Catoni (2012) holds for estimators of a one dimensional mean using one-dimensional data and not d-dimensional ones. Nevertheless, Olivier Catoni showed us how to adapt the proof of Proposition 6.1 in Catoni (2012) by using the sufficiency of the empirical mean in the Gaussian shift model in $\mathbb{R}^d$ to get this deviation dependent lower bound term.

lower bounds revealing that Gaussian mean width is the right way to measure the statistical complexity for the mean estimation problem w.r.t. any $\|\cdot\|_S$.

The chapter is organized as follows. In the next section, we obtain the deviation-minimax optimal rate in the i.i.d. Gaussian case. In Section 7.3 we show that the rate from Theorem 7.1 can be improved and that it can be achieved by a solution to a convex program in the adversarial contamination model and in under weak or no moment assumptions. All the proofs have been gathered in Section 7.4.

7.2 Deviation minimax rates in the Gaussian case: benchmark subgaussian rates for the mean estimation w.r.t. $\|\cdot\|_S$

In this section, we obtain the optimal deviation-minimax rates of estimation of a mean vector μ^* when we are given N i.i.d. $X_1, \dots, X_N$ distributed like $\mathcal{N}(\mu^*, \Sigma)$ when $\Sigma \succeq 0$ is some unknown covariance matrix. In the following, $\mathbb{P}_{\mu^*}^N$ denotes the probability distribution of $(X_1, \dots, X_N)$; it is a Gaussian measure on $\mathbb{R}^{Nd}$ with mean $((\mu^*)^\top, \dots, (\mu^*)^\top)$ and a block $(Nd) \times (Nd)$ covariance matrix with $d \times d$ diagonal blocks given by Σ repeated N times and 0 outside of these blocks.

Unlike classical minimax results holding in expectation or with constant probability (see Chapter 2 in Tsybakov (2009)) we want, in this section, the deviation parameter δ to appear explicitly in the minimax lower bound. Moreover, this dependency of the convergence rate with respect to δ should be of the right order given by the subgaussian $\sqrt{\log(1/\delta)}$ rate and not other polynomial dependency such as $\sqrt{1/\delta}$ as one gets for the empirical mean for L_2 variables (see Proposition 6.2 in Catoni (2012)). This subtle behavior of the rate in terms of δ cannot be seen in expectation or constant deviation minimax lower bounds. In particular, this makes such results (like Theorem 7.3 or 7.4 below) unachievable via classical information theoretic arguments as in Chapter 2 in Tsybakov (2009).

Fortunately, in Lecué and Mendelson (2013), a minimax lower bound has been proved thanks to the Gaussian shift theorem which makes the deviation parameter δ appearing explicitly in the minimax lower bound. We use the same strategy here to prove our main result Theorem 7.3 below and its corollary Theorem 7.4 in the classical Euclidean $S = B_2^d$ case.

We consider the general problem of estimating μ^* w.r.t. $\|\cdot\|_S$. Let $S \subset \mathbb{R}^d$ be a symmetric set. We first obtain an upper bound result revealing the subgaussian rate. We use the empirical mean $\bar{X}_N = N^{-1} \sum_i X_i$ as an estimator of μ^*. Using Borell TIS's inequality (Theorem 7.1 in Ledoux (2001) or pages 56-57 in Talagrand (2014)) we get: for all $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\left\|\bar{X}_N - \mu\right\|_S = \sup_{v \in S} \langle v, \bar{X}_N - \mu \rangle \leq \mathbb{E} \sup_{v \in S} \langle v, \bar{X}_N - \mu \rangle + \sigma_S \sqrt{2 \log(1/\delta)}$$

where $\sigma_S = \sup_{v \in S} \sqrt{\mathbb{E} \langle v, \bar{X}_N - \mu \rangle^2}$ is called the weak variance. It follows that with probability at least $1 - \delta$,

$$\left\|\bar{X}_N - \mu\right\|_S \leq \frac{\ell^*(\Sigma^{1/2} S)}{\sqrt{N}} + \frac{\sup_{v \in S} \left\|\Sigma^{1/2} v\right\|_2 \sqrt{\log(1/\delta)}}{\sqrt{N}} \tag{7.4}$$

where $\ell^*(\Sigma^{1/2} S) = \sup \left(\langle G, x \rangle : x \in \Sigma^{1/2} S \right) = \mathbb{E} \left\|\Sigma^{1/2} G\right\|_S$, for $G \sim \mathcal{N}(0, I_d)$, is the Gaussian mean width of the set $\Sigma^{1/2} S$. In particular, in the case where $S = B_2^d$, we recover the subgaussian rate (7.2) in (7.4). Our aim is now to show that the rate in (7.4) is deviation-minimax optimal. This is what is obtained in the next result.

Theorem 7.3. *Let S be a symmetric subset of $\mathbb{R}^d$ such that $\mathrm{span}(S) = \mathbb{R}^d$. If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4]$,*

$$\mathbb{P}_{\mu^*}^N \left[\|\hat{\mu} - \mu^*\|_S \leq r^* \right] \geq 1 - \delta$$

then

$$r^* \geq \max\left(\frac{1}{24}\sqrt{\frac{\log 2}{\log(5/4)}}\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \frac{\sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2}{12}\sqrt{\frac{\log(1/\delta)}{\sqrt{N}}} \right).$$

It follows from the upper bound (7.4) and the deviation-minimax lower bound from Theorem 7.3 that it is now possible to know exactly (up to absolute constants) the subgaussian rate for the problem of mean estimation in $\mathbb{R}^d$ w.r.t. $\|\cdot\|_S$, it is given by

$$\max\left(\frac{\ell^*(\Sigma^{1/2}S)}{\sqrt{N}}, \frac{\sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\log(1/\delta)}}{\sqrt{N}} \right). \tag{7.5}$$

We may identify the two complexity and deviation terms in this rate. In particular, the complexity term is measured here via the Gaussian mean width of the set $\Sigma^{1/2}S$ and not its entropy as it was previously known following Theorem 7.2. Theorem 7.3 together with (7.4) show that the right way to measure the statistical complexity in the problem of mean estimation in $\mathbb{R}^d$ w.r.t. to any $\|\cdot\|_S$ is via the Gaussian mean width. This differs from other statistical problems such as the regression model with random design where the entropy has been proved to be the right statistical complexity in several examples Mendelson (2017b); Lecué and Mendelson (2013). Following the later results in the regression model, Theorem 7.3 is a bit unexpected because one may though that by taking an ERM over an epsilon net of $\mathbb{R}^d$ for the right choice of ϵ one could obtain a better rate than the one driven by the Gaussian mean width in (7.5); indeed, for this type of procedure, one may expect a rate depending on the (smaller) entropy instead of the (larger) Gaussian mean width. Theorem 7.3 shows that this is not the case: even discretized ERM cannot achieve a better rate than the one driven by the Gaussian mean width in the mean estimation problem.

An important consequence of Theorem 7.3 is obtained when $S = B_2^d$ that is for the problem of multivariate mean estimation w.r.t. the ℓ_2^d-norm which is the problem that has been extensively considered during the last decade. In the following result, we recover the well-known subgaussian rate (7.2) showing that all the upper bound results where this rate has been proved to be achieved are actually deviation-minimax optimal and therefore could not have been improved uniformly over all $\mu^* \in \mathbb{R}^d$.

Theorem 7.4. *If $\hat{\mu} : \mathbb{R}^{Nd} \to \mathbb{R}^d$ is an estimator such that $\mathbb{P}_{\mu^*}^N \left[\|\hat{\mu} - \mu^*\|_2 \leq r^* \right] \geq 1 - \delta$ for all $\mu^* \in \mathbb{R}^d$ and all $\delta \in (0, 1/4]$, then*

$$r^* \geq \max\left(\frac{1}{24}\sqrt{\frac{\log 2}{2\log(5/4)}}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}}, \frac{1}{12}\sqrt{\frac{\|\Sigma\|_{op}\log(1/\delta)}{N}} \right).$$

Given that the empirical mean $\bar{X}_N$ is such that for all $\mu \in \mathbb{R}^d$ with $\mathbb{P}_\mu^N$-probability at least $1 - \delta$,

$$\left\|\bar{X}_N - \mu\right\|_2 \leq \sqrt{\frac{\mathrm{Tr}\left(\Sigma\right)}{N}} + \sqrt{\frac{2\left\|\Sigma\right\|_{op}\log(1/\delta)}{N}}$$

we conclude from Theorem 7.4 that the sub-gaussian rate (7.2) is the deviation-minimax rate of convergence for the multivariate mean estimation problem w.r.t. ℓ_2^d and that it is achieved by

the empirical mean. In particular, there are no statistical procedure that can do better than the empirical mean uniformly over all mean vectors $\mu^* \in \mathbb{R}^d$ up to constant, this includes in particular all discretized versions of $\bar{X}_N$.

7.3 Convex programs

In this section, we introduce statistical procedures which are solutions to convex programs and which can achieve the rate from Theorem 7.1 without the unnecessary Gaussian mean width term $\mathbb{E}\left\|\Sigma^{1/2}G\right\|$. We also show that these procedures handle adversarial corruption and may still perform optimally in some situations where there is not even a first moment.

7.3.1 Construction of the Fenchel-Legendre minimum estimators.

Definition 7.1. *Let S be a subset of $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$. The Fenchel-Legendre transform of f on S is the function f_S^* defined for all $\mu \in \mathbb{R}^d$ by $f_S^*(\mu) = \sup_{v\in S}\left(\langle \mu, v\rangle - f(v)\right)$.*

For our purpose, the main property of a Fenchel-Legendre transform we will use is that it is a convex function as it is the maximal function of the family $(\mu \in \mathbb{R}^d \to \langle \mu, v\rangle - f(v) : v \in S)$ of linear functions.

We are now defining two examples of functions such that by taking the minimum of their Fenchel-Legendre transform over S will lead to optimal estimators of μ^* w.r.t. $\|\cdot\|_S$. The construction of these two functions are based on the median-of-means principle: the dataset $\{X_1, \dots, X_N\}$ is split into K equal size blocks of data indexed by $(B_k)_k$ forming an equipartition of $[N]$. On each block, an empirical mean is constructed $\bar{X}_k = |B_k|^{-1}\sum_{i\in B_k} X_i$. The two functions we are considering are using the K bucketed means $(\bar{X}_k)_k$ and are defined, for all $v \in \mathbb{R}^d$, by

$$f(v) = \frac{1}{|I_K|}\sum_{k\in I_K}\langle \bar{X}_k, v\rangle_{(k)}^* \text{ and } g(v) = \mathrm{Med}(\langle \bar{X}_k, v\rangle) = \langle \bar{X}_k, v\rangle_{(\frac{K+1}{2})}^* \tag{7.6}$$

where if $a_k = \langle \bar{X}_k, v\rangle, k \in [K]$ then $\langle \bar{X}_k, v\rangle_{(k)}^*, k \in [K]$ are the rearrangement of $(a_k)_k$ such that $a_{(1)}^* \le \ldots \le a_{(K)}^*$ (this is the rearrangement of the values a_k's themselves and not of their absolute values) and

$$I_K = \left[\frac{K+1}{4}, \frac{3(K+1)}{4}\right] = \left\{\frac{K+1}{2} \pm k : k = 0, 1, \cdots, \frac{K+1}{4}\right\}$$

is the inter-quartiles interval – w.l.o.g. we assume that $K+1$ can be divided by 4. In other words, $f(v)$ is the average sum over all inter-quartile values of the vector $(\langle \bar{X}_k, v\rangle)_{k\in[K]}$ and $g(v)$ is the median of this vector. Note that both functions f and g are homogeneous i.e. $f(\theta v) = \theta f(v)$ and $g(\theta v) = \theta g(v)$ for every $v \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$ and in particular they are odd functions; two facts we will use later.

We are now considering the Fenchel-Legendre transform of the functions f and g over a symmetric set S:

$$f_S^* : \mu \in \mathbb{R}^d \to \sup_{v\in S}\left(\langle \mu, v\rangle - f(v)\right) \text{ and } g_S^* : \mu \in \mathbb{R}^d \to \sup_{v\in S}\left(\langle \mu, v\rangle - g(v)\right). \tag{7.7}$$

As mentioned previously the two functions f_S^* and g_S^* are convex functions. We are now using them to define convex programs whose solutions will be proved to be robust and subgaussian

estimators of the mean / location vector μ^* w.r.t. $\|\cdot\|_S$:

$$\hat{\mu}_S^f \in \operatorname*{argmin}_{\mu\in\mathbb{R}^d} f_S^*(\mu) \text{ and } \hat{\mu}_S^g \in \operatorname*{argmin}_{\mu\in\mathbb{R}^d} g_S^*(\mu). \tag{7.8}$$

For some special choices of S, the Fenchel-Legendre minimization estimator $\hat{\mu}_S^g$ coincides with some classical procedures. This is for instance the case when $S = B_1^d$ (the unit ball of the ℓ_1^d-norm) or $S = B_2^d$. Indeed, when $S = B_1^d$, $\hat{\mu}_S^g$ is the coordinate-wise Median of Means:

$$\hat{\mu}_S^g = \operatorname*{argmin}_{\mu=(\mu_j)\in\mathbb{R}^d} \max_{j\in[d]} \left|\mu_j - \operatorname{Med}\left(\langle \bar{X}_k, e_j\rangle\right)\right| = \left(\operatorname{Med}\left(\langle \bar{X}_k, e_j\rangle\right) : j \in [d]\right) \tag{7.9}$$

where $(e_j)_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$, because $\|\cdot\|_S = \|\cdot\|_{\mathrm{conv}(S)}$ where $\mathrm{conv}(S)$ is the convex hull of S and so one may just take $S = \{\pm e_j : j \in [d]\}$. It is therefore possible to derive deviation-minimax optimal bounds for the coordinate-wise Median of Means w.r.t. the ℓ_∞^d-norm from general upper bounds on $\hat{\mu}_S^g$ since in that case $\|\cdot\|_S = \|\cdot\|_\infty$.

In the case $S = B_2^d$ (that is for the mean/location estimation problem w.r.t. ℓ_2^d), the Fenchel-Legendre minimum estimator $\hat{\mu}_S^g$ is a minmax MOM estimator Lecué and Lerasle (2020). This connection allows to write $\hat{\mu}_S^g$ (as well as $\hat{\mu}_S^f$) as a non-constraint estimator, it also shows that this minmax MOM estimator is actually solution to a convex optimization problem and how minmax MOM estimator can be generalized to other estimation risks.

Minmax MOM estimators have been introduced as a systematic way to construct robust and subgaussian estimators in Lecué and Lerasle (2020). They have been proved to be deviation-minimax optimal for the mean estimation problem in Lerasle et al. (2019) w.r.t. $\|\cdot\|_2$. Their definition only requires to consider a loss function; here we take for all $\mu \in \mathbb{R}^d$, $\ell_\mu : x \in \mathbb{R}^d \to \|x-\mu\|_2^2$ and the minmax MOM estimator is then defined as

$$\tilde{\mu} \in \operatorname*{argmin}_{\mu\in\mathbb{R}^d} \sup_{\nu\in\mathbb{R}^d} \operatorname{Med}\left(P_{B_k}(\ell_\mu - \ell_\nu) : k \in [K]\right) \tag{7.10}$$

where P_{B_k} is the empirical measure on the data in block B_k. The minmax MOM estimator $\tilde{\mu}$ was proved to achieve the subgaussian rate in (7.2) with confidence $1-\delta$ when the number of blocks is $K \sim \log(1/\delta)$ and $K \gtrsim |\mathcal{O}|$ in Lerasle et al. (2019).

Even though the minmax formulation of $\tilde{\mu}$ suggests a robust version of a descent/ascent gradient method over the median block (see Lecué and Lerasle (2020); Lerasle et al. (2019) for more details), no proof of convergence of this algorithm is known so far. Moreover, the main drawback of the minmax MOM estimator seems to be that it is solution of a non-convex optimization problem and may therefore be likely to be rather difficult to compute in practice. In the next result, we show that this is not the case since the minmax MOM estimator (7.10) is in fact equal to $\hat{\mu}_S^g$ for $S = B_2^d$ and it is therefore solution to a convex optimization problem.

Proposition 7.1. *The minmax MOM estimator $\tilde{\mu}$ defined in (7.10) satisfies $\tilde{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} g_{B_2^d}^*(\mu)$. The minmax MOM estimator is therefore solution to a convex optimization problem.*

Proof. We show that $\tilde{\mu} \in \operatorname{argmin}_{\mu\in\mathbb{R}^d} \sup_{\|v\|_2=1} \operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle)$. We consider the quadratic/multiplier decomposition of the difference of loss functions: for all $\mu,\nu \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$, we have $(\ell_\mu - \ell_\nu)(x) = \|x-\mu\|_2^2 - \|x-\nu\|_2^2 = -2\langle x-\mu, \mu-\nu\rangle - \|\mu-\nu\|_2^2$. Hence, for all $\mu \in \mathbb{R}^d$, we have

$$\sup_{\nu\in\mathbb{R}^d} \operatorname{Med}\left(P_{B_k}(\ell_\mu - \ell_\nu)\right) = \sup_{\nu\in\mathbb{R}^d} \left(-2\operatorname{Med}(\langle \bar{X}_k - \mu, \mu-\nu\rangle) - \|\mu-\nu\|_2^2\right)$$

$$= \sup_{\|v\|_2=1} \sup_{\theta\geq 0} \left(2\theta \operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle) - \theta^2\right) = \sup_{\|v\|_2=1} \left(\operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle)\right)^2 = \left(\sup_{\|v\|_2=1} \operatorname{Med}(\langle \bar{X}_k - \mu, v\rangle)\right)^2.$$

We conclude since

$$\operatorname*{argmin}_{\mu\in\mathbb{R}^d}\left(\sup_{\|v\|_2=1}\operatorname{Med}(\langle\bar{X}_k-\mu,v\rangle)\right)^2=\operatorname*{argmin}_{\mu\in\mathbb{R}^d}\sup_{\|v\|_2=1}\operatorname{Med}\left(\langle\bar{X}_k-\mu,v\rangle\right).$$

■

It follows from Proposition 7.1 that the minmax MOM estimator $\tilde{\mu}$ is solution to a convex optimization problem. This fact is far from being obvious given the definition of $\tilde{\mu}$ in (7.10).

Proposition 7.1 suggests a new formulation for $\hat{\mu}_S^g$ and $\hat{\mu}_S^f$. It is indeed possible to write these estimators as regularized estimators instead of their original constraint formulation (note that the Fenchel-Legendre transforms in (7.7) are suprema over S and are therefore constraint optimization problems). We now show that we may write them as suprema over all $\mathbb{R}^d$ if we add an ad hoc regularization function.

Let us introduce the two following functions which may be seen as regularized versions of the two f and g functions from (7.6): for all $\nu\in\mathbb{R}^d$,

$$F_S(\nu)=f(\nu)+\frac{\|\nu\|_S^2}{4}\text{ and }G_S(v)=g(\nu)+\frac{\|\nu\|_S^2}{4}.\tag{7.11}$$

We also consider their Fenchel-Legendre transforms over the entire set $\mathbb{R}^d$: for all $\mu\in\mathbb{R}^d$,

$$F_S^*(\mu)=\sup_{\nu\in\mathbb{R}^d}\left(\langle\mu,\nu\rangle-F_S(\nu)\right)\text{ and }G_S^*(\mu)=\sup_{\nu\in\mathbb{R}^d}\left(\langle\mu,\nu\rangle-G_S(\nu)\right).$$

The next result shows that the later two Fenchel-Legendre transforms can be used to define the two estimators $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$. The proof of Proposition 7.2 is similar to the one of Proposition 7.1 where the ℓ_2-norm is replaced by $\|\cdot\|_S$ and is therefore omitted.

Proposition 7.2. *Let S be a symmetric subset of $\mathbb{R}^d$ such that $\operatorname{span}(S)=\mathbb{R}^d$. We have $\hat{\mu}_S^f\in\operatorname{argmin}_{\mu\in\mathbb{R}^d}F_S^*(\mu)$ and $\hat{\mu}_S^g\in\operatorname{argmin}_{\mu\in\mathbb{R}^d}G_S^*(\mu)$.*

As a consequence of Proposition 7.2, one can write the two estimators $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ as solutions to unconstrained minmax optimization problems like the minmax MOM estimator (7.10) and in particular, one may design an alternating ascent/descent sub-gradient algorithm similar to the one from Lecué and Lerasle (2020) – we expect the one associated with $\hat{\mu}_S^f$ which uses half of the dataset at each iteration to be more efficient than the one associated with $\hat{\mu}_S^g$ which uses only the N/K data in the median block at each iteration. That is the reason why we provide in Figure 13 this algorithm only for

$$\hat{\mu}_S^f\in\operatorname*{argmin}_{\mu\in\mathbb{R}^d}\sup_{\nu\in\mathbb{R}^d}\left(\langle\mu,\nu\rangle-\frac{1}{|I_K|}\sum_{k\in I_K}\langle\bar{X}_k,\nu\rangle_{(k)}^*-\frac{\|\nu\|_S^2}{4}\right).$$

We also recall that by the Danskin-Bertekas theorem the subgradient of $\|\cdot\|_S$ at $\nu\in\mathbb{R}^d$ when S is a compact and non empty set is given by the convex hull of all $x\in S$ such that $\|\nu\|_S=\langle x,\nu\rangle$.

input : the data $X_1, \dots, X_N$, a number K of blocks, two decreasing steps size sequences $(\eta_t)_t, (\theta_t)_t \subset \mathbb{R}_+^*$ and $\epsilon > 0$ a stopping parameter
output : A robust estimator of the mean μ

Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$ at random
Construct the K empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
Compute $\tilde{\mu}^{(0)}$ the coordinate-wise median-of-means and put $\mu^{(0)} = \tilde{\mu}^{(0)}$ and $\nu^{(0)} = \tilde{\mu}^{(0)}$
while $\left\|\mu^{(t)} - \mu^{(t+1)}\right\|_S \geq \epsilon$ **do**
    Construct an equipartition $B_1 \sqcup \cdots \sqcup B_K = \{1, \cdots, N\}$ at random
    Construct the K empirical means $\bar{X}_k = (N/K) \sum_{i \in B_k} X_i, k \in [K]$
    Find the inter-quartile block numbers $k_1, \dots, k_{(K+1)/2} \in [K]$ such that

$$f(\nu^{(t)}) = \frac{1}{|I_K|} \sum_{j=1}^{(K+1)/2} \langle \bar{X}_{k_j}, \nu^{(t)} \rangle.$$

    Construct $g^{(t)}$ a subgradient of $\|\cdot\|_S$ at $\nu^{(t)}$ and the ascent direction

$$\nabla_\nu^{(t+1)} = \mu^{(t)} - \frac{1}{|I_K|} \sum_{j=1}^{(K+1)/2} \bar{X}_{k_j} - \frac{\left\|\nu^{(t)}\right\|_S g^{(t)}}{2}.$$

    Update $\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t \nabla_\nu^{(t+1)}$.
    Make one descent step: $\mu^{(t+1)} \leftarrow \mu^{(t)} - \theta_t \nu^{(t+1)}$.
end
Return $\mu^{(t+1)}$

Algorithm 13: An alternating ascent/descent algorithm for the robust mean estimation problem w.r.t. $\|\cdot\|_S$ with randomly chosen blocks of data at each step.

7.3.2 The adversarial corruption model and two models for inlier.

In this section, we introduce the assumptions under which we will obtain some statistical upper bounds for the Fenchel-Legendre minimum estimators introduced above. We are considering two types of assumptions: one for the outliers which will be the adversarial corruption model and one for the inlier which will be either the existence of a second moment or a regularity assumption on a family of cdf around 0. We start with the adversarial corruption model.

Assumption 7.1. *There exists N independent random vectors $(\tilde{X}_i)_{i=1}^N$ in $\mathbb{R}^d$. The N random vectors $(\tilde{X}_i)_{i=1}^N$ are first given to an "adversary" who is allowed to modify up to $|\mathcal{O}|$ of these vectors. This modification does not have to follow any rule. Then, the "adversary" gives the modified dataset $(X_i)_{i=1}^N$ to the statistician. Hence, the statistician receives an "adversarially" contaminated dataset of N vectors in $\mathbb{R}^d$ which can be partitioned into two groups: the modified data $(X_i)_{i \in \mathcal{O}}$, which can be seen as outliers and the "good data" or inlier $(X_i)_{i \in \mathcal{I}}$ such that $\forall i \in \mathcal{I}, X_i = \tilde{X}_i$. Of course, the statistician does not know which data has been modified or not so that the partition $\mathcal{O} \cup \mathcal{I} = \{1, \dots, N\}$ is unknown to the statistician.*

In the adversarial contamination model from Assumption 7.1, the set $\mathcal{O} \subset [N]$ can depend arbitrarily on the initial data $(\tilde{X}_i)_{i=1}^N$; the corrupted data $(X_i)_{i \in \mathcal{O}}$ can have any arbitrary dependence structure; and the informative data $(X_i)_{i \in \mathcal{I}}$ may also be correlated (for instance, it

is, in general, the case when the $|\mathcal{O}|$ data $\tilde{X}_i$ with largest ℓ_2^d-norm are modified by the adversary). The adversarial corruption model covers the Huber ϵ-contamination model Huber and Ronchetti (2009) and also the $\mathcal{O} \cup \mathcal{I}$ framework from Lecué and Lerasle (2019); Lecué and Lerasle (2020); M. Lerasle and Lecué (2017).

Assumption 7.1 does not grant any property of the inlier data $(\tilde{X}_i)_{i\in[N]}$ except that they are independent. We will obtain a general result under only Assumption 7.1 in Section 7.4. However, to recover convergence rates similar to the one in Theorem 7.1 or the subgaussian rate in (7.5), we will grant some assumptions on the inlier as well. We are now considering two assumptions on the inlier which are of different nature.

The two assumptions on the inlier we are now considering are related to a subtle property of the Median-of-Means (MOM) principle which somehow benefits from its two components: the empirical median and the empirical mean. Indeed, MOM is en empirical median of empirical means and so if we refer to the classical asymptotic normality (a.n.) results of the empirical mean and the empirical median, the first one holds under the existence of a second moment and the second one holds under the assumption that the cdf is differentiable at the median with positive derivative at the median (see Corollary 21.5 in van der Vaart (1998)). We therefore recover these two types of assumptions when we work with estimators using the MOM principle. A nice feature of MOM based estimators is that their estimation results hold under either one of the two conditions and do not require the two assumptions to hold simultaneously. We can therefore consider the two assumptions independently and get two estimation results for the Fenchel-Legendre minimum estimators introduced above (which are based on the MOM principle). We start with the moment assumption.

Assumption 7.2. *The N independent random vectors $(\tilde{X}_i)_{i=1}^N$ have mean μ^* and there exists a SDP matrix $\Sigma \in \mathbb{R}^{d\times d}$ such that $\mathbb{E}(\tilde{X}_i - \mu^*)(\tilde{X}_i - \mu^*)^\top \preceq \Sigma$.*

Most of the statistical bounds obtained on MOM based estimators have focused on the heavy-tailed setup and have therefore consider Assumption 7.2 as their main assumption. This is the 'empirical mean component' of the MOM principle which has been the most exploited so far. It is however also possible to use the 'empirical median component' of the MOM principle to get statistical bounds in cases where a first moment does not even exist. In that case, μ^* is called a *location parameter* and Σ a *scale parameter*. Also, a natural assumption is similar to the one used to get the a.n. of the empirical median, that is an assumption on the cdf at the median adapted to the multidimensional and non-asymptotic setup. We are now introducing such an assumption.

Assumption 7.3. *The inlier data $(\tilde{X}_i)_{i=1}^N$ are i.i.d.. There exists $\mu^* \in \mathbb{R}^d$ and two absolute constants $c_0 > 0$ and $c_1 > 0$ such that the following holds: for all $v \in S$ and all $0 < r \leq c_0$, $H_{N,K,v}(r) \leq 1/2 - c_1 r$ where*

$$H_{N,K,v}(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \tilde{X}_i - \mu^*, v\rangle > r\right]. \tag{7.12}$$

A typical example where Assumption 7.3 holds is when $S = \mathcal{S}_2^{d-1}$ (that is for the location estimation problem w.r.t. the Euclidean ℓ_2^d norm) and the $\tilde{X}_i$'s are rotational invariant that is when for all $v \in \mathcal{S}_2^{d-1}$, $\langle \tilde{X}_1 - \mu^*, v\rangle$ has the same distribution as $\langle \tilde{X}_1 - \mu^*, e_1\rangle$ where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$. In that case, $\tilde{X}_1$ has the same distribution as $\mu^* + RU$ where R is a real-valued random variable on $\mathbb{R}_+$ independent of U a random vector uniformly distributed over $\mathcal{S}_2^{d-1}$. In that case and for $K = N$, for all $v \in \mathcal{S}_2^{d-1}$ and all $r \in \mathbb{R}$,

$$H_{N,K=N,v}(r) = H(r) := \mathbb{P}[R\langle U, e_1\rangle \geq r] = \int_r^{+\infty} f(x)dx \text{ where } f : x \in \mathbb{R} \to C_d \int_{|x|}^{+\infty} \frac{1}{u}\left(1 - \frac{x^2}{u^2}\right)^{\frac{d-3}{2}} d\mathbb{P}_R(u),$$

$\mathbb{P}_R$ is the probability distribution of R and C_d is a normalization constant which can be proved to satisfy $\sqrt{d} \leq C_d \leq 6\sqrt{d}$ (see for instance, Chapter 4 in Bryc (1995)). In particular, it follows from the mean value theorem that for all $r \geq 0$, $H(r) \leq H(0) - \min_{0\leq x\leq r} f(x)r = 1/2 - f(r)r$. Therefore, Assumption 7.3 holds in that case when there exists constants $c_0', c_1' > 0$ such that $f(c_0') \geq c_1'$, which in turn holds when there exists constants $c_0, c_1 > 0$ such that $H(c_0) \leq 1/2 - c_1$.

Furthermore, we have, for all $t > 0$

$$\begin{aligned}\mathbb{P}[R\langle U, e_1\rangle \geq c_0] &\leq \mathbb{P}[\langle U, e_1\rangle \geq t/\sqrt{d}] + \frac{1}{2}\mathbb{P}[R \geq c_0\sqrt{d}/t] \\ &\leq e^{-t^2/2} + \frac{1}{2}\mathbb{P}[R \geq c_0\sqrt{d}/t],\end{aligned}$$

where the classical second inequality can be found for instance in Vershynin (2018), Chapter 5. So if for some constants $\tilde{c}_0, \tilde{c}_1 > 0$, $bP[R \geq \tilde{c}_0\sqrt{d}] \leq 1 - \tilde{c}_1$, then Assumption 7.3 holds. This is for instance the case, when R is distributed like $\|G\|_2$ for $G \sim \mathcal{N}(0, I_d)$ (by Borell-TIS inequality) but as well when R is the positive part of a Cauchy variable for instance. As a consequence, Assumption 7.3 has nothing to do with the existence of any moment and it may hold even when there is not a first moment and even for $K = N$.

Another example where Assumption 7.3 holds, that we will use in the following to obtain statistical bounds for the coordinate-wise median of means for the location problem is when $S = \{\pm e_j : j \in [d]\}$ and $\tilde{X}_1 = \mu^* + Z$ where $Z = (z_j)_{j=1}^d$ is random vector in $\mathbb{R}^d$ with coordinates $z_1, \ldots, z_d$ having a symmetric around 0 Cauchy distribution. In that case, $\tilde{X}_1$ does not have a first moment and μ^* is a location parameter as the center of symmetry of the distribution of $\tilde{X}_1$. We have for all $j \in [d]$,

$$H_{N,K=N,\pm e_j}(r) = \mathbb{P}\left[\langle \tilde{X}_1 - \mu^*, \pm e_j\rangle \geq r\right] = \mathbb{P}[z_j \geq r] = \int_r^{+\infty} \frac{dx}{\pi(1+x^2)} \leq \frac{1}{2} - \frac{r}{\pi(1+r^2)} \leq \frac{1}{2} - \frac{r}{2\pi}$$

for all $0 < r \leq 1$. Therefore, Assumption 7.3 holds in that case as well.

7.3.3 Statistical bounds for $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$

In this section, we obtain estimation bounds w.r.t. $\|\cdot\|_S$ for $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ in the adversarial contamination model with either the L_2 moment Assumption 7.1 or the regularity at 0 Assumption 7.3.

Estimation properties of $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ under Assumption 7.2. In this section, we obtain high probability estimation upper bounds satisfied by $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$ w.r.t. $\|\cdot\|_S$ in the adversarial contamination and heavy-tailed inlier model. The rate of convergence is given by the quantity

$$r_S^* = \max\left(\frac{64}{\sqrt{N}}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu^*)\right\|_S, \sup_{v\in S}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\frac{64K}{N}}\right). \tag{7.13}$$

The key metric property satisfied by the two Fenchel-Legendre transforms f_S^* and g_S^* in the adversarial contamination and heavy-tailed inlier model is the following isomorphic result.

Lemma 7.1. *Grant Assumption 7.1 and Assumption 7.2. Let S be a symmetric subset of $\mathbb{R}^d$. Assume that $|\mathcal{O}| < K/16$. With probability at least $1 - \exp(-K/512)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq g_S^*(\mu^*) \leq r_S^*$ and $|f_S^*(\mu) - \|\mu - \mu^*\|_S| \leq f_S^*(\mu) \leq r_S^*$.*

Lemma 7.1 shows that if $\|\mu - \mu^*\|_S \geq 2r_S^*$ then $\|\mu - \mu^*\|_S \leq g_S^*(\mu) \leq 2\|\mu - \mu^*\|_S$ and the same holds for f_S^*. It means that both g_S^* and f_S^* are two convex functions equivalent (up to

absolute constants) to $\mu \to \|\mu - \mu^*\|_S$ on $\mathbb{R}^d \backslash (2r_S^*)B_S$, where B_S is the unit ball associated with $\|\cdot\|_S$ and, on $(2r_S^*)B_S$, they are both smaller than $2r_S^*$. Hence, both $g_S^*(\cdot - \mu^*)$ and $f_S^*(\cdot - \mu^*)$ provide a good approximation of the metric space $(\mathbb{R}^d, \|\cdot\|_S)$. In particular, any minimum of g_S^* and f_S^* will be close (up to r_S^*) to a minimum of $\mu \to \|\mu - \mu^*\|_S$ which is μ^*. This explains the statistical properties of $\hat{\mu}_S^f$ and $\hat{\mu}_S^g$: from Lemma 7.1,

$$\left\|\hat{\mu}_S^f - \mu^*\right\|_S \leq f_S^*(\hat{\mu}_S^f) + f_S^*(\mu^*) \leq 2f_S^*(\mu^*) \leq 2r_S^*$$

and the same holds for $\hat{\mu}_S^g$. This leads to the following result.

Theorem 7.5. *Grant Assumption 7.1 and Assumption 7.2. Let S be a symmetric subset of $\mathbb{R}^d$ and r_S^* be defined in (7.13). For all $K > 16|\mathcal{O}|$, with probability at least $1 - \exp(-K/512)$,*

$$\left\|\hat{\mu}_S^f - \mu^*\right\|_S \leq 2r_S^* \text{ and } \|\hat{\mu}_S^g - \mu^*\|_S \leq 2r_S^*.$$

The rate r_S^* obtained in Theorem 7.5 can be split into two terms: the complexity term given by the Rademacher complexity and a deviation term exhibiting the weak variance term as in the Gaussian case. Compare with Theorem 7.1 from Lugosi and Mendelson (2019b), this result shows that the Gaussian mean width term appearing in Theorem 7.1 is actually not necessary, it also shows that this improved rate can be obtained by a procedure solution to a convex program and that it can also handle adversarial corruption. When $S = B_2^d$, we recover the classical subgaussian rate because in that case the Rademacher complexity term in r_S^* is less or equal to $\sqrt{\mathrm{Tr}(\Sigma)}$, see Koltchinskii (2006). In particular, since $\hat{\mu}_S^g$ is the minmax MOM estimator in that case, we recover the main result from Lerasle et al. (2019).

Estimation properties of $\hat{\mu}_S^g$ under Assumption 7.3. In this section, we consider some cases where a first moment may not exist; in that case, μ^* is a location parameter so that Assumption 7.3 holds. The rate of convergence we obtain in that case is given by

$$r^\diamond = \frac{C_0}{c_1}\left(\sqrt{\frac{d+1}{N}} + \sqrt{\frac{u}{N}}\right) + \frac{|\mathcal{O}|}{c_1\sqrt{KN}} \tag{7.14}$$

where c_1 is the absolute constant from Assumption 7.3, C_0 the absolute constant from (7.28) and $u > 0$ a confidence parameter.

The following result is an isomorphic result satisfied by the Fenchel-Legendre transforms g_S^* under Assumption 7.3. It is similar to the one of Lemma 7.1 but with the rate $r^\diamond$.

Lemma 7.2. *Let S be a symmetric subset of $\mathbb{R}^d$. Grant Assumption 7.1 and Assumption 7.3 for some $K \in [N]$. Let $u > 0$. Assume that $C_0\left(\sqrt{(d+1)/K} + \sqrt{u/K}\right) + |\mathcal{O}|/K \leq c_0c_1$. With probability at least $1 - \exp(-u)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq r^\diamond$.*

As explained below Lemma 7.1, a result such as Lemma 7.2 may be used to upper bound the $\|\cdot\|_S$ distance between $\hat{\mu}_S^g$, a minimum of g_S^*, and μ^*, a minimum of $\mu \to \|\mu - \mu^*\|_S$. This yields to the following result.

Theorem 7.6. *Let S be a symmetric subset of $\mathbb{R}^d$. Grant Assumption 7.1 and Assumption 7.3 for some $K \in [N]$. Let $u > 0$ and assume that $C_0\left(\sqrt{(d+1)/K} + \sqrt{u/K}\right) + |\mathcal{O}|/K \leq c_0c_1$. With probability at least $1 - \exp(-u)$, $\|\hat{\mu}_S^g - \mu^*\|_S \leq 2r^\diamond$ where $r^\diamond$ is defined in (7.14).*

Unlike Theorem 7.5, Theorem 7.6 may hold even when there is not a first moment. The result from Theorem 7.6 holds for all $0 < u \lesssim K$ whereas Theorem 7.5 holds only for $u = K$ (even though one may use a Lepski's adaptive scheme to chose adaptively K in that case). The price for adversarial corruption in (7.14) is between $|\mathcal{O}|/N$ (for $K \sim N$) and $\sqrt{|\mathcal{O}|/N}$ (for $K \sim |\mathcal{O}|$). It therefore depends on the choice of K for which Assumption 7.3 holds. As shown after Assumption 7.3 for spherically symmetric random variables one can take $K = N$ and so the best possible price $|\mathcal{O}|/N$ for adversarial corruption may be achieved even when a first moment does not exist. If one needs some averaging effect so that Theorem 7.6 holds, then one should take K as small as possible that is $K \sim |\mathcal{O}|$ and then $\sqrt{|\mathcal{O}|/N}$ will be the price for adversarial corruption as in the L_2 case described in Theorem 7.6.

Subgaussian rates under weak or no moment assumption. It is possible to recover (up to absolute constants) the subgaussian rate (7.5) in Theorem 7.5 for $K \sim \log(1/\delta)$ when the Rademacher complexity term from (7.13) and the Gaussian mean width from (7.5) satisfy

$$\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu^*)\right\|_S \lesssim \ell^*\left(\Sigma^{1/2}S\right). \tag{7.15}$$

Such a result (i.e. Rademacher complexity is smaller than the Gaussian mean width up to constant) depends on the set S and the number of moments granted on the $\tilde{X}_i$'s as well as the sample size. It obviously holds when the $\tilde{X}_i$'s are i.i.d. $\mathcal{N}(\mu^*, \Sigma)$, so that we recover the deviation-minimax optimal subgaussian rate (7.5) in that case. It is also true when the $\tilde{X}_i$'s are subgaussian vectors. There are other situations under weaker moment assumption where (7.15) holds.

For instance, when $S = B_2^d$, (7.15) holds under only a L_2-moment assumption (see Koltchinskii (2006)). It also holds for $S = B_1^d$ when the $\tilde{X}_i$'s are isotropic with coordinates having $\log d$ subgaussian moments (i.e. $\left\|\langle \tilde{X}_i, e_j\rangle\right\|_{L_p} \leq L\sqrt{p}$ for all $1 \leq p \leq \log d$ and $j \in [d]$) and $N \gtrsim \log d$. Together with (7.9) and Theorem 7.5, this implies that the coordinate-wise MOM is a subgaussian estimator of the mean under a $\log d$ subgaussian moment assumption. Upper bounds such as (7.15) have been extended in Mendelson (2017c) to general unconditional norms.

It is also possible to recover the subgaussian rate (7.5) in situations where there is not even a first moment thanks to Theorem 7.6. Indeed, for the case $S = B_1^d$ and $\tilde{X}_1 = \mu^* + Z$ where $Z = (z_j)_{j=1}^d$ has symmetric around 0 Cauchy distributed coordinates, we showed that Assumption 7.3 holds for $K = N$ and that $\hat{\mu}_S^g$ is the coordinate-wise median (here $K = N$) in (7.9). It follows from Theorem 7.6 that, when $d \lesssim N$ and $|\mathcal{O}| \lesssim N$ then for all $d \leq u \lesssim N$, with probability at least $1 - \exp(-u)$,

$$\|\hat{\mu}_S^g - \mu^*\|_\infty \leq 2C_0\left(\sqrt{\frac{d+1}{N}} + \sqrt{\frac{u}{N}}\right) + \frac{2\pi|\mathcal{O}|}{N} \tag{7.16}$$

which is the deviation-minimiax optimal subgaussian rate (7.5) we would have gotten if the $\tilde{X}_i$ were i.i.d. isotropic Gaussian vectors centered in μ^* corrupted by $|\mathcal{O}|$ adversarial outliers (up to absolute constants). But here, (7.16) is obtained without the existence of a first moment. Moreover, in (7.16), the number of outliers is allowed to be proportional to N and the price for adversarial corruption is of the order of $|\mathcal{O}|/N$ which is the same price we have to pay when inlier have a Gaussian distribution – this differs from the $\sqrt{|\mathcal{O}|/N}$ information theoretical lower bound that has been obtained for some non-symmetric inlier. Furthermore, the computational cost of the coordinate-wise MOM is $\mathcal{O}(Nd)$ since the cost for computing the bucketed means

is $\mathcal{O}(Nd)$, the one of finding the median of K numbers is $\mathcal{O}(K)$, see Blum et al. (1973), it is therefore the same computational cost as the one of the empirical mean. It is therefore possible to achieve the same computational and statistical properties as the empirical mean in a setup where a first moment does not even exist.

7.4 Proofs

Proof of Theorem 7.3. The minimax lower bound rate r^* exhibits two quantities: one which is a *complexity term* depending on the Gaussian mean width of $\Sigma^{1/2}S$ and a *deviation term* depending on δ. The two terms come from two arguments. We start with the deviation term.

Let $v_1 \in \mathbb{R}^d$ be such that $\|v_1\|_S = 1$. We consider two Gaussian measures on $\mathbb{R}^{dN}$: $\mathbb{P}_0 = \mathcal{N}(0,\Sigma)^{\otimes N}$ and $\mathbb{P}_1 = \mathcal{N}(3r^*v_1,\Sigma)^{\otimes N}$. They are the distributions of a sample of N i.i.d. Gaussian vectors in $\mathbb{R}^d$ with the same covariance matrix Σ and the first one with mean 0 and the second one with mean $3r^*v_1$. We set $A_0 = (\hat{\mu})^{-1}(B_S(0,r^*)) = \{(x_1,\dots,x_N) \in \mathbb{R}^{Nd} : \|\hat{\mu}(x_1,\dots,x_N)\|_S \leq r^*\}$ and $A_1 = (\hat{\mu})^{-1}(B_S(3r^*v_1,r^*))$. It follows from the statistical properties of $\hat{\mu}$ that $\mathbb{P}_0[A_0] \geq 1-\delta$ and $\mathbb{P}_1[A_1] \geq 1-\delta$.

The key ingredient for the deviation lower bound term is a slightly generalization of Lemma 3.3 in Lecué and Mendelson (2013) which is based on a version of the Gaussian shift Theorem from Li and Kuelbs (1998).

Lemma 7.3. *Let $t \mapsto \Phi(t) = \mathbb{P}(g \leq t)$ be the cumulative distribution function of a standard gaussian random variable on $\mathbb{R}$. Let $\Sigma_0 \succeq 0$ be in $\mathbb{R}^{(Nd)\times(Nd)}$ and $u,v \in \mathbb{R}^{dN}$. Let two gaussian measures $\nu_u \sim \mathcal{N}(u,\Sigma_0)$ and $\nu_v \sim \mathcal{N}(v,\Sigma_0)$ on $\mathbb{R}^{Nd}$. If $A \subset \mathbb{R}^{dN}$ is measurable, then*

$$\nu_v(A) \geq 1 - \Phi\big(\Phi^{-1}(1-\nu_u(A)) + \|\Sigma_0^{-1/2}(u-v)\|_2\big) \tag{7.17}$$

where $\Sigma_0^{-1/2}$ is the square root of the pseudo-inverse of Σ_0.

Proof of Lemma 7.3. When $\Sigma_0 = I_{Nd}$, Lemma 7.3 is exactly Lemma 3.3 in Lecué and Mendelson (2013) for $\sigma = 1$. To prove Lemma 7.3, we observe that $\nu_v(A) = \mathbb{P}[G + \Sigma_0^{-1/2}v \in B]$ where $B = \Sigma_0^{-1/2}A$ and G is a standard Gaussian variable in $\mathrm{Im}(\Sigma_0)$. Hence, it follows from Lemma 3.3 in Lecué and Mendelson (2013) that

$$\mathbb{P}[G + \Sigma_0^{-1/2}v \in B] \geq 1 - \Phi\big(\Phi^{-1}(1-\mathbb{P}[G+\Sigma_0^{-1/2}u \in B]) + \|\Sigma_0^{-1/2}(u-v)\|_{\ell_2^N}\big)$$

which is exactly (7.17). ∎

It follows from Lemma 7.3 that

$$\mathbb{P}_1[A_0] \geq 1 - \Phi\left[\Phi^{-1}(1-\mathbb{P}_0[A_0]) + \left\|\Sigma_0^{-1/2}(0-(3r^*v_1,\dots,3r^*v_1))\right\|_2\right]. \tag{7.18}$$

Moreover, we have $\Phi^{-1}(1-\mathbb{P}_0[A_0]) \leq \Phi^{-1}(\delta)$ (because $1-\mathbb{P}_0[A_0] \leq \delta$) and

$$\left\|\Sigma_0^{-1/2}(0-(3r^*v_1,\dots,3r^*v_1))\right\|_2 = 3r^*\sqrt{N}\left\|\Sigma^{-1/2}v_1\right\|_2. \tag{7.19}$$

As a consequence, if $3r^*\sqrt{N}\left\|\Sigma^{-1/2}v_1\right\|_2 \leq -\Phi^{-1}(\delta)$ then, in (7.18), we get $\mathbb{P}_1[A_0] \geq 1-\Phi[0] \geq 1/2$ which is not possible because $\mathbb{P}_1[A_1] \geq 1-\delta > 3/4$ and $A_1 \cap A_0 = \emptyset$. As a consequence, we necessarily have $3r^*\sqrt{N} \geq (-\Phi^{-1}(\delta))\left\|\Sigma^{-1/2}v_1\right\|_2^{-1}$. The later holds for any $v_1 \in \mathbb{R}^d$ such that $\|v_1\|_S = 1$ hence $3r^*\sqrt{N} \geq (-\Phi^{-1}(\delta))[1/\inf_{\|v\|_S=1}\left\|\Sigma^{-1/2}v\right\|_2]$. It also follows from the bound on

the Mill's ratio from Komatu (1955) (here we use that for all $x \geq 0$, $\Phi(-x) \geq 2\varphi(x)/\sqrt{4+x^2}+x$ where φ is the standard Gaussian density function) that for all $0 < \delta < 1/4$, $-\Phi^{-1}(\delta) \geq 1/4\sqrt{\log(1/\delta)}$. This shows that

$$r^* \geq \frac{1}{12}\sqrt{\frac{\log(1/\delta)}{N}}\frac{1}{\inf_{\|v\|_S=1}\left\|\Sigma^{-1/2}v\right\|_2}. \tag{7.20}$$

To conclude on the deviation term, we use the following duality argument.

Lemma 7.4. *Let $A \in \mathbb{R}^{d\times d}$ be a symmetric and invertible matrix. Let $\|\cdot\|$ be a norm and its dual norm $\|\cdot\|^*$ on $\mathbb{R}^d$. Let S be a symmetric subset of $\mathbb{R}^d$ such that $\operatorname{span}(S) = \mathbb{R}^d$. We have*

$$\frac{1}{\inf_{\|v\|_S=1}\|A^{-1}v\|} \geq \sup_{w\in S}\|Aw\|^*.$$

Proof of Lemma 7.4. Let v be such that $\|v\|_S = 1$ and $w \in S$. We have $|\langle v, w\rangle| \leq 1$ and so $|\langle A^{-1}v/\|A^{-1}v\|, Aw\rangle| \leq 1/\|A^{-1}v\|$. The later holds for all v such that $\|v\|_S = 1$ and $\{A^{-1}v/\|A^{-1}v\| : \|v\|_S = 1\}$ is the unit sphere of $\|\cdot\|$. Hence, we conclude by taking the sup over v such that $\|v\|_S = 1$ and $w \in S$. ∎

It follows from (7.20) and Lemma 7.4 for $\|\cdot\| = \|\cdot\|_2$ and $A = \Sigma^{1/2}$ that

$$r^* \geq \frac{1}{12}\sqrt{\frac{\log(1/\delta)}{N}}\sup_{w\in S}\left\|\Sigma^{1/2}w\right\|_2. \tag{7.21}$$

Let us now turn to the second part of the lower bound; the one coming from the complexity of the problem (here, it is the Gaussian mean width of $\Sigma^{1/2}S$). We know that $\hat{\mu}$ is an estimator such that for all $\mu \in \mathbb{R}^d$, $\mathbb{P}_\mu^N\left[\|\hat{\mu}-\mu\|_S \leq r^*\right] \geq 1-\delta$ which is equivalent to say that

$$\delta \geq \sup_{\mu\in\mathbb{R}^d}\mathbb{E}_\mu^N\phi\left(\frac{\|\hat{\mu}-\mu\|_S}{r^*}\right) \tag{7.22}$$

where we set $\phi : t \in \mathbb{R} \to I(t > 1)$ and $\mathbb{E}_\mu^N$ is the expectation with respect to $X_1, \dots X_N \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \Sigma)$. Next, we consider a Gaussian distribution γ over the set of parameters $\mu \in \mathbb{R}^d$: for $s > 0$, we assume that $\mu \sim \mathcal{N}(0, s\Sigma)$. It follows from (7.22) that

$$\delta \geq \int_{\mu\in\mathbb{R}^d}\mathbb{E}_\mu^N\phi\left(\frac{\|\hat{\mu}-\mu\|_S}{r^*}\right)\gamma(\mu)d\mu = \mathbb{E}\left[\mathbb{E}\left[\phi\left(\frac{\|\hat{\mu}(X_1,\dots,X_N)-\mu\|_S}{r^*}\right)|X_1,\dots,X_N\right]\right]. \tag{7.23}$$

In other words, we lower bound the minmax risk by a Bayesian risk. We now use Anderson's lemma to lower bound the Bayesian risk appearing in (7.23). We first recall Anderson's Lemma.

Theorem 7.7 (Anderson's Lemma)**.** *Let Γ be a semi-definite $d \times d$ matrix and $Z \sim \mathcal{N}(0, \Gamma)$. Let $w : \mathbb{R}^d \to \mathbb{R}$ be such that all its level sets (i.e. $\{x \in \mathbb{R}^d : w(x) \leq c\}$ for $c \in \mathbb{R}$) are convex and symmetric around the origin. Then for all $x \in \mathbb{R}^d$, $\mathbb{E}w(Z + x) \geq \mathbb{E}w(Z)$.*

We remark that $\mu - \mathbb{E}[\mu|X_1, \dots, X_N]$ is distributed according to $\mathcal{N}(0, (s/(1+Ns)\Sigma))$ conditionally to $X_1, \dots, X_N$. Therefore, applying Anderson's Lemma conditionally to $X_1, \dots, X_N$, we obtain in (7.23) that

$$\delta \geq \mathbb{E}\left[\phi\left(\frac{\|\mathbb{E}[\mu|X_1,\dots,X_N]-\mu\|_S}{r^*}\right)\right] = \mathbb{P}\left[\left\|\Sigma^{1/2}G\right\|_S \geq \sqrt{\frac{1+Ns}{s}}r^*\right]$$

where $G \sim \mathcal{N}(0, I_d)$. This result is true for all $s > 0$ so taking $s \uparrow +\infty$, we obtain

$$\delta \geq \mathbb{P}\left[\left\|\Sigma^{1/2}G\right\|_S \geq \sqrt{N}r^*\right].$$

Using Borell-TIS's inequality (Theorem 7.1 in Ledoux (2001) or pages 56-57 in Talagrand (2014)), we know that with probability at least 4/5, $\left\|\Sigma^{1/2}G\right\|_S \geq \mathbb{E}\left\|\Sigma^{1/2}G\right\|_S - \sigma_S\sqrt{2\log(5/4)}$ where we set $\sigma_S = \sup_{\|v\|_S=1}\left\|\Sigma^{1/2}v\right\|_2$. As a consequence, for $\delta = 1/4$, we necessarily have $\sqrt{N}r^* \geq \mathbb{E}\left\|\Sigma^{1/2}G\right\|_S - \sigma_S\sqrt{2\log(5/4)}$ and so $\sqrt{N}r^* \geq (1/2)\mathbb{E}\left\|\Sigma^{1/2}G\right\|_S$ when $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_S \geq 2\sigma_S\sqrt{2\log(5/4)}$. Finally, when $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_S < 2\sigma_S\sqrt{2\log(5/4)}$, we know from (7.21) for $\delta = 1/4$ that

$$r^* \geq \frac{1}{12}\sqrt{\frac{\log 4}{N}}\sigma_S \geq \frac{1}{24}\sqrt{\frac{\log 2}{\log(5/4)}}\frac{\mathbb{E}\left\|\Sigma^{1/2}G\right\|_S}{\sqrt{N}}.$$

■

Proof of Theorem 7.4. Theorem 7.4 follows from Theorem 7.3 and the following lower bound on $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_{B_2^d}$. We have from Borell-TIS's inequality that

$$\begin{aligned}
&\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2^2 - \left(\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right)^2 = \mathbb{E}\left(\left\|\Sigma^{1/2}G\right\|_2 - \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right)^2 \\
&= \int_0^\infty \mathbb{P}\left[\left|\left\|\Sigma^{1/2}G\right\|_2 - \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right| \geq \sqrt{t}\right] dt \leq 2\sigma_{B_2^d}^2
\end{aligned}$$

where $\sigma_{B_2^d}^2 = \sup_{\|v\|_2=1}\left\|\Sigma^{1/2}v\right\|_2^2 = \|\Sigma\|_{op}$. Since $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2^2 = \mathrm{Tr}(\Sigma)$, we have $\left(\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2\right)^2 \geq \mathrm{Tr}(\Sigma) - 2\|\Sigma\|_{op}$. Therefore, $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \geq \sqrt{\mathrm{Tr}(\Sigma)/2}$ when $\mathrm{Tr}(\Sigma) \geq 4\|\Sigma\|_{op}$ and when $\mathrm{Tr}(\Sigma) < 4\|\Sigma\|_{op}$, we use the lower bound from (7.21) and an argument similar to the one appearing in the end of the proof of Theorem 7.3 to get the result. ■

Proof of Lemma 7.1. We first prove the result for the g_S^* function. The one for the f_S^* is similar up to constants and will be sketched after. The proof of Lemma 7.1 for the g_S^* function is a corollary of the general fact which holds under only Assumption 7.1. Let $u > 0$ be a confidence parameter and define R_S^* such that

$$\frac{4}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i - \mu)\right\|_S + \sqrt{\frac{2u}{K}} + \sup_{v\in S}H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right) + \frac{|\mathcal{O}|}{K} < \frac{1}{2}. \tag{7.24}$$

Let us show that with large probability for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq R_S^*$.

We have for all $\mu \in \mathbb{R}^d$,

$$|g_S^*(\mu) - \|\mu - \mu^*\|_S| = \left|\sup_{v\in S}\left(\langle\mu, v\rangle - g(v)\right) - \sup_{v\in S}\langle v, \mu - \mu^*\rangle\right| \leq \sup_{v\in S}|\langle\mu^*, v\rangle - g(v)| = g_S^*(\mu^*) \tag{7.25}$$

where we used that S is symmetric and g is odd. It only remains to show that $g_S^*(\mu^*) \leq R_S^*$ with large probability. To that end, it is enough to prove that, with large probability, for all $v \in S$,

$$\sum_{k\in[K]} I(\langle\bar{X}_k - \mu^*, v\rangle > R_S^*) < \frac{K}{2}. \tag{7.26}$$

We use the notation introduced in Assumption 7.1 and we consider $\overline{\tilde{X}}_k = |B_k|^{-1}\sum_{i\in B_k}\tilde{X}_i$ for $k\in[K]$ which are the K bucketed means constructed on the N independent vectors $\tilde{X}_i, i\in[N]$ before contamination (whereas $\bar{X}_k$ are the ones constructed after contamination). We also set $\mathcal{K} = \{k\in[K] : B_k\cap\mathcal{O}=\emptyset\}$ the indices of the non corrupted blocks. We have

$$\sum_{k\in[K]} I(\langle\bar{X}_k-\mu^*,v\rangle > R_S^*) = \sum_{k\in\mathcal{K}} I(\langle\bar{X}_k-\mu^*,v\rangle > R_S^*) + \sum_{k\notin\mathcal{K}} I(\langle\bar{X}_k-\mu^*,v\rangle > R_S^*)$$
$$\leq \sum_{k\in[K]} I(\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*) + |\mathcal{O}|. \tag{7.27}$$

It only remains to show that with probability at least $1-\exp(-u)$, for all $v\in S$,

$$\sum_{k\in[K]} I(\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*) \leq \frac{4K}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu^*)\right\|_S + \sqrt{2uK} + K\sup_{v\in S} H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right).$$

We define $\phi(t)=0$ if $t\leq 1/2$, $\phi(t)=2(t-1/2)$ if $1/2\leq t\leq 1$ and $\phi(t)=1$ if $t\geq 1$. We have $I(t\geq 1)\leq\phi(t)\leq I(t\geq 1/2)$ for all $t\in\mathbb{R}$ and so

$$\sum_{k\in[K]} I(\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*)$$
$$\leq \sum_{k\in[K]} I(\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*) - \mathbb{P}[\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*/2] + \mathbb{P}[\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*/2]$$
$$\leq \sum_{k\in[K]} \phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right) + \mathbb{P}[\langle\overline{\tilde{X}}_k-\mu^*,v\rangle > R_S^*/2]$$
$$\leq \sup_{v\in S}\left(\sum_{k\in[K]}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right)\right) + K\sup_{v\in S} H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right).$$

Next, we use several tools from empirical process theory and in particular, for a symmetrization argument, we consider a family of N independent Rademacher variables $(\epsilon_i)_{i=1}^N$ independent of the $(\tilde{X}_i)_{i=1}^N$. In *(bdi)* below, we use the bounded difference inequality (Theorem 6.2 in Boucheron et al. (2013)). In *(sa-cp)*, we use the symmetrization argument and the contraction principle (Chapter 4 in Ledoux and Talagrand (2011)) – we refer to the supplementary material of M. Lerasle and Lecué (2017) for more details. We have, with probability at least $1-\exp(-u)$,

$$\sup_{v\in S}\left(\sum_{k\in[K]}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right)\right)$$
$$\overset{(bdi)}{\leq} \mathbb{E}\sup_{v\in S}\left(\sum_{k\in[K]}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right) - \mathbb{E}\phi\left(\frac{\langle\overline{\tilde{X}}_k-\mu^*,v\rangle}{R_S^*}\right)\right) + \sqrt{2uK}$$
$$\overset{(sa-cp)}{\leq} \frac{4K}{NR_S^*}\mathbb{E}\sup_{v\in S}\langle v, \sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu^*)\rangle + \sqrt{2uK}$$
$$= \frac{4K}{\sqrt{N}R_S^*}\mathbb{E}\left\|\frac{1}{\sqrt{N}}\sum_{i\in[N]}\epsilon_i(\tilde{X}_i-\mu^*)\right\|_S + \sqrt{2uK}.$$

We therefore showed that under Assumption 7.1, with probability at least $1-\exp(-u)$, for all $\mu\in\mathbb{R}^d$, $|g_S^*(\mu)-\|\mu-\mu^*\|_S| \leq R_S^*$.

Now, if Assumption 7.2 holds then for all $v \in S$, we have from Markov's inequality that

$$H_{N,K,v}\left(\frac{R_S^*}{2}\sqrt{\frac{N}{K}}\right) \leq \frac{\mathbb{E}\langle \overline{\tilde{X}}_k - \mu, v\rangle^2}{(r_S^*/2)^2} = \frac{4Kv^\top \Sigma v}{N(r_S^*)^2} \leq \frac{4K \sup_{v\in S} \left\|\Sigma^{1/2}v\right\|_2^2}{N(r_S^*)^2} \leq \frac{1}{8}$$

and therefore (7.24) holds for $R_S^* = r_S^*$ when $|\mathcal{O}| < K/8$ and $u = K/128$. This proves the result of Lemma 7.1 for g_S^* under Assumption 7.2.

Finally, for the function f_S^* one needs to control the average of the $K/2$ inter-quartiles. One way to do it is to control the value of all elements $\langle \bar{X}_k - \mu^*, v\rangle$ in the inter-quartiles interval. This can be done by defining an R_S^* similar to the one in (7.24) but where the right-hand side value $1/2$ is replaced by $1/4$ in (7.24). This only modifies the absolute constants which are the one used in Lemma 7.1. ∎

Proof of Lemma 7.2. Unlike in Lemma 7.1 where we used the Rademacher complexities as a complexity measure, in this proof, the complexity measure we are using is the Vapnik and Chervonenkis (VC) dimension Vapnik and Chervonenkis (2015); Vapnik (2000) of a class $\mathcal{F}$ of Boolean functions, i.e. of functions from $\mathbb{R}^d$ to $\{0,1\}$ in our case. We recall that the Vapnik and Chervonenkis dimension of $\mathcal{F}$, denoted by $VC(\mathcal{F})$, is the maximal integer n such that there exists $x_1, \ldots, x_n \in \mathbb{R}^d$ for which the set $\{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F})\}$ is of maximal cardinality, that is of size 2^n. The VC dimension of the set of all indicators of half affine spaces in $\mathbb{R}^d$ is $d+1$ (see Example 2.6.1 in van der Vaart and Wellner (1996)). We also know (see, for instance, Chapter 3 in Koltchinskii (2011)) the following concentration bound: let $Y_1, \ldots, Y_n$ be independent random vectors in $\mathbb{R}^d$, there exists an absolute constant C_0 such that for all $u > 0$, with probability at least $1 - \exp(-u)$,

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}f(Y_i)\right) \leq C_0\left(\sqrt{\frac{VC(\mathcal{F})}{n}} + \sqrt{\frac{u}{n}}\right). \tag{7.28}$$

Lemma 7.2 is a corollary of a general result which holds under the only Assumption 7.1. This general result says that for all $u > 0$, with probability at least $1 - \exp(-u)$, for all $\mu \in \mathbb{R}^d$, $|g_S^*(\mu) - \|\mu - \mu^*\|_S| \leq R^\diamond$ where $R^\diamond$ is any point such that

$$C_0\left(\sqrt{\frac{d+1}{K}} + \sqrt{\frac{u}{K}}\right) + \sup_{\|v\|_2=1} H_{N,K,v}\left(R^\diamond\sqrt{\frac{N}{K}}\right) + \frac{|\mathcal{O}|}{K} < \frac{1}{2} \tag{7.29}$$

where C_0 is the constant from (7.28). In particular, when Assumption 7.3 holds then one can check that (7.29) holds for $R^\diamond = r^\diamond$ when $r^\diamond \leq c_0$ proving the result of Lemma 7.2. It only remains to show the general result. To that end we follow the same strategy as in the proof of Lemma 7.1 up to (7.27) (and with R_S^* replaced by $R^\diamond$). From that point, we use (7.28) and the VC dimension of the set of affine half spaces to get that with probability at least $1 - \exp(-u)$, for all $v \in S$,

$$\sum_{k\in[K]} I(\langle \overline{\tilde{X}}_k - \mu^*, v\rangle > R^\diamond) \leq H_{N,K,v}\left(R^\diamond\sqrt{\frac{N}{K}}\right) + C_0\left(\sqrt{\frac{d+1}{N/K}} + \sqrt{\frac{u}{N/K}}\right)$$

and so by definition of $R^\diamond$, on the same event, for all $v \in S$, $\sum_{k\in[K]} I(\langle \bar{X}_k - \mu^*, v\rangle > R^\diamond) < 1/2$. This concludes the proof. ∎

Bibliography

Ahsen, M. E. and Vidyasagar, M. (2019). An approach to one-bit compressed sensing based on probably approximately correct learning theory. *Journal of Machine Learning Research*, 20(11):1–23.

Allen-Zhu, Z., Gelashvili, R., and Razenshteyn, I. (2014). The restricted isometry property for the general p-norms. *arXiv:1407.2178*.

Allen-Zhu, Z., Lee, Y. T., and Orecchia, L. (2015). Using optimization to obtain a width-independent, parallel, simpler, and faster positive sdp solver.

Alon, N., Matias, Y., and Szegedy, M. (1999). The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

Arora, S., Hazan, E., and Kale, S. (2012). The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164.

Artstein, S., Milman, V., and Szarek, S. J. (2004). Duality of metric entropy. *Ann. of Math. (2)*, 159(3):1313–1328.

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794.

Barak, B., Hardt, M., and Kale, S. (2009). *The Uniform Hardcore Lemma via Approximate Bregman Projections*, pages 1193–1200.

Bárány, I. and Mustafa, N. H. (2020). An application of the universality theorem for tverberg partitions to data depth and hitting convex sets. *Computational Geometry*, page 101649.

Bartlett, P. L. and Mendelson, S. (2003). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482.

Bernholt, T. (2006). Robust Estimators are Hard to Compute. Technical Reports 2005,52, Technische Universität Dortmund, Sonderforschungsbereich 475: Komplexitätsreduktion in multivariaten Datenstrukturen.

Birgé, L. (1984). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.*, 20(3):201–223.

Blum, M., Floyd, R. W., Pratt, V. R., Rivest, R. L., and Tarjan, R. E. (1973). Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Bryc, W. o. (1995). *The normal distribution*, volume 100 of *Lecture Notes in Statistics*. Springer-Verlag, New York. Characterizations with applications.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Trans. Inform. Theory*, 59(11):7711–7717.

Cai, T. T., Liu, W., and Zhou, H. H. (2016). Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *Ann. Statist.*, 44(2):455–488.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185.

Catoni, O. and Giulini, I. (2017). Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. Technical report, CNRS and LSPM.

Chen, L. H. Y. and Shao, Q.-M. (2001). A non-uniform Berry-Esseen bound via Stein's method. *Probab. Theory Related Fields*, 120(2):236–254.

Chen, M., Gao, C., and Ren, Z. (2017). A general decision theory for huber's ϵ-contamination model.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960.

Chen, Y. and Shao, Q.-M. (2012). Berry-Esseen inequality for unbounded exchangeable pairs. In *Probability approximations and beyond*, volume 205 of *Lect. Notes Stat.*, pages 13–30. Springer, New York.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019a). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, Philadelphia, PA.

Cheng, Y., Diakonikolas, I., Ge, R., and Woodruff, D. (2019b). Faster algorithms for high-dimensional robust covariance estimation.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates.

Cherapanamjeri, Y., Hopkins, S. B., Kathuria, T., Raghavendra, P., and Tripuraneni, N. (2020a). Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 601–609, New York, NY, USA. Association for Computing Machinery.

Cherapanamjeri, Y., Tripuraneni, N., Bartlett, P. L., and Jordan, M. I. (2020b). Optimal mean estimation without a variance.

Chinot, G., Guillaume, L., and Matthieu, L. (2018). Statistical learning with lipschitz and convex loss functions. *arXiv preprint arXiv:1810.01090*.

Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using l1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198.

Dalalyan, A. S. and Minasyan, A. (2020). All-in-one robust estimator of the gaussian mean. *arXiv preprint arXiv:2002.01432*.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, 15(3):1269–1292.

de la Peña, V. H. and Giné, E. (1999). *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

Debruyne, M. (2009). An outlier map for support vector machine classification. *Ann. Appl. Stat.*, 3(4):1566–1580.

Depersin, J. (2020a). Robust subgaussian estimation with vc-dimension. *arXiv preprint arXiv:2004.11734*.

Depersin, J. (2020b). A spectral algorithm for robust regression with subgaussian rates. Technical report, CREST - ENSAE.

Depersin, J. and Lecué, G. (2020). Convex programs and algorithms for robust subgaussian estimation of a mean vector with respect to any norm. Technical report, IPParis, Crest, ENSAE.

Depersin, J. and Lecué, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019a). Robust Estimators in High-Dimensions Without the Computational Intractability. *SIAM J. Comput.*, 48(2):742–864.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*, pages 655–664. IEEE Computer Soc., Los Alamitos, CA.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017). Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2018a). Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. Society for Industrial and Applied Mathematics.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

Diakonikolas, I., Kane, D. M., and Pensia, A. (2020). Outlier robust mean estimation with subgaussian rates via stability. *arXiv preprint arXiv:2007.15618*.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2018b). List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. ACM, New York.

Diakonikolas, I., Karmalkar, S., Kane, D., Price, E., and Stewart, A. (2019b). Outlier-robust high-dimensional sparse estimation via iterative filtering.

Diakonikolas, I., Kong, W., and Stewart, A. (2019c). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, Philadelphia, PA.

Donoho, D. (1982a). *L.(1982) Breakdown properties of multivariate location estimators*. PhD thesis, Ph. D. Qualifying Paper, Dept. of Statistics, Harvard Univ.

Donoho, D. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA.

Donoho, D. L. (1982b). Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford ....

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827.

Du, S. S., Balakrishnan, S., and Singh, A. (2017). Computationally efficient robust estimation of sparse functionals.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.*, 6(6):899–929.

Gao, C. (2017). Robust regression via mutivariate regression depth.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pages 81–124.

Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145.

Grothendieck, A. (1953). Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. São Paulo*, 8:1–79.

Haldane, J. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.

Hampel, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 27:87–104.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Holmes, R. B. (2012). *Geometric functional analysis and its applications*, volume 24. Springer Science & Business Media.

Hopkins, S. B. (2018). Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*.

Hopkins, S. B., Li, J., and Zhang, F. (2020). Robust and heavy-tailed mean estimation made simple, via regret minimization. *arXiv preprint arXiv:2007.15839*.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101.

Huber, P. J. (1981). Robust statistics. *Wiley Series in Probability and Statistics*.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition.

Jerrum, M. R., Valiant, L. G., and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188.

Johnson, D. and Preparata, F. (1978). The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107.

Karnin, Z., Liberty, E., Lovett, S., Schwartz, R., and Weinstein, O. (2011). On the furthest hyperplane problem and maximal margin clustering.

Karnin, Z., Liberty, E., Lovett, S., Schwartz, R., and Weinstein, O. (2012). Unsupervised svms: On the complexity of the furthest hyperplane problem. In Mannor, S., Srebro, N., and Williamson, R. C., editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 2.1–2.17, Edinburgh, Scotland. PMLR.

Klartag, B. (2009). A Berry-Esseen type inequality for convex bodies with an unconditional basis. *Probab. Theory Related Fields*, 145(1-2):1–33.

Klartag, B. (2017). Super-Gaussian directions of random vectors. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 187–211. Springer, Cham.

Klartag, B. and Sodin, S. (2011). Variations on the Berry-Esseen theorem. *Teor. Veroyatn. Primen.*, 56(3):514–533.

Kolmogorov, A. N. (1959). Entropy per unit time as a metric invariant of automorphisms. In *Dokl. Akad. Nauk SSSR*, volume 124, pages 754–755.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *Information Theory, IEEE Transactions on*, 47:1902 – 1914.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, Berlin.

Koltchinskii, V. and Mendelson, S. (to appear). Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Notices.* arXiv:1312.3580.

Koltchinskii, V., Panchenko, D., Lozano, F., et al. (2003). Bounding the generalization error of convex combinations of classifiers: balancing the dimensionality and the margins. *The Annals of Applied Probability*, 13(1):213–252.

Komatu, Y. (1955). Elementary inequalities for Mills' ratio. *Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs.*, 4:69–70.

Kothari, P. K. and Steurer, D. (2017). Outlier-robust moment-estimation via sum-of-squares.

Lecué, G. and Lerasle, M. (2019). Learning from mom's principles: Le cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: theory and practice. *Ann. Statist.*, 48(2):906–931.

Lecué, G. and Mendelson, S. (2013). Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825.*

Lecué, G. and Mendelson, S. (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534.

Lecué, G. and Mendelson, S. (to appear). Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.* ArXiv:1401.2188.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.*, 48(2):906–931.

Lecué, G. and Mendelson, S. (2016). Regularization and the small-ball method i: sparse recovery.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

Ledoux, M. and Talagrand, M. (2011). *Probability in Banach spaces.* Classics in Mathematics. Springer-Verlag, Berlin. Isoperimetry and processes, Reprint of the 1991 edition.

Lei, Z., Luh, K., Venkat, P., and Zhang, F. (2020). A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 2598–2612.

Lepskiĭ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659.

Lerasle, M. (2019). Lecture notes: Selected topics on robust statistical learning theory.

Lerasle, M. and Oliveira, R. (2011). Robust empirical mean estimators. Technical report, IMPA and CNRS.

Lerasle, M., Szabo, Z., Mathieu, T., and Lecue, G. (2019). Monk – outlier-robust mean embedding estimation by median-of-means.

Li, J. (2017). Robust sparse estimation tasks in high dimensions.

Li, W. V. and Kuelbs, J. (1998). Some shift inequalities for Gaussian measures. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pages 233–243. Birkhäuser, Basel.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.*, 18(1):405–414.

Liu, R. Y. (1992). Data depth and multivariate rank tests. In L_1*-statistical analysis and related methods (Neuchâtel, 1992)*, pages 279–294. North-Holland, Amsterdam.

Liu, R. Y. and Singh, K. (1992). Ordering directional data: concepts of data depth on circles and spheres. *Ann. Statist.*, 20(3):1468–1484.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Lu, J., Han, F., and Liu, H. (2020). Robust scatter matrix estimation for high dimensional distributions with heavy tail. *IEEE transactions on information theory*.

Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments.

Lugosi, G. and Mendelson, S. (2017). Regularization, sparse recovery, and median-of-means tournaments.

Lugosi, G. and Mendelson, S. (2018). Near-optimal mean estimators with respect to general norms.

Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.*, 19(5):1145–1190.

Lugosi, G. and Mendelson, S. (2019b). Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields*, 175(3-4):957–973.

Lugosi, G. and Mendelson, S. (2019c). Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794.

M. Lerasle, T. Matthieu, Z. S. and Lecué, G. (2017). Monk – outliers-robust mean embedding estimation by median-of-means. Technical report, CNRS, University of Paris 11, Ecole Polytechnique and CREST.

Maronna, R. A., Martin, R. D., and J., Y. V. (2006a). *Robust statistics: theory and methods*. Wiley.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006b). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and methods.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

McDiarmid, C. (1997). Centering sequences with bounded differences. *Combinatorics, Probability and Computing*, 6(1):79–86.

Mendelson, S. (2017a). Extending the small-ball method.

Mendelson, S. (2017b). "Local" vs. "global" parameters—breaking the Gaussian complexity barrier. *Ann. Statist.*, 45(5):1835–1862.

Mendelson, S. (2017c). On multiplier processes under weak moment assumptions. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 301–318. Springer, Cham.

Mendelson, S. and Zhivotovskiy, N. (2018). Robust covariance estimation under $l_4 - l_2$ norm equivalence.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335.

Minsker, S. (2018a). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, 46(6A):2871–2903.

Minsker, S. (2018b). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

Minsker, S. and Strawn, N. (2017). Distributed statistical estimation and rates of convergence in normal approximation. Technical report, arXiv: 1704.02658.

Nagy, S., Schütt, C., Werner, E. M., et al. (2019). Halfspace depth and floating body. *Statistics Surveys*, 13:52–118.

Nemirovsky, A. S. and Yudin, D. B. a. (1983). *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Nesterov, Y. (1997). Semidefinite relaxation and non-convex quadratic optimization. *Optimization Methods and Software*, 12:1–20.

Oliveira, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3-4):1175–1194.

Peña, D. and Prieto, F. J. (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, 16(1):228–254.

Peng, R., Tangwongsan, K., and Zhang, P. (2012). Faster and simpler width-independent parallel algorithms for positive semidefinite programming.

Petrov, V. V. (1995). *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York. Sequences of independent random variables, Oxford Science Publications.

Pisier, G. (1989). *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge.

Pisier, G. (2012). Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society*, 49(2):237–323.

Prasad, A., Balakrishnan, S., and Ravikumar, P. (2019). A unified approach to robust mean estimation.

Prasad, A., Balakrishnan, S., and Ravikumar, P. (2020). A robust univariate mean estimator is all you need.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation.

Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145 – 147.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277.

Stahel, W. A. (1981). *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, ETH Zurich.

Talagrand, M. (1996). Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103.

Talagrand, M. (2014). *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg. Modern methods and classical problems.

Tsukuma, H. (2010). Proper bayes minimax estimators of the normal mean matrix with common unknown variances. *Journal of Statistical Planning and Inference*, 140(9):2596 – 2606.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, pages 448–485. Stanford Univ. Press, Stanford, Calif.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531.

Tyler, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.*, 22(2):1024–1044.

Van Aelst, S. (2016). Stahel–donoho estimation for high-dimensional data. *International Journal of Computer Mathematics*, 93(4):628–639.

Van Aelst, S., Vandervieren, E., and Willems, G. (2011). Stahel-donoho estimators with cellwise weights. *Journal of Statistical Computation and Simulation*, 81(1):1–27.

van de Geer, S. and Muro, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.*, 8(2):3031–3061.

van der Vaart, A. and Wellner, J. A. (2009). *A note on bounds for VC dimensions*, volume Volume 5 of *Collections*, pages 103–107. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

van Handel, R. (2016). *Probability in high dimension*.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. N. (2000). *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

Vapnik, V. N. and Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, Cham. Reprint of Theor. Probability Appl. **1**6 (1971), 264–280.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vidyasagar, M. (1997). *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer-Verlag, Berlin, Heidelberg.

Warren, H. E. (1968). Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178.

Wei, X. and Minsker, S. (2017). Estimation of the covariance structure of heavy-tailed distributions. In *NIPS*.

Wolf, L., Jhuang, H., and Hazan, T. (2007). Modeling appearances with low-rank svm. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6.

Zinodiny, S., Rezaei, S., and Nadarajah, S. (2017). Bayes minimax estimation of the mean matrix of matrix-variate normal distribution under balanced loss function. *Statistics and Probability Letters*, 125:110 – 120.

Zinodiny, S., Rezaei, S., and Nadarajah, S. (2018). Minimax estimation of the mean matrix of the matrix variate normal distribution under the divergence loss function. *Statistica*, 77(4):369–384.

Zuo, Y., Cui, H., and He, X. (2004a). On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Ann. Statist.*, 32(1):167–188.

Zuo, Y., Cui, H., and Young, D. (2004b). Influence function and maximum bias of projection depth based estimators. *Ann. Statist.*, 32(1):189–218.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Complexités statistiques et informatiques des problèmes d'estimation robustes en grandes dimensions.

Mots clés : Robustesse ; Grandes dimensions ; Algorithmes

Résumé : La théorie de l'apprentissage statistique vise à fournir une meilleure compréhension des propriétés statistiques des algorithmes d'apprentissage. Ces propriétés sont souvent dérivées en supposant que les données sous-jacentes sont recueillies par échantillonnage de variables aléatoires gaussiennes (ou sous-gaussiennes) indépendantes et identiquement distribuées. Ces propriétés peuvent donc être radicalement affectées par la présence d'erreurs grossières (également appelées "valeurs aberrantes") dans les données, et par des données à queue lourde. Nous sommes intéressés par les procédures qui ont de bonnes propriétés même lorsqu'une partie des données est corrompue et à forte queue, procédures que nous appelons *robustes*, que nous obtenons souvent dans cette thèse en utilisant l'heuristique Median-Of-Mean.
Nous sommes particulièrement intéressés par les procédures qui sont robustes dans des configurations à haute dimension, et nous étudions (i) comment la dimensionnalité affecte les propriétés statistiques des procédures robustes, et (ii) comment la dimensionnalité affecte la complexité computationnelle des algorithmes associés. Dans l'étude des propriétés statistiques (i), nous trouvons que pour une large gamme de problèmes, la complexité statistique des problèmes et sa "robustesse" peuvent être en un sens "découplées", conduisant à des limites où le terme dépendant de la dimension est ajouté au terme dépendant de la corruption, plutôt que multiplié par celui-ci. Nous proposons des moyens de mesurer les complexités statistiques de certains problèmes dans ce cadre corrompu, en utilisant par exemple la dimension VC. Nous fournissons également des limites inférieures pour certains de ces problèmes.
Dans l'étude de la complexité computationnelle de l'algorithme associé (ii), nous montrons que dans deux cas particuliers, à savoir l'estimation robuste de la moyenne par rapport à la norme euclidienne et la régression robuste, on peut relaxer les problèmes d'optimisation associés qui deviennent exponentiellement difficiles avec la dimension pour obtenir un algorithme tractable qui se comporte de manière polynomiale dans la dimension.

Title : Statistical and Computational Complexities of Robust and High-Dimensional Estimation Problems

Keywords : Robustness ; High Dimension ; Algorithms

Abstract : Statistical learning theory aims at providing a better understanding of the statistical properties of learning algorithms. These properties are often derived assuming the underlying data are gathered by sampling independent and identically distributed gaussian (or subgaussian) random variables. These properties can thus be drastically affected by the presence of gross errors (also called "outliers") in the data, and by data being heavy-tailed. We are interested in procedures that have good properties even when part of the data is corrupted and heavy-tailed, procedures that we call *robusts*, that we often get in this thesis by using the Median-Of-Mean heuristic.
We are especially interested in procedures that are robust in high-dimensional set-ups, and we study (i) how dimensionality affects the statistical properties of robust procedures, and (ii) how dimensionality affects the computational complexity of the associated algorithms. In the study of the statistical properties (i), we find that for a large range of problems, the statistical complexity of the problems and its "robustness" can be in a sense "decoupled", leading to bounds where the dimension-dependent term is added to the term that depends on the corruption, rather than multiplied by it. We propose ways of measuring the statistical complexities of some problems in that corrupted framework, using for instance VC-dimension. We also provide lower bounds for some of those problems.
In the study of computational complexity of the associated algorithm (ii), we show that in two special cases, namely robust mean-estimation with respect to the euclidean norm and robust regression, one can relax the associated optimization problems that becomes exponentially hard with the dimension to get tractable algorithm that behaves polynomially in the dimension.

Institut Polytechnique de Paris
91120 Palaiseau, France