Robust machine learning by median-of-means : theory and practice

G. Lecué and M. Lerasle

November 30, 2017

Abstract

We introduce new estimators for robust machine learning based on median-of-means (MOM) estimators of the mean of real valued random variables. These estimators achieve optimal rates of convergence under minimal assumptions on the dataset. The dataset may also have been corrupted by outliers on which no assumption is granted. We also analyze these new estimators with standard tools from robust statistics. In particular, we revisit the concept of breakdown point. We modify the original definition by studying the number of outliers that a dataset can contain without deteriorating the estimation properties of a given estimator. This new notion of *breakdown number*, that takes into account the statistical performances of the estimators, is non-asymptotic in nature and adapted for machine learning purposes. We proved that the breakdown number of our estimator is of the order of *number of observations * rate of convergence*. For instance, the breakdown number of our estimators for the problem of estimation of a d-dimensional vector with a noise variance σ^2 is $\sigma^2 d$ and it becomes $\sigma^2 s \log(ed/s)$ when this vector has only s non-zero component. Beyond this breakdown point, we proved that the rate of convergence achieved by our estimator is *number of outliers* divided by *number of observations*.

Besides these theoretical guarantees, the major improvement brought by these new estimators is that they are easily computable in practice. In fact, basically any algorithm used to approximate the standard Empirical Risk Minimizer (or its regularized versions) has a robust version approximating our estimators. On top of being robust to outliers, the "MOM version" of the algorithms are even faster than the original ones, less demanding in memory resources in some situations and well adapted for distributed datasets which makes it particularly attractive for large dataset analysis. As a proof of concept, we study many algorithms for the classical LASSO estimator. It turns out that the original algorithm can be improved a lot in practice by randomizing the blocks on which "local means" are computed at each step of the descent algorithm. A byproduct of this modification is that our algorithms come with a measure of *depth* of data that can be used to detect outliers, which is another major issue in Machine learning.

1 Introduction

Recent technological developments have allowed companies and state organizations to collect and store huge datasets. These yield amazing achievements in artificial intelligence such as self driving cars or softwares defeating humans in highly complex games such as chess or Go. In fact, most big organizations have realized that data will have a major role in the future economy. Most companies in banks, electric or oil companies, etc. have a digital branch developing new data-based services for customers and new companies even build all their business on data collected on Twitter, Facebook or Google.

Big datasets have also challenged scientists in statistics and computer science to develop new methods. In particular, Machine Learning has attracted a lot of attention over the past few years. As explained in [80], machine learning can be seen as a branch of statistical learning dealing with large datasets where any procedure besides presenting optimal statistical guarantees should be provided with at least a tractable algorithm for its practical implementation. This new constraint raised several interesting problems while revisiting the classical learning theory of Vapnik [95]. In particular, to name just a few, algorithms minimizing convex relaxations of the classical empirical $0-1$-risk counting the number of misclassification have been proposed and studied over the last few years. More generally, optimization algorithms are now routinely used and analyzed by statisticians.

1.1 Corruption of big datasets

Our theoretical understanding of many classical procedures of machine learning such as LASSO [85] heavily rely on two assumptions on the data, that should both be i.i.d. and have subgaussian behaviors. As noted in [7], "data from real-world experiments oftentimes tend to be corrupted with outliers and/or exhibit heavy tails". For example, in finance, heavy-tailed processes are routinely used and, in biology or medical experiments datasets are regularly subject to some corruption by outliers. These outliers are even in some applications the data of actual interests, one can think of fraud detections for example. The need for robust procedures in data science can be appreciated, for instance, by the recently posted challenges on "kaggle", the most popular data science competition platform. The 1.5 million dollars problem "Passenger Screening Algorithm Challenge" is about to find terrorist activity from 3D images. The "NIPS 2017: Defense Against Adversarial Attack" is about constructing algorithms robust to adversarial data.

1.1.1 Current ML solutions are highly sensitive to dataset's corruption

It is easy to check that most routinely used and theoretically understood algorithms break down when even a single "outlier" corrupts the dataset. This is the case of the standard least-squares estimator or its ℓ_1 penalized version known as LASSO for example as one can see from the simulation in Figure 1

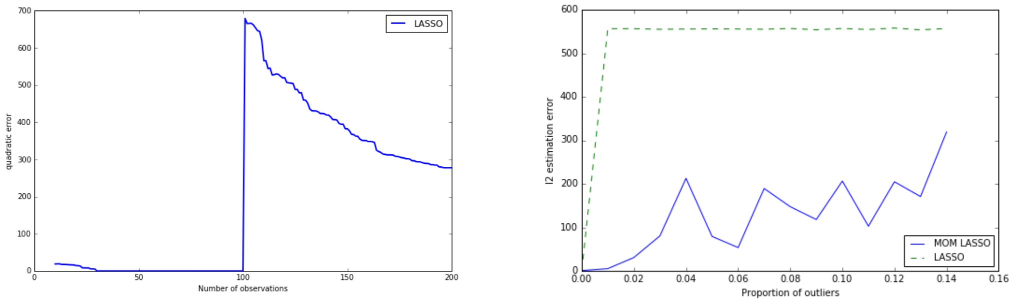

Figure 1: (left): Estimation error of the LASSO after one outliers was added at observation number 100. (right) Estimation error versus proportion of outliers for LASSO and its "Median of Means (MOM)" version.

LASSO is not an isolated example, most algorithms are actually designed to approximate minimizers of penalized empirical loss and empirical means of unbounded functions are highly sensitive to corrupted data. In statistics, building robust estimators has been an issue for a long time at least since the work of John Tukey, Frank Hampel and Peter Huber. It has yield interesting "robust" alternatives to the standard maximum likelihood estimator, one can refer to [41] for an overview on robust statistics and to Baraud, Birgé and Sart [10] or Baraud and Birgé [8] see also [20, 5] for deep new developments in robust statistics.

1.1.2 Resist or detect outliers : a raising challenge in ML

There are mainly two types of outliers in practice : those corrupting a dataset which are not interesting (outliers can appear in datasets due to storage issues, they can also be adversarial data as fake news, false declarative data, etc.) and those that are rare but important observations like frauds, terrorist activities, tumors in medical images,... A famous example of the latter type of "outlier" discovered unexpectedly was the ozone hole [62, p.2]. In the former case, the challenge is to construct predictions as sharp as if the dataset was clean and in the latter, the main question is to detect outliers.

Of course, any dataset is preprocessed to be "cleaned" from outliers. This step, called "data jujitsu", "data massage", "data wrangling" or "data munging" usually takes 80% of data scientists time. It requires an important quantity of expertise to guide data scientists [12]. Some platforms like "Amazon mechanical turk" allow to preprocess collectively datasets with millions of lines. Even when done with maximal care, some contamination always affect modern datasets, because outliers are particularly hard to detect in high dimension and because the distinction between outliers and informative data is often delicate in practice.

For these reasons, robustness has recently received a lot of attention in machine learning. Theoretical results have been obtained when "outliers" appear because the underlying distribution exhibits heavier tails than subgaussian distributions, producing anomalies. In this case, many attempts have been made to design estimators presenting subgaussian behaviors under minimal assumptions for this to make sense. This program was initiated by Catoni [20] for the estimation of the mean on the real line and Audibert and Catoni [5] in the least-squares regression framework. To the best of our knowledge, besides for the estimation of the mean of a real valued random variable where median-of-means estimators [2, 43, 73] provide a computationally tractable estimator satisfying optimal subgaussian concentration bounds [25], the other estimators are either suboptimal see for example [71] for the estimation of the mean of a vector in $\mathbb{R}^d$ or [18] for more general learning frameworks, or completely untractable see for example [61, 59, 50] for some examples in regression frameworks.

1.2 Robustness : state of the art

The book [62] traces back the origins of robust statistics to the fundamental works of John Tukey [86, 87], Peter Huber [39, 40] and Frank Hampel [33, 35]. Tukey asked the following question for the location problem of a Gaussian distribution "what happens if the distribution slightly deviates from the normal distribution?". Tukey's 1960 example shows the dramatic lack of distributional robustness of some classical procedures like Maximum Likelihood Estimator (MLE) [42, Example 1.1]. A related question was asked by Hodges [38] about "tolerance to extreme values".

In fact, one doesn't define one robustness property in general. Estimation, prediction or more generally any statistical properties like consistency, asymptotic normality, minimax optimality, etc. require assumptions (cf. the "no free lunch theorem" [24]). Such assumptions are naturally questionable on real databases. What properties of an estimator pertain when one or several assumptions are not satisfied? As there are several types of assumptions, one can define several types of robustness. For example, Tukey's question is about robustness to a misspecification of the Gaussian model while Hodges question can be understood as robustness to contamination of the dataset by large outliers or robustness to heavy-tailed distributions in the model that may have caused the appearance of large data.

1.2.1 Natural robustness properties of learning procedures

Robustness issues are particularly sensitive when data and risk are unbounded. To understand that, suppose first that the loss ℓ is the classical (bounded) $0-1$ loss in classification: $\ell_f(x,y) = \mathbf{1}_{y\neq f(x)}$. Assume that the dataset $\mathcal{D}$ is made of $N-|\mathcal{O}|$ i.i.d. data $(X_i,Y_i)_{i\in\mathcal{I}}$ and $|\mathcal{O}|$ "outliers" $(X_i,Y_i)_{i\in\mathcal{O}}$ that can be anything. The empirical risk

$$P_N\ell_f = \frac{1}{N}\sum_{i=1}^{N}\ell_f(X_i,Y_i)$$

satisfies for all $f:\mathcal{X}\to\{-1,1\}$ (where $\mathcal{X}$ is the space where the X_i take their values and the $Y_i\in\{-1,1\}$)

$$|(P_N-P_{\mathcal{I}})\ell_f| \leqslant \frac{2|\mathcal{O}|}{N}, \qquad \text{where} \qquad P_{\mathcal{I}}\ell_f = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\ell_f(X_i,Y_i) \ . \tag{1}$$

If the fraction $|\mathcal{O}|/N$ is small enough, the ERM over a class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\{-1,1\}$

$$\widehat{f}_N = \operatorname*{argmin}_{f\in\mathcal{F}} P_N\ell_f$$

performs theoretically as well as the ERM based only on the "good" data $(X_i,Y_i)_{i\in\mathcal{I}}$

$$\widehat{f}_{\mathcal{I}} = \operatorname*{argmin}_{f\in\mathcal{F}} P_{\mathcal{I}}\ell_f \ ,$$

which is statistically optimal in some problems [24]. In that case, the number of outliers $|\mathcal{O}|$ has to be less than N times the *rate of convergence of* $\widehat{f}_{\mathcal{I}}$ – this is a condition we will meet later for our estimators.

The problem with $\widehat{f}_N$ is that minimizing the empirical loss is NP-hard. To overcome this issue practitioners have introduced several convex relaxation of the $0-1$ loss. As a result, what is actually computed is a minimizer of empirical risk associated with an unbounded loss function ℓ (like the square, logistic, hinge or exponential losses) over an unbounded set of functions (as the set of linear functions $\{\langle \cdot, t\rangle, \ t \in \mathbb{R}^d\}$). This actual minimization problem is much more sensitive to outliers (no bounds can be provided between the empirical processes such as in (1)) and *practical* ERM (actually its proxies) are completely wrong even if the *theoretical* ERM performs well. In the following, we assume that $Y, \mathcal{F}$ are unbounded and ℓ is the (unbounded) square loss : $\ell_f(x,y) = (y - f(x))^2$ that is known to be sensitive to outliers.

Classical estimators, such as the maximum likelihood estimator (MLE) are designed for some very specific choice of statistical model and are therefore extremely model dependent : in the set of all Gaussian distributions with mean $\mu \in \mathbb{R}$ and variance $\sigma > 0$, the MLE of μ is the empirical mean while it is the empirical median in the set of all Laplace distributions with position parameters $\mu \in \mathbb{R}$. Tukey's example showed that MLE may perform very poorly under model misspecification. ML and learning theory tackle this issue by not assuming a statistical model on data. Instead of imposing strong assumptions allowing to fit all data distribution, learning procedures typically focus on a simpler task that can be achieved under much weaker restrictions on the randomness of data.

The typical example is as follows. A dataset $\mathcal{D} := \{(X_i, Y_i)_{i=1}^N\}$ of N independent and identically distributed (i.i.d.) variables with values in $\mathbb{R}^d \times \mathbb{R}$ is given. In statistics, the classical (parametric) approach is to start with a statistical model, i.e. a family $(P_\theta)_{\theta\in\Theta}$ indexed by some subset $\Theta \subset \mathbb{R}^d$ of distributions that should contain that of (X_1, Y_1), or at least that of $Y_1|X_1$. In the Gaussian linear regression model [83], one assumes that $Y_i = \langle X_i, t^*\rangle + \sigma\zeta_i$, where $t^* \in \mathbb{R}^{d-1}$ and $\sigma > 0$ are unknown parameters and $(\zeta_i)_{i=1}^N$ are i.i.d. standard Gaussian random variables independent of $(X_i)_{i=1}^N$. The problem is to estimate the unknown parameters $\theta = (t^*, \sigma) \in \mathbb{R}^d$ given the dataset $\mathcal{D}$ *assuming that the distribution of* $(Y_1|X_1)$ *belongs to the set of Gaussian distributions* $(P_\theta)_{\theta\in\Theta}$.

In Learning Theory, no assumption on the conditional expectation $\mathbb{E}[Y_1|X_1]$ nor any particular form for the conditional distribution of $Y_1|X_1$ is assumed. For instance, Y_i can be any function of X_i, it may also be independent of X_i. Given a new input X, one still wants to guess the associated output Y. To proceed, a class of functions F is proposed containing functions $f \in F$ used to guess the value $f(X)$ of Y. It is not assumed that the regression function $x \mapsto \mathbb{E}[Y|X = x] \in F$, we just hope that the ideal choice f^* in F allows to build a good guess $f^*(X)$ of Y. The learning task is to provide from the dataset a good approximation of f^*. If $f^*(X)$ is actually a good guess of Y, then so will be any good approximation of f^* applied in X, if not then we simply end up with a bad guess of Y. The function f^* providing the best guess $f^*(X)$ of Y is called an oracle and the difference $Y - f^*(X)$ is called the noise. Assume that $F = \{\langle \cdot, t\rangle : t \in \mathbb{R}^d\}$ is the class of linear functionals. If the regression function $x \mapsto \mathbb{E}[Y|X = x]$ is assumed to be in F then the tasks in Machine Learning and Statistics are essentially the same because the oracle equal the regression function. The difference is that learning procedures consider the estimation of f^* that may not be the regression function without assuming a parametric family of distributions for the noise. As such, the misspecification problem is naturally tackled in Machine Learning.

1.2.2 Robustness to heavy-tailed data

The Gaussian assumption on the noise $Y_1 - f^*(X_1)$ in statistics is usually weakened in learning theory to a subgaussian assumption (the L_p-moments of the noise are bounded by those of a Gaussian variable). This subgaussian assumption is already much more flexible and essentially allows to extend the use of least-squares estimators and their penalized versions to a much broader class of possible distributions for the noise. Going beyond the subgaussian assumption is in general way more technical. One way to proceed is to assume a subexponential assumption – random variables whose L_p moments are smaller than those of an exponential variable– as in the Bernstein's condition in [92]. More recently, the subexponential assumption has also been relaxed into L_p-moment assumptions. Least-squares estimators have been studied as well as some modifications less sensitive to large data for instance in [20, 5, 51, 54]. It appears that the statistical properties of least-squares estimators are deteriorated when the noise satisfies weaker assumptions [54]. A modification of the ERM based on Le Cam's estimation by tests principle [48, 49] has also been considered in [60, 50].

1.3 Robustness in regression

Most of the robustness literature of the 1980s was on the regression problem [22], [42, Section 7.12]. These methods make a distinction between the problem of robustness with respect to the outputs $(Y_i)_{i=1}^N$ on one side and with respect to the inputs $(X_i)_{i=1}^N$ on the other side [42, Chapter 7]. Outliers in the inputs remain extremely difficult to handle while many solutions have been proposed to deal with outliers in the outputs.

1.3.1 Outliers in the inputs in regression

Even if some solutions have been proposed in particular cases, see for example [97], the problem of robustness with respect to outliers in the inputs $X_1, \dots, X_N$ (also called *leverage points*), has not been solved in general neither in theory nor in practice. In the book [42], a three steps approach is suggested to solve the robustness problem in regression

1. build redundancy into the design matrix $\mathbb{X}$ if possible,
2. find routine methods when there are only few leverage points,
3. find analytical methods for identifying leverage points and, if possible, leverage groups.

In other words, [42] suggest to make an important "datacleaning" step before any analysis for solving the leverage point problem. As already discussed, this may not be possible for nowadays large datasets. Therefore, designing procedures that can handle outliers such as leverage points is an important task that requires new ideas.

In statistics, an elegant solution to this issue (and many others) called ρ-estimators [10, 9] has recently been proposed. However, ρ-estimators do not provide a satisfying solution for learning issues since they require a statistical model (even if almost no assumption besides a controlled massiveness are required on this model), they are specifically built for the Hellinger loss (although they seem to work extremely well for other risk function in some special cases) and they remain far from being computable in general which is dead-end for Machine Learning.

1.3.2 Outliers in the outputs

Robust procedures can be analyzed under the assumption that "we can act as if the x_{ij} are free of gross errors" [42, paragraph 3 in Section 7.3] (x_{ij} being the j-th coordinate of X_i for $i \in [N]$ and $j \in [d]$). In this approach, the square loss is replaced by a convex differentiable function ρ and the estimator minimizes

$$\sum_{i=1}^N \rho\left(Y_i - \left\langle X_i, t \right\rangle\right) \quad . \tag{2}$$

If $\psi = \rho'$, the estimator is also solution of the equation

$$\sum_{i=1}^N X_i \psi\left(Y_i - \left\langle X_i, t \right\rangle\right) = 0 \quad . \tag{3}$$

This estimator is robust when ψ is bounded since even gross errors in a few Y_i's don't affect the right hand side of (3). Classical examples of ψ functions include the Huber function $\psi(t) = t \min(1, c/|t|)$ or Tukey's bisquare function $\psi(t) = t(1-(t/c)^2)^2 I(|t| \leqslant c)$ for all $t \in \mathbb{R}$. In both cases, a tuning parameter c needs to be specified in advance to ultimately balance between resistance to outliers and bias of the estimation. Earlier regression estimate go back to Laplace and its Least Absolute Deviation (LAD) or L_1-estimate that minimizes $\sum_i |Y_i - \langle X_i, t\rangle|$ yielding the median in the location problem [26]. These M-estimation strategies work well both in theory and practice [42, 62], but fail totally when even one X_i has a gross error. In the regression problem $Y_i = \langle X_i, t^*\rangle + \xi_i$ where $\mathbb{X}^\top \mathbb{X} = I_{d\times d}$ and $\xi_1, \dots, \xi_N$ are i.i.d. random errors such that $\mathbb{E}\psi(\xi_i) = 0$, asymptotic results for M-estimators can be found in [42] when d is fixed and N goes to infinity. When d is allowed to grow with N, some results have also been obtained but, as noted in [42, p168], under conditions that are, for all "*practical purposes worthless*" since "*already for a moderately large number of parameters, we would need an impossibly large number of observations*". When the inputs are clean of outliers, subgaussian random variables, or when there is no design, non asymptotic results on minimizers of the Huber loss have also been obtained recently in the small and large dimension settings [30, 75].

Many alternative robust procedures can be found in the regression literature, among other, one can mention the least median of squared residual estimator [76] and the least trimmed sum of squares estimator [77] which minimizes the sum of the $N/2+1$ smallest squared residuals. Finally, [36] introduced Least Median of Squares in regression which naturally extends the median estimator in the location problem :

$$\hat{t} \in \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}} \operatorname{Median}\left(Y_1 - \langle X_1, \theta\rangle, \cdots, Y_N - \langle X_N, \theta\rangle\right) .$$

Its computational complexity rises exponentially with d [37, p.330], [63].

1.3.3 Other challenges

We discussed the problems of detecting outliers and resisting corruption of the inputs earlier. Besides these, robust statistic has left some interesting open questions that have to be addressed.

Most results in robust statistics are asymptotic but the asymptotic approach may be quite misleading : "*[...] approaches based on asymptotics are treacherous, and it could be quite misleading to transfer insights gained from asymptotic variance theory or from infinitesimal approaches (gross error sensitivity and the like) by heuristics to leverage points with* $h_i > 0.2$" [42, p.189]. In other words, robustness issues ultimately have to be tackled from a non-asymptotic point of view. Again, ρ-estimators [10, 9] treat this issue from a statistical perspective but cannot be applied in our learning setting.

The actual solution to get rid of some data a priori and work on what is supposed to be clean data seems quite dangerous. First, if "outliers" are due to heavy-tailed data, this can induce some bias in the estimation. Second, data can be typical for the estimation of some means $P\ell_f$ but not for other choices of loss function ℓ' and/or function f'. An estimator $\widehat{P\ell_f}$ may even get better if one uses some data while another $\widehat{P\ell_f}'$ would get worse. The definition of outliers is thus not clear, it may depend on the estimators, the risk function and the parameter of interest. Therefore, if some subsampling is performed to get more robustness, it should be done with respect to the loss ℓ and point f and not on systematic preprocessing algorithms blind to the specific learning problem we want to solve.

When designing robust strategies for ML, one should keep in mind that these should be computationally tractable. To the best of our knowledge, no estimator robust to outlier in the outputs nor any general learning procedure robust to heavy-tailed data can currently be used in a high dimensional learning problems. *There is currently no available algorithm for high dimensional learning whose good statistical performance don't break down in the presence of one outlier in the dataset.*

1.4 Contributions

In this paper, we propose a new estimator of the learning function

$$f^* = \underset{f\in F}{\operatorname{argmin}}\, P[(Y - f(X))^2] \ .$$

This new estimator has minimax-optimal non-asymptotic risk bounds even when the dataset has been corrupted by outliers (that may have corrupted indifferently the inputs, the outputs or both of them) and when informative data (those that are not outliers) only satisfy somehow minimal assumptions relating their L_2 geometry on F to that of P (see Sections 1.4.1 and 3.1 for details). This estimator solves therefore all challenges related to outliers in the learning theory literature including "adversarial outliers" – which are data satisfying no assumptions – and "stochastic outliers" – which are informative data that may look like outliers because of some weak concentration properties due to heavy tail distributions.

We also show that our estimator can be used for ML purposes by providing several algorithms to compute proxies of our estimators and illustrate their performance in an extensive simulation study (see Section A). This opens several interesting conjectures for example, to study and compare theoretically convergence of these algorithms, and to automatically calibrate our estimators.

1.4.1 Going beyond the i.i.d. setup: the $\mathcal{O}\cup\mathcal{I}$ framework

We propose the following setup that allows the presence of outliers and relax the standard i.i.d. setup. We assume that data are partitioned in two groups (unknown from the statistician), more precisely $\{1,\ldots,N\}=\mathcal{O}\cup\mathcal{I}$, with $\mathcal{O}\cap\mathcal{I}=\emptyset$

- data $(X_i,Y_i)_{i\in\mathcal{O}}$ ($\mathcal{O}$ stands for "outliers") are not assumed to be independent nor independent to the other ones $(X_i,Y_i)_{i\in\mathcal{I}}$, they may not be identically distributed, in fact *nothing is required on these data*; they can even be adversarial in the sense that may have been designed to make the learning task harder,

- data $(X_i,Y_i)_{i\in\mathcal{I}}$ (called *informative*) are the only data on which we can rely on to solve our learning task. In what follows, we only need very weak assumptions on those data to make MOM estimators achieving optimal statistical properties. In particular, we will always assume that the informative data are independent.

As we allow the presence of outliers, our procedure will be in a sense robust to the i.d. (identically distributed) assumption. This robustness goes actually further, since even informative data won't be assumed identically distributed. Instead, our procedure will be shown to work under the much weaker requirement that the distributions of (X_i,Y_i) for all $i\in\mathcal{I}$ induce an equivalent L_2 metric over the class F and equivalent covariance between functions in F and the output. If various distributions have equivalent first and second moments for all functions $f\in F$, our estimators will exhibit the same optimal performance as if they were i.i.d. Gaussian with similar first and second moments on all $f\in F$. Our new assumption feels more natural than the usual i.i.d one as distributions inducing the same "risk geometry" on F seem interesting to solve our learning task and any other higher moment assumption above 2 seems unnatural. However, the standard ERM once again would fail to achieve this goal as the behavior of the supremum over subclasses of F of the empirical process used to bound the risk of the ERM depends in general on higher moments of these distributions.

1.4.2 Quantifying robustness by breackdown point

The breakdown point of an estimator [32, 29, 42] is the smallest proportion of corrupted observations necessary to push an estimator to infinity [28, p.1809], [42, p.279]. The notion has been introduced by Hampel [32] in his 1968 Ph. D. thesis. In [29, 27], the notion was revisited with the following non-asymptotic definition, see also [81, 34]. Let $\mathcal{D}_\mathcal{I}$ denote a dataset and let T be an estimator. If there exists a strategic / malicious / adversarial choice of another dataset $\mathcal{D}_\mathcal{O}$ such that $T(\mathcal{D}_\mathcal{I}\cup\mathcal{D}_\mathcal{O})-T(\mathcal{D}_\mathcal{I})$ is arbitrarily large, the estimator is said to break down under the contamination fraction $|\mathcal{D}_\mathcal{O}|/(|\mathcal{D}_\mathcal{I}|+|\mathcal{D}_\mathcal{O}|)$. The **Donoho-Hampel-Huber breakdown point** $\epsilon^*(T,\mathcal{D}_\mathcal{I})$ is then defined as the smallest contamination fraction under which estimator T breaks down:

$$\epsilon^*(T,\mathcal{D}_\mathcal{I})=\min_{m\in\mathbb{N}}\left\{\frac{m}{|\mathcal{D}_\mathcal{I}|+m}:\sup_{\mathcal{D}_\mathcal{O}:|\mathcal{D}_\mathcal{O}|=m}|T(\mathcal{D}_\mathcal{I}\cup\mathcal{D}_\mathcal{O})-T(\mathcal{D}_\mathcal{I})|=\infty\right\}. \tag{4}$$

For the estimation of the mean μ of a random variable from N i.i.d. observations, the empirical mean can be made arbitrarily large by adding a single "bad" observation. The breakdown point of the empirical mean estimator is therefore $1/(N+1)$ and this is the worst possible value for an estimator. The empirical mean is not robust to outliers, it performs very poorly in corrupted environments. On the other hand, the empirical median has a breakdown value of $1/2$ for the location problem.

Constant estimators have breakdown points equal to 1. To avoid these degenerate examples, one restricts the set of estimators to consider optimality from the breakdown point of view. For instance, in the regression problem, one can consider **equivariant** estimators [22, 21] and [62, p.92], that is estimators T invariant by affine transformation of the data: for all $a \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$,

$$T((X_i, \lambda(Y_i + \langle X_i, a\rangle))_{i=1}^N) = \lambda \left(T((X_i, Y_i)_{i=1}^N) + a\right) . \tag{5}$$

Among equivariant estimators, the common sense heuristic applies : "when more than half data are bad, one cannot learn from good ones" [21]. In particular, among equivariant estimators, $1/2$ is the best possible breakdown point for an estimator and the empirical median is optimal from this perspective. When the parameter of interest belongs to a bounded set, one can adapt the definition of breakdown point as in the discussion papers of [21]. Other generalization to quantile estimation can be found in [78].

Similar results are much harder to obtain in higher dimensions, the main reason being the absence of a natural notion of median or quantiles in dimension $d \geqslant 2$. This problem gave birth to the construction of *depths*, the most famous one being Tukey's depth [89, 88]. The notions of median, trimmed mean and covariance resulting from Tukey's depth yield robust estimators in high dimensions. For instance, in [27, 28], the depth-trimmed mean and the deepest point are affine invariant estimators with breakdown points $1/3$. It is even possible to achieve a breakdown point for the location problem in $\mathbb{R}^d$ close to $1/2$ by using a weight function based on Tukey's depth [82, 27].

1.4.3 A modified notion of breakdown point for ML

We introduce a new concept of breakdown point for Learning. The standard breakdown point (4) does not come with any estimation or prediction property. As already mentioned, this leads to consider restricted classes of estimators to study optimality from the breakdown point of view to avoid degenerate cases. Instead of imposing algebraic restrictions, our new definition focus on the risk of the procedures.

Definition 1. *Let $\delta \in (0,1)$, $\mathcal{R} > 0$, $N \geqslant 1$, F be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{P}$ be a set of distributions on $\mathcal{X} \times \mathbb{R}$. Let $T : \cup_{n\geqslant 1}(\mathcal{X} \times \mathbb{R})^n \to F$ denote an estimator and let $\mathcal{D} = \{(X_i, Y_i)_{i=1}^N\}$ be a dataset made of N i.i.d. random variables with a common distribution in $\mathcal{P}$. For any $P \in \mathcal{P}$, let $f_P^* \in \operatorname{argmin}_{f\in F} \mathbb{E}_{(X,Y)\sim P}[(Y - f(X))^2]$. The breakdown number of the estimator T on the class $\mathcal{P}$ at rate $\mathcal{R}$ with confidence δ is*

$$K^*_{ML}(T, N, \mathcal{R}, \delta, \mathcal{P}) = \min\left\{k \in \mathbb{Z}_+ : \inf_{P\in\mathcal{P}} \mathbb{P}_{\mathcal{D}\sim P^{\otimes N}}\left(\sup_{|\mathcal{O}|=k} \|T(\mathcal{D}\cup\mathcal{O}) - f_P^*\|_{L^2(P_X)} \leqslant \mathcal{R}\right) \geqslant 1-\delta\right\}$$

where P_X denotes the marginal on $\mathcal{X}$ of P.

In words, the breakdown number is the minimal number of points one has to add to a dataset to break statistical performance of an estimator T. In this definition, the rate $\mathcal{R}$ one can achieve is of particular importance. Ultimately, we would like to take it as small as possible. To that end, recall the definition of a minimax rate of convergence.

Definition 2. *Let $\delta > 0$, $N \geqslant 1$, $\sigma > 0$, F be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and, for any $f \in F$, let P_f denote the distribution of (X, Y) when X is a standard Gaussian process on $\mathcal{X}$, ξ is a standard Gaussian random variable on $\mathbb{R}$ independent of X and $Y = f(X) + \sigma\xi$. For any $f \in F$, denote by $\mathbb{P}_f = P_f^{\otimes N}$*

the distribution of a sample $(X_i, Y_i)_{i=1,\dots,N}$ of N i.i.d. random variables distributed according to P_f. Any estimator $\widehat{f}_N = T((X_i, Y_i)_{i=1,\dots,N})$ is said to perform with accuracy $\mathcal{R}$ and confidence δ uniformly on F if

$$\forall f \in F, \qquad \mathbb{P}_f \left(\left\| \widehat{f}_N - f \right\|_{L^2(P_X)}^2 \leqslant \mathcal{R} \right) \geqslant 1 - \delta \ .$$

The minimax accuracy $\mathcal{R}(\delta, F)$ is the smallest accuracy that can be achieved uniformly over F with confidence δ by some estimator.

The minimax rate of convergence defined in Definition 2 is obtained in the "Gaussian model" (that is when the dataset is made of N i.i.d. observations in the regression model with Gaussian design and Gaussian noise independent of the design) which is somehow the benchmark model in statistics in which one can derive minimax rates of convergence over F. We will therefore use the minimax rates obtained in this model as benchmark rates of convergence.

For any class $\mathcal{P}$ containing the Gaussian model indexed by F, that is $(P_f)_{f\in F}$ (we shall always work under this assumption in the following) and any $\mathcal{R} < \mathcal{R}(\delta, F)$, it is clear that $K^*_{\text{ML}}(T, N, \mathcal{R}, \delta, \mathcal{P}) = 0$ as no estimator can even achieve the rate $\mathcal{R}$ even on datasets containing no outliers. The notion is therefore interesting only for rates $\mathcal{R} \geqslant \mathcal{R}(\delta, F)$ and we shall be particularly interested by rates $\mathcal{R} = C\mathcal{R}(\delta, F)$ for some absolute constant $C \geqslant 1$ since estimators with $K^* = K^*_{\text{ML}}(T, N, \mathcal{R}, \delta, \mathcal{P}) > 0$ are minimax (therefore statistically optimal) even when the dataset has been corrupted by up to K^* outliers. Furthermore, we will be interested in classes $\mathcal{P}$ much larger than $(P_f)_{f\in F}$, typically, one would like $\mathcal{P}$ to be the class of all distributions P of (X, Y) such that Y and all $f(X)$ with $f \in F$ have finite second order moments since these are the minimal properties giving sense to the square risk $\mathbb{E}[(Y - f(X))^2]$. Finally, the following link between the classical breakdown point and the new breakdown number always holds : for any $\delta > 0$, $\mathcal{R} > 0$, any class $\mathcal{P}$, any dataset $\mathcal{D}$ and any estimator T, one has

$$\epsilon^*(T, \mathcal{D}) \geqslant \frac{1 + K^*_{\text{ML}}(T, |\mathcal{D}|, \mathcal{R}, \delta, \mathcal{P})}{1 + K^*_{\text{ML}}(T, |\mathcal{D}|, \mathcal{R}, \delta, \mathcal{P}) + N} \ .$$

Using this new definition, one can be interested in various existence or optimality properties such as:

1. Given $0 < \delta < 1$, $C > 0$, what is the largest set of sample distributions $\mathcal{P}_{\delta,C}$ where one can build estimators with accuracy $C\mathcal{R}(\delta, F)$ of the same order as in the Gaussian model ?

2. given $0 < \delta < 1$, $\mathcal{R} > \mathcal{R}(\delta, F)$ and $\mathcal{P} \subset \mathcal{P}_{\delta,C}$, is there some procedure T such that $K^*_{\text{ML}}(T, N, \mathcal{R}, \delta, \mathcal{P}) > 0$?

3. given $0 < \delta < 1$, $\mathcal{R} > \mathcal{R}(\delta, F)$ and $\mathcal{P} \subset \mathcal{P}_{\delta,C}$, how large can be $\sup_T K^*_{\text{ML}}(T, N, \mathcal{R}, \delta, \mathcal{P})$? and is this maximum achievable by a computationally tractable estimator T?

Most recent theoretical works on robust learning approaches focus on the first question, showing that statistical optimality can actually be achieved over much broader classes of distributions than the Gaussian model, see [60, 59] for example. In this paper, we build on the approach introduced in [50] where we proved that, for any δ smaller than $\exp(-CNr_N^2)$, there exists estimators such that $K^*_{\text{ML}}(T, N, r_N^2, \delta, \mathcal{P}) \geqslant CNr_N^2$ where r_N^2 is the minimax rate of convergence in the Gaussian model for exponential deviation. It is for instance given by $r_N^2 = \sigma^2 d/N$ for the problem of estimation of a d-dimensional vector and $r_N^2 = \sigma^2 s \log(ed/s)/N$ when this vector has only s non-zero coordinates. But we go further in this paper, showing that these performance can be achieved by a procedure that is computationally tractable, hence by a Machine Learning algorithm. This ultimate breakthrough is due to a new estimator that we shall now introduce.

1.4.4 Our estimators

The key equation to understand our procedure is that the target

$$f^* = \operatorname*{argmin}_{f\in F} \mathbb{E}[(Y - f(X))^2]$$

can be obtained as the solution of the following minimaximization problem

$$f^* = \operatorname*{argmin}_{f\in F} \sup_{g\in F} \mathbb{E}[(Y - f(X))^2 - (Y - g(X))^2] \enspace .$$

This remark is obvious since the expectation operator is linear, it is fundamental though in our construction as we use non-linear estimators of the expectation. More precisely, unknown expectations are estimated by median-of-means (MOM) estimators [2, 43, 73] : given a partition of the dataset into blocks of equal sizes, MOM is the median of the empirical means over each block, see Section 2.2 for details. MOM estimators are naturally more robust than the empirical mean thanks to the median step : if the number of blocks is at least twice that of the outliers, outliers can only affect less than half blocks so MOM remains an acceptable estimator of the mean. MOM's estimators are non linear, therefore plugging a MOM estimator into the minimization problem doesn't yield the same estimator as the plugging estimator on the minimaximization problem. More precisely, it is not so hard to see that MOM minimizers cannot achieve (fast) minimax rates $\mathcal{R}(\delta, F)$. This is why we focus in the following on the minimaximization problem

MOM estimators of the increments of criteria have already been used in [60, 59, 50]. What is new here is that, instead of combining these estimators to build a "tournament" as in [60, 59] or using the approach of Le Cam as in [50], we simply plug these estimators in the minimaximization problem. Our construction is therefore extremely simplified compared to these previous papers, we'll show that our new estimator achieves the same theoretical properties as these alternatives.

The main difference is that this new estimator is easy to implement. Given (f_t, g_t) and $\ell_f(x, y) = (y - f(x))^2$, one can find a median block B_{med} such that

$$P_{B_{\text{med}}}[\ell_{f_t} - \ell_{g_t}] = \text{median}(P_{B_k}[\ell_{f_t} - \ell_{g_t}],\ k \in \{1, \dots, K\}) \enspace .$$

Then one can perform a descent/ascent algorithm over the block B_{med} to get the next iteration (f_{t+1}, g_{t+1}). In practice though, this basic idea can be substantially improved by shuffling the blocks at each time step, cf. Section A.8. The first advantage of this shuffling step is that the algorithm doesn't converge to local minimaxima.

Furthermore, our algorithm defines a natural notion of depth of data : deep data are typically regularly chosen as member of the median block B_{med} while "outliers" on the other hand are typically left aside. This notion of depth, based on the risk function, is natural in our learning framework and should probably be investigated more carefully in future works. It also suggests an empirical definition of outliers and therefore an outliers detection algorithm as a by-product.

The paper is organised as follow. Section 2 introduces formally the classical least-squares learning framework and presents our new estimator, Section 3 details our main theoretical results, that are proved in Section 4. An extended simulation study is provided in Section A, including the presentation of many robust versions of standard optimization algorithms. Theoretical abstract rates derived in the main results are particularized in Section B into important problems in Machine Learning to show the extent of our results. These main results focus on penalized version of the basic tests presented in this introduction that are well suited for high dimensional learning frameworks, we complete these results in Section C, providing results for the basic estimators without penalizations in small dimension. Finally, Section D provides some optimality results for our procedures.

2 Setting

Let $\mathcal{X}$ denote a measurable space and let $(X, Y), (X_i, Y_i)_{i\in[N]}$ denote random variables taking values in $\mathcal{X} \times \mathbb{R}$. Let P denote the distribution of (X, Y) and, for $i \in [N]$, let P_i denote the distribution of (X_i, Y_i).

Let F denote a convex class of functions $f : \mathcal{X} \to \mathbb{R}$ and suppose that $F \subset L_P^2$, $\mathbb{E}[Y^2] < \infty$. For any $(x, y) \in \mathcal{X} \times \mathbb{R}$, let $\ell_f(x, y) = (y - f(x))^2$ denote the square loss function and let f^* denote an oracle

$$f^* \in \operatorname*{argmin}_{f \in F} P\ell_f \qquad \text{where} \qquad \forall g \in L_P^1,\ Pg = \mathbb{E}[g(X, Y)] \ .$$

For any $Q \in \{P, (P_i)_{i \in [N]}\}$ and any $p \geqslant 1$, let $\|f\|_{L_Q^p} = (Q|f|^p)^{1/p}$ the L_Q^p-norm of f whenever it's defined. Finally, let $\|\cdot\|$ be a norm defined on the span of F; $\|\cdot\|$ will be used as a regularization norm.

2.1 Minimaximization

Since f^* minimizes $f \to P\ell_f$ and the distribution P is unknown, it is estimated in empirical risk minimization [96, 92, 44] by $\widehat{f}_{\text{ERM}}$ minimizing $f \mapsto P_N\ell_f$, where for any function g, $P_N g = N^{-1}\sum_{i=1}^N g(X_i, Y_i)$. This approach works well when $P_N\ell_f$ is close to $P\ell_f$ for all functions in F. This uniform proximity requires strong concentration properties. Moreover, a single outlier can break down all estimators $P_N\ell_f$.

A key observation in our analysis is that f^* is solution of a minimaximization problem:

$$f^* = \operatorname*{argmin}_{f \in F} P\ell_f = \operatorname*{argmin}_{f \in F} \sup_{g \in F} P(\ell_f - \ell_g) \ . \tag{6}$$

To estimate the unknown expectations $T_{\text{id}}(g, f) = P(\ell_f - \ell_g)$, we use a Median-Of-Means (MOM) estimator of the expectation. As this estimator is non linear, plugging a MOM estimator in the minimaximization problem or in the minimization problem in (6) does not yield the same estimator.

The minimaximization point of view has been considered in [5] and [10, 8]. These papers used bounded modification of the quadratic loss. As for classification, the impact of outliers on the empirical mean is negligible if the loss is bounded and the number of outliers controlled, which ensures robustness. The cost to pay is that the function to be minimized is not convex and is therefore hard to implement. The idea of estimation by aggregation of tests is due to Le Cam [48, 49] and was further developed by Birgé [15], Baraud [11], Baraud, Birgé and Sart [10]. MOM estimators have been recently considered see [60, 50, 59, 61] and Section 2.2 for details. Nevertheless, to the best of our knowledge, this paper is the first to consider the estimators obtained by plugging MOM's estimators in the minimaximization problem (6). This approach will proved to be efficient for designing algorithms.

2.2 MOM estimator

Let K denote an integer smaller than N and let $B_1, \ldots, B_K$ denote a partition of $[N]$ into blocks of equal size N/K (w.l.o.g. we assume that K divides N and that B_1 is made of the first N/K data, B_2 of the next N/K data, etc.). For all function $\mathcal{L} : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and $k \in [K]$, let $P_{B_k}\mathcal{L} = |B_k|^{-1}\sum_{i \in B_k} \mathcal{L}(X_i, Y_i)$. Then $\text{MOM}_K(\mathcal{L})$ is a median of the set of K real numbers $\{P_{B_1}\mathcal{L}, \cdots, P_{B_K}\mathcal{L}\}$.

We will make an extensive use of empirical medians and quantiles in the following. We now precise some conventions used repeatedly hereafter. For all $\alpha \in (0, 1)$ and real numbers $x_1, \ldots, x_K$, we denote by

$$\mathcal{Q}_\alpha(x_1, \ldots, x_K) = \{u \in \mathbb{R} : \quad |\{k \in [K] : x_k \geqslant u\}| \geqslant (1 - \alpha)\ell, \quad |\{k \in [K] : x_k \leqslant u\}| \geqslant \alpha\ell\} \ .$$

Any element in $\mathcal{Q}_\alpha(x)$ is a $(1 - \alpha)$-empirical quantile of the vector $x_1, \ldots, x_K$. Hereafter, $Q_\alpha(x)$ denotes an element in $\mathcal{Q}_\alpha(x)$. For all $x = (x_1, \ldots, x_K)$, $y = (y_1, \ldots, y_K)$ and $t \in \mathbb{R}$,

$$\begin{aligned} Q_\alpha(x) \geqslant t \qquad &\text{iff} \qquad \sup \mathcal{Q}_\alpha(x) \geqslant t \ , \\ Q_\alpha(x) \leqslant t \qquad &\text{iff} \qquad \inf \mathcal{Q}_\alpha(x) \leqslant t \ , \\ z = Q_\alpha(x) + Q_\alpha(y) \qquad &\text{iff} \qquad z \in \mathcal{Q}_\alpha(x) + \mathcal{Q}_\alpha(y) \end{aligned}$$

where in the last inequality we use the Minkowsky sum of two sets. More generally, inequalities involving empirical quantiles are always understood in the worst possible case. For example, one can check that, for all vectors x and y

$$\sup \mathcal{Q}_{1/4}(x) + \sup \mathcal{Q}_{1/4}(y) \leqslant \inf \mathcal{Q}_{1/2}(x + y) \ .$$

In the following, we will simply write with some abuse of notation that for all vectors x and y:

$$Q_{1/4}(x) + Q_{1/4}(y) \leqslant Q_{1/2}(x+y) \enspace .$$

Likewise, one can check that $Q_{1/2}(x-y) \leqslant Q_{3/4}(x) - Q_{1/4}(y)$ and $Q_{1/2}(x) \geqslant -Q_{1/2}(-x)$. Moreover, to prove that $Q_\alpha(x) \geqslant t$ it is enough to prove that $\exists J \subset [K], |J| \geqslant (1-\alpha)K, \forall k \in J, x_k \geqslant t$ and to prove that $Q_\alpha(x) \leqslant t$ it is enough to prove that $\exists J \subset [K], |J| \geqslant \alpha K, \forall k \in J, x_k \leqslant t$.

We conclude this section with the definition of MOM tests and their regularized versions.

Definition 3. *Let $\alpha \in (0,1)$ and $K \in [N]$. For all functions $\mathcal{L} : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ the **α-quantile on K blocks of $\mathcal{L}$** is $Q_{\alpha,K}(\mathcal{L}) = Q_\alpha((P_{B_k}\mathcal{L})_{k\in[K]})$. In particular, the Median-of-Means (MOM) of $\mathcal{L}$ on K blocks is defined as $MOM_K(\mathcal{L}) = Q_{1/2,K}(\mathcal{L})$. For all $f, g \in F$, the **MOM test on K blocks of g against f** is defined by*

$$T_K(g,f) = MOM_K(\ell_f - \ell_g)$$

and, for a given regularization parameter $\lambda \geqslant 0$, its regularized version is

$$T_{K,\lambda}(g,f) = MOM_K(\ell_f - \ell_g) + \lambda(\|f\| - \|g\|) \enspace .$$

2.3 Minimaximization of MOM tests

We are now in position to define our basic estimators. These are simply obtained by replacing the unknown expectations $P(\ell_f - \ell_g)$ in (6) by the (regularized version) of the MOM tests on K blocks of g against f,

Definition 4. *For any $K \in [N/2]$, let*

$$\hat{f}_K \in \operatorname*{argmin}_{f\in F} \max_{g\in F} T_K(g,f) \text{ and } \hat{f}_{K,\lambda} \in \operatorname*{argmin}_{f\in F} \max_{g\in F} T_{K,\lambda}(g,f). \tag{7}$$

One can rewrite our estimators as cost minimization estimators

$$\hat{f}_K \in \operatorname*{argmin}_{f\in F} \mathcal{C}_K(f) \text{ and } \hat{f}_{K,\lambda} \in \operatorname*{argmin}_{f\in F} \mathcal{C}_{K,\lambda}(f)$$

where for all $f \in F$, $\mathcal{C}_K(f) = \sup_{g\in F} T_K(g,f)$ and $\mathcal{C}_{K,\lambda}(f) = \sup_{g\in F} T_{K,\lambda}(g,f)$. These cost functions play a central role in the theoretical analysis of $\hat{f}_K$ and $\hat{f}_{K,\lambda}$ (cf. Section 4).

Compared to the aggregation of tests by “tournaments” [60, 59], or “Le Cam type” aggregation of tests [50], minimaximization estimators have several avantages. First they are very natural from Eq (6) and they do not require a proxy for the excess risk or for the $L^2(P)$-norm, which simplifies the presentation of this estimator. Finally, while building the estimators of [60, 59] or [50] is totally impossible, these new estimators are minimaximization procedures of locally convex-concave functions that could be approximated by a numerical scheme. It is actually easy to “convert” most of the classical algorithms used for “cost minimization” into an algorithm for minimaximization solving (7) (see Section A below).

Remark 1 ($K = 1$ and ERM)**.** *If one chooses $K = 1$ then for all $f, g \in F$, $T_K(g,f) = P_N(\ell_f - \ell_g)$ and it is straightforward to check that $\hat{f}_K$ and $\hat{f}_{K,\lambda}$ are respectively the Empirical risk Minimization (ERM) and its regularized version (RERM). It turns out that, when we choose K on a data-driven basis as we do in Section A.7, when the proportion of outliers is small and good data have light tails, then the selected number of blocks $\hat{K}$ is equal or close to 1 (cf. Figure 3 below). Therefore, for clean datasets where ERM performs optimally, our procedure matches this optimal estimator.*

3 Assumptions and main results

Denote by $\mathcal{O} \cup \mathcal{I}$ a partition of $[N]$, where $\mathcal{O}$ has cardinality $|\mathcal{O}|$. Data $(X_i, Y_i)_{i\in\mathcal{O}}$ are considered as *outliers*, no assumptions on the joint distribution of these data or on their distribution conditionally on data $(X_i, Y_i)_{i\in\mathcal{I}}$ is made. These data may not be independent, nor independent from the remaining data. The remaining set $(X_i, Y_i)_{i\in\mathcal{I}}$ is the set of *informative* data, that is the ones one can use for estimation. These are hereafter assumed *independent*. Given the data $(X_i, Y_i)_{i\in[N]}$ no one knows in advance which data is informative or not.

3.1 Assumptions

All the assumptions we need to get the results involve only first and second moment of the distributions $P, (P_i)_{i\in\mathcal{I}}$ on functions in F and Y. In particular, this setting strongly relaxes usual strong concentration assumptions made on the informative data to study ERM estimators.

Assumption 1. *There exists $\theta_{r0} > 0$ such that for all $f \in F$ and all $i \in \mathcal{I}$,*

$$\sqrt{P_i(f - f^*)^2} \leqslant \theta_{r0}\sqrt{P(f - f^*)^2}.$$

Of course, Assumption 1 holds in the i.i.d. framework, with $\theta_{r0} = 1$ and $\mathcal{I} = [N]$. The second assumption bounds the correlation between the “noise” $\zeta_i = Y_i - f^*(X_i)$ and the shifted class $F - f^*$.

Assumption 2. *There exists $\theta_m > 0$ such that for all $i \in \mathcal{I}$ and all $f \in F$,*

$$var(\zeta_i(f - f^*)(X_i)) \leqslant \theta_m^2 \|f - f^*\|_{L_P^2}^2 \quad .$$

Assumption 2 typically holds in the i.i.d. setup when the noise $\zeta = Y - f^*(X)$ has uniformly bounded L^2-moments conditionally to X, which holds in the classical framework when ζ is independent of X and ζ has a finite L^2-moment bounded by θ_m. In non-i.i.d. setups, assumption 2 also holds if for all $i \in \mathcal{I}$, $\|\zeta\|_{L_{P_i}^4} \leqslant \theta_2 < \infty$ – where $\zeta(x, y) = y - f^*(x)$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}$ – and, for every $f \in F$, $\|f - f^*\|_{L_{P_i}^4} \leqslant \theta_1 \|f - f^*\|_{L_P^2}$, because, in that case,

$$\sqrt{\text{var}_{P_i}(\zeta(f - f^*))} \leqslant \|\zeta(f - f^*)\|_{L_{P_i}^2} \leqslant \|\zeta\|_{L_{P_i}^4} \|f - f^*\|_{L_{P_i}^4} \leqslant \theta_1\theta_2 \|f - f^*\|_{L_P^2},$$

and so Assumption 2 holds for $\theta_m = \theta_1\theta_2$.

Now, let us introduce a norm equivalence assumption over $F - f^*$: we call it a L^2/L^1 assumption.

Assumption 3. *There exists $\theta_0 \geqslant 1$ such that for all $f \in F$ and all $i \in \mathcal{I}$*

$$\|f - f^*\|_{L_P^2} \leqslant \theta_0 \|f - f^*\|_{L_{P_i}^1} \quad .$$

Note that $\|f - f^*\|_{L_{P_i}^1} \leqslant \|f - f^*\|_{L_{P_i}^2}$ for all $f \in F$ and $i \in \mathcal{I}$. Therefore, Assumption 1 and Assumption 3 are together equivalent to assume that all the norms $L_P^2, L_{P_i}^2, L_{P_i}^1, i \in \mathcal{I}$ are equivalent over $F - f^*$. Note also that Assumption 3 is equivalent to the small ball property (cf. [46, 67]) which has been recently used in Learning theory and signal processing. We refer to [47, 55, 66, 68, 69, 79] for examples of distributions satisfying the small ball assumption.

Proposition 1. *Let Z be a real-valued random variable. The following holds:*

1. *If there exists κ_0 and u_0 such that $\mathbb{P}(|Z| > \kappa_0 \|Z\|_2) \geqslant u_0$ then $\|Z\|_2 \leqslant (u_0\kappa_0)^{-1} \|Z\|_1$;*
2. *if there exists θ_0 such that $\|Z\|_2 \leqslant \theta_0 \|Z\|_1$, then for any $\kappa_0 < \theta_0^{-1}$, $\mathbb{P}(|Z| > \kappa_0 \|Z\|_2) \geqslant u_0$ where $u_0 = (\theta_0^{-1} - \kappa_0)^2$.*

Proof. Suppose that there exists κ_0 and u_0 such that $\mathbb{P}(|Z| > \kappa_0 \|Z\|_2) \geqslant u_0$ then

$$\|Z\|_1 \geqslant \int_{|z| \geqslant \kappa_0 \|Z\|_2} |z| dP_Z(z) \geqslant u_0 \kappa_0 \|Z\|_2 \quad ,$$

where P_Z denotes the probability distribution associated with Z. Conversely, assume that $\|Z\|_2 \leqslant \theta_0 \|Z\|_2$. Let $p = \mathbb{P}(|Z| \geqslant \kappa_0 \|Z\|_2)$. It follows from Paley-Zigmund's argument [23, Proposition 3.3.1] that

$$\begin{aligned}\|Z\|_2 &\leqslant \theta_0 \|Z\|_1 \leqslant \theta_0 \left(\mathbb{E}[|Z| I(|Z| \leqslant \kappa_0 \|Z\|_2)] + \mathbb{E}[|Z| I(|Z| \geqslant \kappa_0 \|Z\|_2)\right) \\ &\leqslant \theta_0 \|Z\|_2 (\kappa_0 + \sqrt{p}) \quad ,\end{aligned}$$

therefore, $p \geqslant (\theta_0^{-1} - \kappa_0)^2$. ■

3.2 Complexity measures and minimax rates of convergence

Balls associated with the regularization norm $\|\cdot\|$ and the L^2_P norm play a prominent role in learning theory [53, 51]. In particular, for all $\rho \geqslant 0$, the "sub-models"

$$B(f^*, \rho) = \{f \in F : \|f - f^*\| \leqslant \rho\} = f^* + \rho B$$

where $B = \{f \in \text{span}(F), \|f\| \leqslant \rho\}$ and their "localizations" at various level $r \geqslant 0$, i.e. intersection of $B(f^*, \rho)$ with L^2_P-balls

$$B_2(f^*, r) = \{f \in F : \|f - f^*\|_{L^2_P} \leqslant r\}$$

are key sets because their Rademacher complexities, drives the minimax rates of convergence. Let us introduce these complexity measures.

Definition 5. *Let $(\epsilon_i)_{i \in [N]}$ be independent Rademacher random variables (i.e. uniformly distributed in $\{-1, 1\}$), independent from $(X_i, Y_i)_{i=1}^N$. For all $f \in F$, $r > 0$ and $\rho \in (0, +\infty]$, we denote the intersection of the $\|\cdot\|$-ball of radius r and the L^2_P-norm of radius ρ centered at f by*

$$B_{reg}(f, \rho, r) = B(f, \rho) \cap B_2(f, r) = \left\{g \in F : \|g - f\|_{L^2_P} \leqslant r, \ \|g - f\| \leqslant \rho\right\} \quad .$$

Let $\zeta_i = Y_i - f^*(X_i)$ for all $i \in \mathcal{I}$ and for $\gamma_Q, \gamma_M > 0$ define

$$r_Q(\rho, \gamma_Q) = \inf\left\{r > 0 : \forall J \subset \mathcal{I}, |J| \geqslant \frac{N}{2}, \ \mathbb{E} \sup_{f \in B_{reg}(f^*, \rho, r)} \left|\sum_{i \in J} \epsilon_i (f - f^*)(X_i)\right| \leqslant \gamma_Q |J| r\right\} \quad ,$$

$$r_M(\rho, \gamma_M) = \inf\left\{r > 0 : \forall J \subset \mathcal{I}, |J| \geqslant \frac{N}{2}, \ \mathbb{E} \sup_{f \in B_{reg}(f^*, \rho, r)} \left|\sum_{i \in J} \epsilon_i \zeta_i (f - f^*)(X_i)\right| \leqslant \gamma_M |J| r^2\right\} \quad ,$$

and let $\rho \to r(\rho, \gamma_Q, \gamma_M)$ be a continuous and non decreasing function such that for every $\rho > 0$,

$$r(\rho) = r(\rho, \gamma_Q, \gamma_M) \geqslant \max\{r_Q(\rho, \gamma_Q), r_M(\rho, \gamma_M)\}.$$

It follows from Lemma 2.3 in [53] that r_M and r_Q are continuous and non decreasing functions. Note that $r_M(\cdot), r_Q(\cdot)$ depend on f^*. According to [53], if one can choose $r(\rho)$ equal to the maximum of $r_M(\rho)$ and $r_Q(\rho)$ then $r(\rho)$ is the minimax rate of convergence over $B(f^*, \rho)$. Note also that r_Q and r_M are well defined when $\mathcal{I} \geqslant N/2$, which implies that at least half data are informative.

3.3 The sparsity equation

To control the risk of our estimator, we bound from above $T_{K,\lambda}(f, f^*)$ for all functions f far from f^* either in L^2_P-norm or for the regularization norm $\|\cdot\|$. Recall that

$$T_{K,\lambda}(f, f^*) = \mathrm{MOM}_K[2\zeta(f - f^*) - (f - f^*)^2] + \lambda(\|f^*\| - \|f\|) \ .$$

The multiplier term "$2\zeta(f - f^*)$" is the one containing the noise and is therefore the term we will try to control from above using either the quadratic process "$(f-f^*)^2$" or the regularization term "$\lambda(\|f^*\|-\|f\|)$". To that end we will need both to control from below the quadratic process and the regularization term.

Let $f \in F$ and denote $\rho = \|f - f^*\|$. When $\|f - f^*\|_{L^2_P}$ is small, the quadratic term $(f - f^*)^2$ will not help to bound from above $T_{K,\lambda}(f, f^*)$, one shall only rely on the penalization term $\lambda(\|f^*\| - \|f\|)$. One can bound from below $\|f^*\| - \|f\| \gtrsim \rho$ for all f close to f^* in L^2_P using the *sparsity equation* of [56]. First, introduce the subdifferentials of $\|\cdot\|$: for all $f \in F$,

$$(\partial \|\cdot\|)_f = \{z^* \in E^* : \|f + h\| \geqslant \|f\| + z^*(h) \text{ for every } h \in E\}$$

where $(E^*, \|\cdot\|^*)$ is the dual normed space of $(E, \|\cdot\|)$.

For any $\rho > 0$, let H_ρ denote the set of functions "close" to f^* in L^2_P and at distance ρ from f^* in regularization norm and let $\Gamma_{f^*}(\rho)$ denote the set of subdifferentials of all vectors close to f^*:

$$H_\rho = \{f \in F : \|f - f^*\| = \rho \text{ and } \|f - f^*\|_{L^2_P} \leqslant r(\rho)\} \text{ and } \Gamma_{f^*}(\rho) = \bigcup_{f\in F:\|f-f^*\|\leqslant \rho/20} (\partial \|\cdot\|)_f \ .$$

If there exists a "sparse" f^{**} in $\{f \in F : \|f^* - f\| \leqslant \rho/20\}$, that is $(\partial \|\cdot\|)_{f^{**}}$ is almost all the unit dual sphere, then $\|f\| - \|f^{**}\|$ is large for any $f \in H_\rho$ so $\|f\| - \|f^*\| \geqslant \|f\| - \|f^{**}\| - \|f^* - f^{**}\|$ is large as well. More precisely, let us introduce, for all $\rho > 0$,

$$\Delta(\rho) = \inf_{f\in H_\rho} \sup_{z^*\in\Gamma_{f^*}(\rho)} z^*(f - f^*) \ .$$

The sparsity equation, introduced in [56], quantifies these notions of "large".

Definition 6. *A radius $\rho > 0$ is said to satisfy the* ***sparsity equation*** *when $\Delta(\rho) \geqslant 4\rho/5$.*

One can check that, if ρ^* satisfies the sparsity equation, so do all $\rho \geqslant \rho^*$. Therefore, one can define

$$\rho^* = \inf\left(\rho > 0 : \Delta(\rho) \geqslant \frac{4\rho}{5}\right) .$$

Note that if $\rho \geqslant 20\,\|f^*\|$ then $0 \in \Gamma_{f^*}(\rho)$. Moreover, $(\partial \|\cdot\|)_0$ equals to the dual ball (i.e. the unit ball of $(E^*, \|\cdot\|^*)$) and so $\Delta(\rho) = \rho$. This implies that any $\rho \geqslant 20\,\|f^*\|$ satisfies the sparsity equation. This simple observation will be used to get "complexity-dependent rates of convergence" as in [52].

3.4 Main results

Our first results study the performance of the estimators $\widehat{f}_{K,\lambda}$ for a fixed value of K. The other ones will provide an adaptive way to select K.

Theorem 1. *Grant Assumptions 1, 2 and 3 and let r_Q, r_M denote the functions introduced in Definition 5. Assume that $N \geqslant 384(\theta_0\theta_{r0})^2$ and $|\mathcal{O}| \leqslant N/(768\theta_0^2)$. Let ρ^* be solution to the sparsity equation from Definition 6. Let K^* denote the smallest integer such that*

$$K^* \geqslant \frac{N\epsilon^2}{384\theta_m^2} r^2(\rho^*) \ ,$$

where $\epsilon = 1/(833\theta_0^2)$ and $r^2(\cdot)$ is defined in Definition 5 for $\gamma_Q = (384\theta_0)^{-1}$ and $\gamma_M = \epsilon/192$. For any $K \geqslant K^*$, define the radius ρ_K and the regularization parameter as

$$r^2(\rho_K) = \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N} \text{ and } \lambda = \frac{16\epsilon r^2(\rho_K)}{\rho_K}.$$

Assume that for every $i \in \mathcal{I}$, $K \in [\max(K^*, |\mathcal{O}|), N]$ and $f \in F$ such that $\|f - f^*\| \leqslant \rho$ for $\rho \in [\rho_K, 2\rho_K]$, one has

$$|P_i\zeta(f - f^*) - P\zeta(f - f^*)| \leqslant \epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right) \quad . \tag{8}$$

Then, for all $K \in [\max(K^*, 8|\mathcal{O}|), N/(96(\theta_0\theta_{r0})^2)]$, with probability larger than $1 - 4\exp(-7K/9216)$, the estimator $\hat{f}_{K,\lambda}$ defined in Section 2.3 satisfies

$$\left\|\widehat{f}_{K,\lambda} - f^*\right\| \leqslant 2\rho_K, \quad \left\|\widehat{f}_{K,\lambda} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K)$$
$$R(\widehat{f}_{K,\lambda}) \leqslant R(f^*) + (1 + 52\epsilon)r^2(2\rho_K) \ .$$

Remark 2 (connexion between the $P_i, i \in \mathcal{I}$ and P)**.** *Assumption (8) holds for instance when for every $i \in \mathcal{I}$ and $f \in F$ one has*

$$\|Y_i - f^*(X_i)\|_{L^2} \leqslant \|Y - f^*(X)\|_{L^2}, \|Y_i - f(X_i)\|_{L^2} \geqslant \|Y - f(X)\|_{L^2}$$

and

$$\|f^*(X_i) - f(X_i)\|_{L^2} \leqslant \|f^*(X) - f(X)\|_{L^2}.$$

These assumptions involve second moments associated with P and $(P_i)_{i\in\mathcal{I}}$. As a consequence, if the metrics $L_{P_i}^2$ for $i \in \mathcal{I}$ and L_P^2 coincide on the functions $x \to (f - f^*)(x)$ and $(x, y) \to y - f(x)$ and the P_i satisfies Assumption 2 and 3 then Theorem 1 holds. This means that we only need informative data to induce the same L^2 metric as P to estimate the oracle f^* even if we do not have any observation coming from P itself. This setup relaxes the classical i.i.d. where all data are generated from P. In this setting, our estimators achieve the same results as the ERM would if data were all *i.i.d. with a noise ζ independent of X and both X and ζ had a Gaussian distribution (see Section D).*

The function $r(\cdot)$ is used to define the regularization parameter, so it cannot depend on f^*. When $r_M(\cdot), r_Q(\cdot)$ depend on f^*, r should be a computable upper bound independent from f^*.

3.4.1 Adaptive choice of K

In Theorem 1, all rates depend on the choice of the tuning parameter K. The following construction inspired from Lepski's method provides an adaptive choice of this parameter. Let us first recall the definition of empirical criterion introduced in Section 2.3 and the associated confidence regions: for all $J \in [K]$, $\lambda > 0$, $f \in F$ and absolute constant $c_{ad} > 0$,

$$\mathcal{C}_{J,\lambda}(f) = \sup_{g\in F} T_{J,\lambda}(g, f) \text{ and } \hat{R}_{J,c_{ad}} = \left\{f \in F : \mathcal{C}_{J,\lambda}(f) \leqslant \frac{c_{ad}}{\theta_0^2} r^2(\rho_J)\right\} \ ,$$

where $T_{J,\lambda}(g, f) = \text{MOM}_J\big(\ell_f - \ell_g\big) + \lambda\left(\|f\| - \|g\|\right)$. For all $J \in [\max(K^*, 8|\mathcal{O}|), N/(96(\theta_0\theta_{r0})^2)]$, let

$$\hat{K}_{c_{ad}} = \inf\left\{K \in [\max(K^*, 8|\mathcal{O}|), N/(96(\theta_0\theta_{r0})^2)] : \cap_{J=K}^{N/(96(\theta_0\theta_{r0})^2)} \hat{R}_{J,c_{ad}} \neq \emptyset\right\}$$
$$\text{and choose } \widehat{f}_{c_{ad}} \in \cap_{J=\hat{K}_{c_{ad}}}^{N/(96(\theta_0\theta_{r0})^2)} \hat{R}_{J,c_{ad}} \ .$$

The following theorem shows the performance of these estimators.

Theorem 2. *Grant the assumptions of Theorem 1 and assume moreover that and $|\mathcal{O}| \leqslant N/(768\theta_0^2\theta_{r0}^2)$. For any $K \in [\max(K^*, 8|\mathcal{O}|), N/(96(\theta_0\theta_{r0})^2)]$, with probability larger than*

$$1 - 4\exp(-K/2304) = 1 - 4\exp\left(-\epsilon^2 N r^2(\rho_K)/884736\right) \ ,$$

one has

$$\left\|\widehat{f}_{c_{ad}} - f^*\right\| \leqslant 2\rho_K, \qquad \left\|\widehat{f}_{c_{ad}} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K) \ ,$$
$$R(\widehat{f}_{c_{ad}}) \leqslant R(f^*) + (1+52\epsilon) r^2(2\rho_K) \ ,$$

where $c_{ad} = 18/833$ and $\epsilon = (833\theta_0^2)^{-1}$. In particular, for $K = K^*$, we have $r(2\rho_{K^*}) = \max\left(r(2\rho^*), \sqrt{|\mathcal{O}|/N}\right)$. Therefore, if $r(2\rho^*) \leqslant c_1 r(\rho^*)$ holds for some absolute constant c_1, then the breakdown number of $\widehat{f}_{c_{ad}}$ is larger than $Nr(\rho^*)^2$.

Remark 3 (deviation parameter)**.** *Note that $r(\cdot)$ is any continuous, non decreasing function such that for all $\rho > 0$, $r(\rho) \geqslant \max\left(r_Q(\rho,\gamma_Q), r_M(\rho,\gamma_M)\right)$. In particular, if $r_* : \rho \to \max\left(r_Q(\rho,\gamma_Q), r_M(\rho,\gamma_M)\right)$ is itself a continuous function (it is clearly non decreasing) then for every $x > 0$, $r(\rho) = \max\left(r_Q(\rho,\gamma_Q), r_M(\rho,\gamma_M)\right) + x/N$ is another non decreasing upper bound. Therefore, one can derive results similar to those in Theorem 2 but with an extra confidence parameter : for all $x > 0$, with probability at least $1 - 4\exp(-c_0 N r_*^2(\rho_{K^*}) + c_0 x)$,*

$$\left\|\widehat{f}_{c_{ad}} - f^*\right\| \leqslant 2\rho_K, \qquad \left\|\widehat{f}_{c_{ad}} - f^*\right\|_{L_P^2} \leqslant r_*(2\rho_K) + \frac{x}{N} \ ,$$
$$R(\widehat{f}_{c_{ad}}) \leqslant R(f^*) + (1+52\epsilon)\left(r^2(2\rho_K) + \frac{x}{N}\right) .$$

Note however that $\widehat{f}_{c_{ad}}$ depends on x through the regularization parameter $\lambda = 16\epsilon(r_*(\rho_K) + x/N)/\rho_K$.

4 Proofs

Recall the quadratic / multiplier decomposition of the difference of losses: for all $f, g \in F$, $x \in \mathcal{X}$ and $y \in \mathbb{R}$,

$$\begin{aligned}\ell_f(x,y) - \ell_g(x,y) &= (y - f(x))^2 - (y - g(x))^2 \\ &= (f(x) - g(x))^2 + 2(y - g(x))(g(x) - f(x)). \end{aligned} \tag{9}$$

Upper and lower bounds on $T_K(\cdot,\cdot)$ follow from a study of "quadratic" and "multiplier" quantiles of means processes. As no assumption is granted on the outliers, any block of data containing one or more of these outliers is "lost" from our perspective meaning that empirical means over these blocks cannot be controlled. Let $\mathcal{K}$ denote the set of blocks which have not been corrupted by outliers:

$$\mathcal{K} = \{k \in [K] : B_k \subset \mathcal{I}\} . \tag{10}$$

If $k \in \mathcal{K}$, all data indexed by B_k are informative. We will show that controls on the blocks indexed by $\mathcal{K}$ are sufficient to insure statistical performance of MOM estimators.

4.1 Bounding quadratic and multiplier processes

The first result is a lower bound on the quantiles of means quadratic processes.

Lemma 1. *Grant Assumptions 1 and 3. Fix $\eta \in (0,1)$, $\rho \in (0,+\infty]$ and let $\alpha, \gamma, \gamma_Q, x$ be positive numbers such that $\gamma\left(1-\alpha-x-16\gamma_Q\theta_0\right) \geqslant 1-\eta$. Assume that $K \in [|\mathcal{O}|/(1-\gamma), N\alpha/(2\theta_0\theta_{r0})^2]$. Then there exists an event $\Omega_Q(K)$ such that $\mathbb{P}\left(\Omega_Q(K)\right) \geqslant 1-\exp\left(-K\gamma x^2/2\right)$ and, on $\Omega_Q(K)$: for all $f \in F$ such that $\|f-f^*\| \leqslant \rho$, if $\|f-f^*\|_{L_P^2} \geqslant r_Q(\rho,\gamma_Q)$ then*

$$\left|\left\{k \in [K] : P_{B_k}(f-f^*)^2 \geqslant (4\theta_0)^{-2}\left\|f-f^*\right\|_{L_P^2}^2\right\}\right| \geqslant (1-\eta)K \ .$$

In particular, $Q_{\eta,K}((f-f^*)^2) \geqslant (4\theta_0)^{-2}\left\|f-f^*\right\|_{L_P^2}^2$.

Proof. Define $F_\rho^* = B(f^*,\rho) = \{f \in F : \|f-f^*\| \leqslant \rho\}$. For all $f \in F_\rho^*$, let $n_f = (f-f^*)/\left\|f-f^*\right\|_{L_P^2}$ and note that for all $i \in \mathcal{I}$, $P_i|n_f| \geqslant \theta_0^{-1}$ by Assumption 3 and $P_i n_f^2 \leqslant \theta_{r0}^2$ by Assumption 1. It follows from Markov's inequality that, for all $k \in \mathcal{K}$ ($\mathcal{K}$ is defined in (10)),

$$\mathbb{P}\left(\left|P_{B_k}|n_f| - \overline{P}_{B_k}|n_f|\right| > \frac{\theta_{r0}}{\sqrt{\alpha|B_k|}}\right) \leqslant \alpha \ ,$$

where $\overline{P}_{B_k}|n_f| = |B_k|^{-1}\sum_{i\in B_k}\mathbb{E}|n_f(X_i)| \geqslant \theta_0^{-1}$ and therefore,

$$\mathbb{P}\left(P_{B_k}|n_f| \geqslant \frac{1}{\theta_0} - \frac{\theta_{r0}}{\sqrt{\alpha|B_k|}}\right) \geqslant 1-\alpha \ .$$

Since $K \leqslant N\alpha/(2\theta_0\theta_{r0})^2$, $|B_k| = N/K \geqslant (2\theta_0\theta_{r0})^2/\alpha$ and so

$$\mathbb{P}\left(2\theta_0 P_{B_k}|n_f| \geqslant 1\right) \geqslant 1-\alpha \ . \tag{11}$$

Let ϕ be the function defined on $\mathbb{R}_+$ by $\phi(t) = (t-1)I(1 \leqslant t \leqslant 2) + I(t \geqslant 2)$, and, for all $f \in F_\rho^*$ let $Z(f) = \sum_{k\in[K]} I(4\theta_0 P_{B_k}|n_f| \geqslant 1)$. Since for all $x \in \mathbb{R}$, $I(x \geqslant 1) \geqslant \phi(x)$,

$$Z(f) \geqslant \sum_{k\in\mathcal{K}} \phi\left(4\theta_0 P_{B_k}|n_f|\right) \ .$$

Now, for any $x \in \mathbb{R}_+$, $\phi(x) \geqslant I(x \geqslant 2)$, thus, according to (11),

$$\mathbb{E}\left[\sum_{k\in\mathcal{K}} \phi\left(4\theta_0 P_{B_k}|n_f|\right)\right] \geqslant \sum_{k\in\mathcal{K}} \mathbb{P}\left(4\theta_0 P_{B_k}|n_f| \geqslant 2\right) \geqslant |\mathcal{K}|(1-\alpha) \ .$$

Therefore,

$$Z(f) \geqslant |\mathcal{K}|(1-\alpha) + \sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right) \ .$$

Denote $\mathcal{F} = \{f \in F : \|f-f^*\| \leqslant \rho, \ \|f-f^*\|_{L_P^2} \geqslant r_Q(\rho,\gamma_Q)\}$. By the bounded difference inequality (see, for instance [17, Theorem 6.2]), there exists an event $\Omega_Q(K)$ with probability larger than $1-\exp(-x^2|\mathcal{K}|/2)$, on which, for all $f \in \mathcal{F}$,

$$\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right)\right| \leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right)\right| + |\mathcal{K}|x \ .$$

By the symmetrization argument,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \left(\phi\left(4\theta_0 P_{B_k}|n_f|\right) - \mathbb{E}\left[\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right]\right)\right| \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}} \epsilon_k\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right| \ .$$

Since the function ϕ is 1-Lipschitz and $\phi(0) = 0$, by the contraction principle (see, for example [57, Chapter 4] or [17, Theorem 11.6]), we have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\epsilon_k\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right| \leqslant 4\theta_0\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\epsilon_k P_{B_k}|n_f|\right| .$$

The family $(\epsilon_{[i]}|n_f(X_i)| : i\in\cup_{k\in\mathcal{K}}B_k)$, where $[i] = \lceil i/K\rceil$ for all $i\in\mathcal{I}$, is a collection of centered random variables. Therefore, if $(\epsilon'_k)_{k\in\mathcal{K}}$ and $(X'_i)_{i\in\mathcal{I}}$ denote independent copies of $(\epsilon_k)_{k\in\mathcal{K}}$ and $(X_i)_{i\in\mathcal{I}}$ then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\epsilon_k P_{B_k}|n_f|\right| \leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right| .$$

Then, as $(X_i)_{i\in\mathcal{I}}$ and $(X'_i)_{i\in\mathcal{I}}$ are two independent families of independent variables therefore, if $(\epsilon''_i)_{i\in\mathcal{I}}$ denote a family of i.i.d. Rademacher variables independent of $(\epsilon_i), (\epsilon'_i), (X_i)_{i\in\mathcal{I}}, (X'_i)_{i\in\mathcal{I}}$ then $(\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|)$ and $(\epsilon''_i\left(\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right))$ have the same distribution. Therefore,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right| \leqslant \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon''_i\left(\epsilon_k|n_f(X_i)| - \epsilon'_k|n_f(X'_i)|\right)\right|$$

$$= \mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\epsilon''_i\left(|n_f(X_i)| - |n_f(X'_i)|\right)\right| \leqslant \frac{2K}{N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right| .$$

By the contraction principle, on $\Omega_Q(K)$,

$$Z(f) \geqslant |\mathcal{K}|\left(1-\alpha-x-\frac{16\theta_0 K}{|\mathcal{K}|N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right|\right) . \tag{12}$$

For any $f\in\mathcal{F}$, $r_Q(\rho,\gamma_Q)n_f + f^*\in F$ because F is convex. Moreover, $\|r_Q(\rho,\gamma_Q)n_f\|_{L_P^2} = r_Q(\rho,\gamma_Q)$ and $\|r_Q(\rho,\gamma_Q)n_f\| = [r_Q(\rho,\gamma_Q)/\|f-f^*\|_{L_P^2}]\,\|f-f^*\| \leqslant \rho$. Therefore, $r_Q(\rho,\gamma_Q)n_f + f^*\in\mathcal{F}$. Therefore, by definition of $r_Q(\rho,\gamma_Q)$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right| = \frac{1}{r_Q(\rho,\gamma_Q)}\mathbb{E}\sup_{f\in F:\|f-f^*\|\leqslant\rho,\ \|f-f^*\|_{L_P^2}=r_Q(\rho,\gamma_Q)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i(f-f^*)(X_i)\right| \leqslant \gamma_Q\frac{|\mathcal{K}|N}{K} .$$

Using the last inequality together with (12) and the assumption $K\geqslant|\mathcal{O}|/(1-\gamma)$ (so that $|\mathcal{K}|\geqslant K-|\mathcal{O}|\geqslant\gamma K$), we get that, on the event $\Omega_Q(K)$, for any $f\in\mathcal{F}$,

$$Z(f)\geqslant|\mathcal{K}|\left(1-\alpha-x-16\theta_0\gamma_Q\right)\geqslant(1-\eta)K .$$

Hence, on $\Omega_Q(K)$, for any $f\in\mathcal{F}$, there exists at least $(1-\eta)K$ blocks B_k for which $P_{B_k}|n_f|\geqslant(4\theta_0)^{-1}$. On these blocks, $P_{B_k}n_f^2\geqslant(P_{B_k}|n_f|)^2\geqslant(4\theta_0)^{-2}$, therefore, on $\Omega_Q(K)$, $Q_{\eta,K}[n_f^2]\geqslant(4\theta_0)^{-2}$. ∎

Now, let us turn to a control of the multiplier process.

Lemma 2. *Grant Assumption 2. Fix $\eta\in(0,1)$, $\rho\in(0,+\infty]$, and let α,γ_M,γ,x and ϵ be positive absolute constants such that $\gamma\left(1-\alpha-x-8\gamma_M/\epsilon\right)\geqslant 1-\eta$. Let $K\in[|\mathcal{O}|/(1-\gamma),N]$. There exists an event $\Omega_M(K)$ such that $\mathbb{P}(\Omega_M(K))\geqslant 1-\exp(-\gamma Kx^2/2)$ and on the event $\Omega_M(K)$: if $f\in F$ is such that $\|f-f^*\|\leqslant\rho$ then*

$$\left|\left\{k\in\mathcal{K} : \left|2(P_{B_k}-\overline{P}_{B_k})(\zeta(f-f^*))\right|\leqslant\epsilon\max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho,\gamma_M), \|f-f^*\|_{L_P^2}^2\right)\right\}\right|\geqslant(1-\eta)K .$$

Proof. For all $k \in [K]$ and $f \in F$, set $W_k = ((X_i, Y_i))_{i \in B_k}$ and define

$$g_f(W_k) = 2(P_{B_k} - \overline{P}_{B_k})\left(\zeta(f - f^*)\right) \text{ and } \gamma_k(f) = \epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2 \alpha}\frac{K}{N}, r_M^2(\rho, \gamma_M), \|f - f^*\|_{L_P^2}^2\right) \ .$$

Let $f \in F$ and $k \in \mathcal{K}$. It follows from Markov's inequality that

$$\begin{aligned}\mathbb{P}\left[2\left|g_f(W_k)\right| \geqslant \gamma_k(f)\right] &\leqslant \frac{4\mathbb{E}\left[\left(2(P_{B_k} - \overline{P}_{B_k})(\zeta(f - f^*))\right)^2\right]}{\frac{16\theta_m^2}{\alpha}\|f - f^*\|_{L_P^2}^2\frac{K}{N}} \\ &\leqslant \frac{\alpha\sum_{i \in B_k}\mathrm{var}_{P_i}(\zeta(f - f^*))}{|B_k|^2\theta_m^2\|f - f^*\|_{L_P^2}^2\frac{K}{N}} \leqslant \frac{\alpha\theta_m^2\|f - f^*\|_{L_P^2}^2}{|B_k|\theta_m^2\|f - f^*\|_{L_P^2}^2\frac{K}{N}} = \alpha \ . \end{aligned} \tag{13}$$

Let $J = \cup_{k \in \mathcal{K}} B_k$ and let $r_M(\rho) = r_M(\rho, \gamma_M)$. We have

$$\begin{aligned}\mathbb{E}\sup_{f \in B(f^*,\rho)}\sum_{k \in \mathcal{K}}\epsilon_k\frac{g_f(W_k)}{\gamma_k(f)} &\leqslant 2\mathbb{E}\sup_{f \in B(f^*,\rho)}\left|\sum_{k \in \mathcal{K}}\frac{\epsilon_k(P_{B_k} - \overline{P}_{B_k})(\zeta(f - f^*))}{\epsilon\max(r_M^2(\rho), \|f - f^*\|_{L_P^2}^2)}\right| \\ &\leqslant \frac{2}{\epsilon r_M^2(\rho)}\mathbb{E}\left[\sup_{f \in B(f^*,\rho):\|f - f^*\|_{L_P^2} \geqslant r_M(\rho)}\left|\sum_{k \in \mathcal{K}}\epsilon_k(P_{B_k} - \overline{P}_{B_k})\left(\zeta r_M(\rho)\frac{f - f^*}{\|f - f^*\|_{L_P^2}}\right)\right|\right. \\ &\qquad \left.\vee \sup_{f \in B(f^*,\rho):\|f - f^*\|_{L_P^2} \leqslant r_M(\rho)}\left|\sum_{k \in \mathcal{K}}\epsilon_k(P_{B_k} - \overline{P}_{B_k})\left(\zeta(f - f^*)\right)\right|\right] \\ &\leqslant \frac{2}{\epsilon r_M^2(\rho)}\mathbb{E}\sup_{f \in B(f^*,\rho):\|f - f^*\|_{L_P^2} \leqslant r_M(\rho)}\left|\sum_{k \in \mathcal{K}}\epsilon_k(P_{B_k} - \overline{P}_{B_k})\left(\zeta(f - f^*)\right)\right| \ ,\end{aligned}$$

where in the last but one inequality, we used that the class F is convex and the same argument as in the proof of Lemma 1. Since $(\epsilon_{[i]}(\zeta_i(f - f^*)(X_i) - P_i\zeta_i(f - f^*)) : i \in \mathcal{I})$ is a family of centered random variables, one can use the symmetrization argument to get

$$\begin{aligned}\mathbb{E}\sup_{f \in B(f^*,\rho)}\sum_{k \in \mathcal{K}}\epsilon_k\frac{g_f(W_k)}{\gamma_k(f)} &\leqslant \frac{4K}{\epsilon r_M^2(\rho)N}\mathbb{E}\sup_{f \in B(f^*,\rho):\|f - f^*\|_{L_P^2} \leqslant r_M(\rho)}\left|\sum_{i \in J}\epsilon_i\zeta_i(f - f^*)(X_i)\right| \\ &\leqslant \frac{4K}{\epsilon N}\gamma_M|\mathcal{K}|\frac{N}{K} = \frac{4\gamma_M}{\epsilon}|\mathcal{K}| \ ,\end{aligned} \tag{14}$$

where the definition of $r_M(\rho)$ has been used in the last but one inequality.

Let $\psi(t) = (2t - 1)I(1/2 \leqslant t \leqslant 1) + I(t \geqslant 1)$. The function ψ is 2-Lipschitz and satisfies $I(t \geqslant 1) \leqslant \psi(t) \leqslant I(t \geqslant 1/2)$, for all $t \in \mathbb{R}$. Therefore, all $f \in B(f^*, \rho)$ satisfies

$$\begin{aligned}\sum_{k \in \mathcal{K}} I\left(|g_f(W_k)| < \gamma_k(f)\right) &= |\mathcal{K}| - \sum_{k \in \mathcal{K}} I\left(\frac{|g_f(W_k)|}{\gamma_k(f)} \geqslant 1\right) \geqslant |\mathcal{K}| - \sum_{k \in \mathcal{K}}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) \\ &= |\mathcal{K}| - \sum_{k \in \mathcal{K}}\mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \sum_{k \in \mathcal{K}}\left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right] \\ &\geqslant |\mathcal{K}| - \sum_{k \in \mathcal{K}}\mathbb{E}I\left(\frac{|g_f(W_k)|}{\gamma_k(f)} \geqslant \frac{1}{2}\right) - \sum_{k \in \mathcal{K}}\left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{P}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right] \\ &\geqslant (1 - \alpha)|\mathcal{K}| - \sup_{f \in B(f^*,\rho)}\left|\sum_{k \in \mathcal{K}}\left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right|\end{aligned}$$

where we used (13) in the last inequality.

The bounded difference inequality ensures that there exists an event $\Omega_M(K)$ satisfying $\mathbb{P}(\Omega_M(K)) \geqslant 1 - \exp(-x^2|\mathcal{K}|/2)$, where

$$\sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right| \\ \leqslant \mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right| + |\mathcal{K}|x \enspace .$$

Furthermore, it follows from by the symmetrization argument that

$$\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \left[\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right) - \mathbb{E}\psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right]\right| \leqslant 2\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \epsilon_k \psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right|$$

and, from the contraction principle and (14), that

$$\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \epsilon_k \psi\left(\frac{|g_f(W_k)|}{\gamma_k(f)}\right)\right| \leqslant 2\mathbb{E} \sup_{f\in B(f^*,\rho)} \left|\sum_{k\in\mathcal{K}} \epsilon_k \frac{|g_f(W_k)|}{\gamma_k(f)}\right| \leqslant \frac{8\gamma_M}{\epsilon}|\mathcal{K}| \enspace .$$

In conclusion, on $\Omega_M(K)$, for all $f \in B(f^*,\rho)$,

$$\sum_{k\in\mathcal{K}} I\left(|g_f(W_k)| < \gamma_k(f)\right) \geqslant \left(1-\alpha-x-8\gamma_M/\epsilon\right)|\mathcal{K}| \geqslant K\gamma\left(1-\alpha-x-8\gamma_M/\epsilon\right) \geqslant (1-\eta)K \enspace .$$

■

4.2 Bounding the empirical criterion $\mathcal{C}_{K,\lambda}(f^*)$

Let us first introduce the event on which the statement of Theorem 1 holds. Denote by $\Omega(K)$ the intersection of the events $\Omega_Q(K)$, $\Omega_M(K)$ defined respectively in Lemmas 1 and 2 for $\rho \in \{\kappa\rho_K : \kappa \in \{1,2\}\}$ and

$$\eta = \frac{1}{4}, \gamma = \frac{7}{8}, \alpha = \frac{1}{24}, x = \frac{1}{24}, \gamma_Q = \frac{1}{384\theta_0}, \epsilon = \frac{1}{c\theta_0^2} \text{ and } \gamma_M = \frac{\epsilon}{192} \tag{15}$$

for some absolute constants $c > 0$ to be specified later. For these values, conditions in both Lemmas 1 and 2 are satisfied:

$$\gamma(1-\alpha-x-16\gamma_Q\theta_0) \geqslant 1-\eta = \frac{3}{4} \text{ and } \gamma(1-\alpha-x-8\gamma_M/\epsilon) \geqslant 1-\eta = \frac{3}{4}.$$

According to Lemmas 1 and 2, the event $\Omega(K)$ satisfies $\mathbb{P}(\Omega(K)) \geqslant 1 - 4\exp\left(-7K/9216\right)$. On $\Omega(K)$, the following holds for all $\rho \in \{\kappa\rho_K : \kappa \in \{1,2\}\}$ and $f \in F$ such that $\|f - f^*\| \leqslant \rho$,

1. if $\|f-f^*\|_{L_P^2} \geqslant r_Q(\rho,\gamma_Q)$ then

$$Q_{1/4,K}((f-f^*)^2) \geqslant \frac{1}{(4\theta_0)^2}\|f-f^*\|_{L_P^2}^2 \quad , \tag{16}$$

2. there exists $3K/4$ block B_k with $k \in \mathcal{K}$, for which

$$|(P_{B_k} - \overline{P}_{B_k})[2\zeta(f-f^*)]| \leqslant \epsilon \max\left(r_M^2(\rho,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_P^2}^2\right) \quad . \tag{17}$$

Moreover, on the blocks B_k where (17) holds, it follows from assumption in (8) that all $f \in F$ such that $\|f - f^*\| \leqslant \rho$ satisfies

$$P_{B_k}[2\zeta(f - f^*)]| \leqslant P[2\zeta(f - f^*)] + 2\epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right) \enspace .$$

It follows from the convexity of F and the nearest point theorem that $P[2\zeta(f - f^*)] \leqslant 0$ for all $f \in F$, therefore, for all $f \in F$ such that $\|f - f^*\| \leqslant \rho$,

$$Q_{3/4,K}(2\zeta(f - f^*)) \leqslant 2\epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right) \enspace . \tag{18}$$

Moreover, still on the blocks B_k where (17) holds, one also has, thanks to assumption (8), that for all $f \in F$ such that $\|f - f^*\| \leqslant \rho$,

$$P[-2\zeta(f - f^*)] \leqslant P_{B_k}[-2\zeta(f - f^*)] + 2\epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right) \enspace .$$

It follows that, for all $f \in F$ such that $\|f - f^*\| \leqslant \rho$,

$$\begin{aligned}
&P[-2\zeta(f - f^*)] \leqslant Q_{1/4,K}[-2\zeta(f - f^*)] + 2\epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right)\\
&\leqslant Q_{1/4,K}[(f - f^*)^2 - 2\zeta(f - f^*)] + \lambda(\|f\| - \|f^*\|) + 2\epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right) + \lambda\rho\\
&\leqslant T_{K,\lambda}(f^*, f) + 2\epsilon \max\left(r_M^2(\rho, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right) + \lambda\rho \enspace .
\end{aligned} \tag{19}$$

The main result of this section is Lemma 3. It will be used to bound from above the criterion $\mathcal{C}_{K,\lambda}(f^*) = \sup_{g \in F} T_{K,\lambda}(g, f^*)$. Recall that ρ_K and λ are defined as

$$r^2(\rho_K) = \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N} \text{ and } \lambda = \frac{c'\epsilon r^2(\rho_K)}{\rho_K} \tag{20}$$

where $\epsilon = (c\theta_0^2)^{-1}$ and $c, c' \geqslant$ are absolute constants. We also need to consider a partition of the space F according to the distance between g and f^* w.r.t. $\|\cdot\|$ and $\|\cdot\|_{L_P^2}$ as in Figure 2: define for all $\kappa \geqslant 1$,

$$\begin{aligned}
F_1^{(\kappa)} &= \left\{g \in F : \|g - f^*\| \leqslant \kappa\rho_K \text{ and } \|g - f^*\|_{L_P^2} \leqslant r(\kappa\rho_K)\right\} \enspace ,\\
F_2^{(\kappa)} &= \left\{g \in F : \|g - f^*\| \leqslant \kappa\rho_K \text{ and } \|g - f^*\|_{L_P^2} > r(\kappa\rho_K)\right\} \enspace ,\\
F_3^{(\kappa)} &= \{g \in F : \|g - f^*\| > \kappa\rho_K\} \enspace .
\end{aligned}$$

Lemma 3. *On the event $\Omega(K)$, it holds for all $\kappa \in \{1, 2\}$,*

$$\sup_{g \in F_1^{(\kappa)}} T_{K,\lambda}(g, f^*) \leqslant (2 + c'\kappa)\epsilon r^2(\kappa\rho_K), \qquad \sup_{g \in F_2^{(\kappa)}} T_{K,\lambda}(g, f^*) \leqslant \left((2 + c'\kappa)\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\kappa\rho_K)$$

and

$$\sup_{g \in F_3^{(\kappa)}} T_{K,\lambda}(g, f^*) \leqslant \kappa \max\left(2\epsilon - \frac{1}{16\theta_0^2} + \frac{11c'\epsilon}{10}, 2\epsilon - \frac{7c'\epsilon}{10}\right) r^2(\rho_K)$$

when $c \geqslant 32$ and $10\epsilon/4 \leqslant c'\epsilon \leqslant ((4\theta_0)^{-2} - 2\epsilon)$.

Proof of Lemma 3. Recall that, for all $g \in F$, $\ell_{f^*} - \ell_g = 2\zeta(g - f^*) - (g - f^*)^2$ where $\zeta(x, y) = y - f^*(x)$. Let us now place ourself on the event $\Omega(K)$ up to the end of proof.

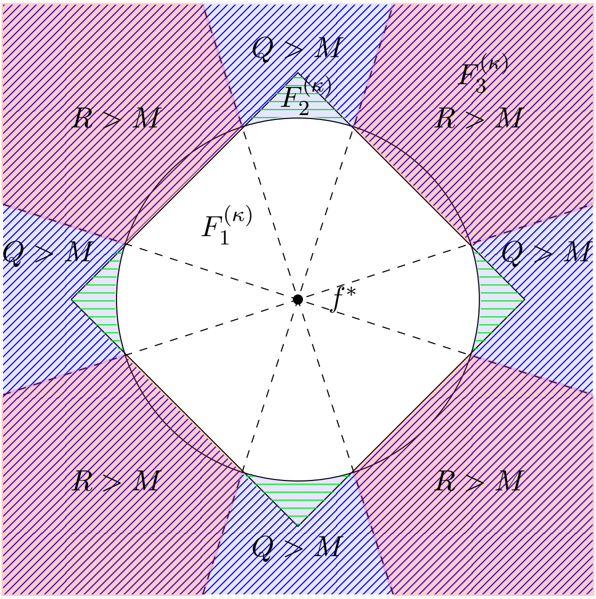

Figure 2: Partition $\{F_1^{(\kappa)}, F_2^{(\kappa)}, F_3^{(\kappa)}\}$ of F and the control of the multiplier MOM process by either the quadratic MOM process (the "$Q > M$" part) or the regularization term (the "$R > M$" part).

Bounding $\sup_{\mathbf{g}\in\mathbf{F_1^{(\kappa)}}} \mathbf{T_{K,\lambda}(g, f^*)}$. Let $g \in F_1^{(\kappa)}$. Since the quadratic process is non negative,

$$T_{K,\lambda}(g, f^*) = \mathrm{MOM}_K\big(2\zeta(g-f^*) - (g-f^*)^2\big) - \lambda\left(\|g\| - \|f^*\|\right) \leqslant Q_{3/4,K}(2\zeta(g-f^*)) + \lambda\left\|f^* - g\right\| \enspace .$$

Therefore, applying (18) for $\rho = \kappa\rho_K$ and the choice of ρ_K and λ as in (20), we get

$$\begin{aligned} T_{K,\lambda}(g, f^*) &\leq 2\epsilon\max\left(r_M^2(\kappa\rho_K, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) + \lambda\kappa\rho_K \leqslant 2\epsilon r^2(\kappa\rho_K) + c'\kappa\epsilon r^2(\rho_K) \\ &\leqslant (2 + c'\kappa)\epsilon r^2(\kappa\rho_K) \enspace . \end{aligned}$$

Bounding $\sup_{\mathbf{g}\in\mathbf{F_2^{(\kappa)}}} \mathbf{T_{K,\lambda}(g, f^*)}$. Let $g \in F_2^{(\kappa)}$. Given that $Q_{1/2}(x-y) \leqslant Q_{3/4}(x) - Q_{1/4}(y)$ for any vector x and y, we have

$$\mathrm{MOM}_K\big(2\zeta(g-f^*) - (g-f^*)^2\big) + \lambda\left(\|f^*\| - \|g\|\right) \leqslant Q_{3/4,K}(2\zeta(g-f^*)) - Q_{1/4,K}((f^*-g)^2) + \lambda\kappa\rho_K \enspace .$$

Moreover $2\epsilon \leqslant (4\theta_0)^{-2}$ when $c \geqslant 32$, so it follows from (16) and (18) for $\rho = \kappa\rho_K$ that

$$\begin{aligned} Q_{3/4,K}(2\zeta(f^*-g)) - Q_{1/4,K}((f^*-g)^2) &\leqslant 2\epsilon\max\left(r_M^2(\kappa\rho_K, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) - \frac{\|f-f^*\|_{L_P^2}^2}{(4\theta_0)^2} \\ &\leqslant \left(2\epsilon - \frac{1}{(4\theta_0)^2}\right)\|f-f^*\|_{L_P^2}^2 \leqslant \left(2\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\kappa\rho_K) \enspace . \end{aligned}$$

Putting both inequalities together and using that $\lambda\kappa\rho_K = c'\kappa\epsilon r^2(\rho_K)$, we get

$$T_{K,\lambda}(g, f^*) \leqslant \left((2 + c'\kappa)\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\kappa\rho_K) \enspace .$$

Bounding $\sup_{\mathbf{g}\in\mathbf{F_3^{(\kappa)}}} \mathbf{T_{K,\lambda}(g, f^*)}$ via an homogeneity argument. Start with two lemmas.

Lemma 4. *Let $\rho \geqslant 0$, $\Gamma_{f^*}(\rho) = \cup_{f \in f^* + (\rho/20)B}(\partial \left\|\cdot\right\|)_f$ (cf.) section 3.3). For all $g \in F$,*

$$\|g\| - \|f^*\| \geqslant \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(g - f^*) - \frac{\rho}{10} \ .$$

Proof. Let $g \in F$, $f^{**} \in f^* + (\rho/20)B$ and $z^* \in (\partial \left\|\cdot\right\|)_{f^{**}}$. We have

$$\|g\| - \|f^*\| \geqslant \|g\| - \|f^{**}\| - \|f^{**} - f^*\| \geqslant z^*(g - f^{**}) - \frac{\rho}{20} = z^*(g - f^*) - z^*(f^{**} - f^*) - \frac{\rho}{20} \geqslant z^*(g - f^*) - \frac{\rho}{10} \ ,$$

where the last inequality follows from $z^*(f^{**} - f^*) \leqslant \|f^{**} - f^*\|$. The result follows by taking supremum over $z^* \in \Gamma_{f^*}(\rho)$. ∎

Lemma 5. *Let $\rho \geqslant 0$. Let $g \in F$ be such that $\|g - f^*\| \geqslant \rho$. Define $f = f^* + \rho(g - f^*)/\left\|g - f^*\right\|$. Then $f \in F$, $\|f - f^*\| = \rho$ and,*

$$\begin{aligned} &MOM_K\big((g - f^*)^2 - 2\zeta(g - f^*)\big) + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(g - f^*) \\ &\qquad \geqslant \frac{\|g - f^*\|_{L_P^2}}{\rho} \left(MOM_K\big((f - f^*)^2 - 2\zeta(f - f^*)\big) + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*) \right) \ . \end{aligned}$$

Proof. The first conclusion holds by convexity of F, the second statement is obvious. For the last one, let $\Upsilon = \|g - f^*\|/\rho$ and note that $\Upsilon \geqslant 1$ and $g - f^* = \Upsilon(f - f^*)$, so we have

$$\begin{aligned} &\mathrm{MOM}_K\big((g - f^*)^2 - 2\zeta(g - f^*)\big) + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(g - f^*) \\ &= \mathrm{MOM}_K\big(\Upsilon^2(f - f^*)^2 - 2\Upsilon\zeta(f - f^*)\big) + \lambda\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*) \\ &\geqslant \Upsilon \left(\mathrm{MOM}_K\big((f - f^*)^2 - 2\zeta(f - f^*)\big) + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*) \right) \ . \end{aligned}$$

∎

Now, let us bound $\sup_{g \in F_3^{(\kappa)}} T_{K,\lambda}(g, f^*)$. Let $g \in F_3^{(\kappa)}$. Apply Lemma 4 and Lemma 5 to $\rho = \rho_K$: there exists $f \in F$ such that $\|f - f^*\| = \rho_K$ and

$$\begin{aligned} T_{K,\lambda}(g, f^*) &= \mathrm{MOM}_K\big(2\zeta(g - f^*) - (g - f^*)^2\big) - \lambda\left(\|g\| - \|f^*\|\right) \\ &\leqslant \mathrm{MOM}_K\big(2\zeta(g - f^*) - (g - f^*)^2\big) - \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g - f^*) + \lambda\frac{\kappa\rho_K}{10} \\ &\leqslant \frac{\|g - f^*\|}{\rho_K} \left(\mathrm{MOM}_K\big(2\zeta(f - f^*) - (f - f^*)^2\big) - \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \right) + \lambda\frac{\kappa\rho_K}{10} \ . \end{aligned} \tag{21}$$

First assume that $\|f - f^*\|_{L_P^2} \leqslant r(\rho_K)$. In that case, $\|f - f^*\| = \rho_K$ and $\|f - f^*\|_{L_P^2} \leqslant r(\rho_K)$ therefore, $f \in H_{\rho_K}$. Moreover, by definition of K^* and since $K \geqslant K^*$, we have $\rho_K \geqslant \rho^*$ which implies that ρ_K satisfies the sparsity equation from Definition 6. Therefore, $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \geqslant \Delta(\rho_K) \geqslant 4\rho_K/5$. Now, it follows from the definition of λ in (20) that

$$-\lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \leqslant -\frac{4c'\epsilon r^2(\rho_K)}{5} \ .$$

Moreover, since the quadratic process is non-negative, by (18) applied to $\rho = \rho_K$,

$$\begin{aligned} &\mathrm{MOM}_K\big(2\zeta(f - f^*) - (f - f^*)^2\big) \leqslant Q_{3/4,K}[2\zeta(f - f^*)] \\ &\leqslant 2\epsilon \max\left(r_M^2(\rho_K, \gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2 \right) \leqslant 2\epsilon r^2(\rho_K) \ . \end{aligned}$$

Finally, noting that $2\epsilon - 4c'\epsilon/5 \leqslant 0$ when $c' \geqslant 10/4$, binding all the pieces together in (21) yields

$$T_{K,\lambda}(g,f^*) \leqslant \kappa\epsilon\left(2-4c'/5\right) r^2(\rho_K) + \lambda\frac{\kappa\rho_K}{10} = \kappa\epsilon\left(2-\frac{7c'}{10}\right) r^2(\rho_K) \ .$$

Second, assume that $\|f-f^*\|_{L_P^2} \geqslant r(\rho_K)$. Since $\|f-f^*\| = \rho_K$, it follows from (16) and (17) for $\rho = \rho_K$ that

$$\begin{aligned} &\text{MOM}_K\big(2\zeta(f-f^*)-(f-f^*)^2\big) \leqslant Q_{3/4,K}(2\zeta(f-f^*)) - Q_{1/4,K}((f^*-f)^2)\\ &\leqslant 2\epsilon\max\left(r_M^2(\rho_K,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f-f^*\|_{L_p^2}^2\right) - \frac{\|f-f^*\|_{L_P^2}^2}{(4\theta_0)^2}\\ &\leqslant \left(2\epsilon - \frac{1}{16\theta_0^2}\right)\|f-f^*\|_{L_P^2}^2 \leqslant \left(2\epsilon-\frac{1}{16\theta_0^2}\right) r^2(\rho_K) \ , \end{aligned}$$

where we used that $2\epsilon \leqslant (16\theta_0)^{-2}$ when $c \geqslant 32$ in the last inequality. Plugging the last result in (21) we get

$$\begin{aligned} T_{K,\lambda}(g,f^*) &\leqslant \frac{\|g-f^*\|}{\rho_K}\left(\left(2\epsilon-\frac{1}{16\theta_0^2}\right) r^2(\rho_K) + \lambda\rho_K\right) + \lambda\frac{\kappa\rho_K}{10}\\ &\leqslant \frac{\|g-f^*\|}{\rho_K}\left((2+c')\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\rho_K) + \frac{c'\kappa\epsilon}{10} r^2(\rho_K) \leqslant \kappa\left(\left(2+\frac{11c'}{10}\right)\epsilon - \frac{1}{16\theta_0^2}\right) r^2(\rho_K) \end{aligned}$$

when $16(2+c')\epsilon \leqslant \theta_0^{-2}$.

4.3 From a control of $\mathcal{C}_{K,\lambda}(\hat{f})$ to statistical performance

The proof follows essentially the one of [56, Theorem 3.2] or [50, Lemma 2].

Lemma 6. *Let $\hat{f} \in F$ be such that, on $\Omega(K)$, $\mathcal{C}_{K,\lambda}(\hat{f}) \leqslant (2+c')\epsilon r^2(\rho_K)$. Then, on $\Omega(K)$, $\hat{f}$ satisfies*

$$\left\|\hat{f}-f^*\right\| \leqslant 2\rho_K, \quad \left\|\hat{f}-f^*\right\|_{L_P^2} \leqslant r(2\rho_K) \quad \textit{and} \quad R(\hat{f}) \leqslant R(f^*) + (1+(4+3c')\epsilon) r^2(2\rho_K) \ ,$$

when $c' = 16$ and $c > 832$.

Proof. Recall that for any $x \in \mathbb{R}^K$, $Q_{1/2}(x) \geqslant -Q_{1/2}(-x)$. Therefore,

$$\mathcal{C}_{K,\lambda}(\hat{f}) = \sup_{g\in F} T_{K,\lambda}(g,\hat{f}) \geqslant T_{K,\lambda}(f^*,\hat{f}) \geqslant -T_{K,\lambda}(\hat{f},f^*) \ .$$

Thus, on $\Omega(K)$, $\hat{f} \in \left\{g \in F : T_{K,\lambda}(g,f^*) \geqslant -(2+c')\epsilon r^2(\rho_K)\right\}$. When $c' = 16$ and $c > 832$,

$$-(2+c')\epsilon > 2(1+c')\epsilon - \frac{1}{16\theta_0^2} \text{ and } -(2+c')\epsilon > 2\max\left(2\epsilon - \frac{1}{16\theta_0^2} + \frac{11c'\epsilon}{10}, 2\epsilon - \frac{7c'\epsilon}{10}\right)$$

therefore, $\hat{f} \in F_1^{(2)}$ on $\Omega(K)$. This yields the results for both the regularization and the L_P^2-norm.

Finally, let us turn to the control on the excess risk. It follows from (19) for $\rho = \kappa\rho_K$ that

$$\begin{aligned} R(\hat{f}) - R(f^*) &= \left\|\hat{f}-f^*\right\|_{L_P^2}^2 + P[-2\zeta(\hat{f}-f^*)]\\ &\leqslant r^2(2\rho_K) + T_{K,\lambda}(f^*,\hat{f}) + 2\epsilon\max\left(r_M^2(2\rho_K,\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \left\|\hat{f}-f^*\right\|_{L_p^2}^2\right) + 2\lambda\rho_K\\ &\leqslant r^2(2\rho_K) + \mathcal{C}_{K,\lambda}(\hat{f}) + 2\epsilon r^2(2\rho_K) + 2c'\epsilon r^2(\rho_K) = (1+(4+3c')\epsilon) r^2(2\rho_K) \ . \end{aligned}$$

■

4.4 End of the proof of Theorem 1

By definition of $\widehat{f}_{K,\lambda}$,

$$\mathcal{C}_{K,\lambda}(\widehat{f}_{K,\lambda}) \leq \mathcal{C}_{K,\lambda}(f^*) = \sup_{g\in F} T_{K,\lambda}(g, f^*) \leq \max_{i\in[3]} \sup_{g\in F_i^{(1)}} T_{K,\lambda}(g, f^*),$$

where $\{F_1^{(1)}, F_2^{(1)}, F_3^{(1)}\}$ is the decomposition of F as in Figure 2. It follows from Lemma 3 (for $\kappa = 1$) that on the event $\Omega(K)$,

$$\mathcal{C}_{K,\lambda}(\widehat{f}_{K,\lambda}) \leqslant (2 + c')\epsilon r^2(\rho_K) \ .$$

Therefore, for $c' = 16$ and $c = 833$ the conclusion of the proof of Theorem 1 follows from Lemma 6.

4.5 Proof of Theorem 2

Define

$$K_1 = \frac{|\mathcal{O}|}{1-\gamma} = 8|\mathcal{O}| \text{ and } K_2 = \frac{N\alpha}{(2\theta_0\theta_{r0})^2} = \frac{N}{96(\theta_0\theta_{r0})^2}.$$

Let $K \in [K_1, K_2]$ and let $\Omega_{K,c_{ad}} = \{f^* \in \cap_{J=K}^{K_2} \hat{R}_{J,c_{ad}}\}$ where we recall that $\hat{R}_{J,c_{ad}} = \{f \in F : \mathcal{C}_{J,\lambda}(f) \leqslant (c_{ad}/\theta_0^2) r^2(\rho_J)\}$. Lemma 3 (for $\kappa = 1$) shows that, for $c_{ad} = (2 + c')/c$, $\Omega_{K,c_{ad}} \supset \cap_{J=K}^{K_2}\Omega(J)$. Therefore, on $\cap_{J=K}^{K_2}\Omega(J)$, $\hat{K}_{c_{ad}} \leqslant K$ which implies that $\widehat{f}_{c_{ad}} \in \hat{R}_{K,c_{ad}}$. By Lemma 6 (for $c' = 16$ and $c = 833$), this implies that

$$\left\|\widehat{f}_{c_{ad}} - f^*\right\| \leqslant 2\rho_K, \qquad \left\|\widehat{f}_{c_{ad}} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K) \quad \text{and} \quad R(\widehat{f}_{c_{ad}}) \leqslant R(f^*) + (1 + (4 + 3c')\epsilon) r^2(2\rho_K) \ .$$

A Simulation study

The aim of this section is to show that the min-max procedure introduced in this work can be computed using alternating descent-ascent algorithms. It appears that all the following algorithms can be recast as **block gradient descent (BGD)** but the major difference with the classical BGD is that blocks are chosen according to their "centrality" via the median operator instead of iterating over all the blocks (resp. at random) in the classical (resp. stochastic) BGD approach.

We test our algorithms in a high-dimensional framework. In this setup, the ℓ_1-norm has played a prominent role through the LASSO estimator. There has been a huge number of algorithms designed to implement the LASSO. The aim of this section is to show that there is a natural equivalent formulation to all of these algorithms for the MOM-LASSO that makes them more robust to outliers as can be appreciated in Figure 1. The choice of hyper-parameters like the number of blocks or the regularization parameter cannot be done via classical Cross-Validation approaches because of the potential presence of outliers in the test sets, CV procedures are adapted using MOM estimators. We also advocate for using random blocks at every iterations of the algorithms. This bypasses a problem of "local saddle points" we have identified. A by product of the latter approach is a definition of depth adapted to the learning task and therefore of an outliers detection algorithm.

A.1 Data generating process and corruption by outliers

We study the performance of all the algorithms on a dataset corrupted by outliers of various forms. The "basic" set of "informative data" is called $\mathcal{D}_1$. This set is corrupted by various datasets of outliers, named $\mathcal{D}_2, \mathcal{D}_3$, $\mathcal{D}_4$ and $\mathcal{D}_5$. Good and bad data have been merged and shuffled in the dataset $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3 \cup \mathcal{D}_4$ given to the statistician. Let us now detail the construction of these datasets.

1. The set $\mathcal{D}_1$ of "informative data" is a set of N_{good} i.i.d. data (X_i, Y_i) with common distribution
$$Y = \left\langle X, t^* \right\rangle + \zeta \ , \tag{22}$$
where $t^* \in \mathbb{R}^d$, $X \sim \mathcal{N}(0, I_{d\times d})$ and $\zeta \sim \mathcal{N}(0, \sigma^2)$ is independent of X.
2. $\mathcal{D}_2$ is a dataset of N_{bad-1} "bad data" (X_i, Y_i) such that $Y_i = 1$ and $X_i = (1)_{j=1}^d$
3. $\mathcal{D}_3$ is a dataset of N_{bad-2} "bad data" (X_i, Y_i) such that $Y_i = 10000$ and $X_i = (1)_{j=1}^d$
4. $\mathcal{D}_4$ is a dataset of N_{bad-3} "bad data" (X_i, Y_i) where Y_i is a $0-1$-Bernoulli random variable and X_i is uniformly distributed over $[0,1]^d$,
5. $\mathcal{D}_5$ is also a set of "outliers" that have been generated according to a linear model as in (22) (i.e. with the same target vector t^*) but for a different choice of design X and noise ζ. Here, we take $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = (\rho^{|i-j|})_{1\leqslant i,j\leqslant d}$ and ζ is a heavy-tailed noise distributed according to a Student distribution with various degrees of freedom.

These different types of "outliers" $\mathcal{D}_j, j = 2,3,4,5$ have been chosen to illustrate that the theory allows for outliers that may have absolutely nothing to do with the oracle t^* that can be neither independent nor random as illustrated by datasets $\mathcal{D}_2$ and $\mathcal{D}_3$.

A.2 From algorithms for the LASSO to their "MOM versions"

There is a huge literature presenting many algorithms to implement the LASSO procedure. We explore several of them to investigate their performance regarding robustness.

Each algorithm designed for the LASSO can be transformed into an algorithm for the min-max estimator we have been studying in this work. Let us now explain how this can be done. Recall that MOM version of the LASSO estimator is
$$\hat{t}_{K,\lambda} \in \underset{t\in\mathbb{R}^d}{\operatorname{argmin}} \sup_{t'\in\mathbb{R}^d} T_{K,\lambda}(t', t) \tag{23}$$
where $T_{K,\lambda}(t', t) = \text{MOM}_K\big(\ell_t - \ell_{t'}\big) + \lambda\left(\|t\|_1 - \|t'\|_1\right)$, $\text{MOM}_K\big(\ell_t - \ell_{t'}\big)$ is a median of the set of real numbers $\{P_{B_1}(\ell_t - \ell_{t'}), \cdots, P_{B_K}(\ell_t - \ell_{t'})\}$ and for all $k \in [K]$,
$$P_{B_k}(\ell_t - \ell_{t'}) = \frac{1}{|B_k|} \sum_{i\in B_k} (Y_i - \left\langle X_i, t\right\rangle)^2 - (Y_i - \left\langle X_i, t'\right\rangle)^2.$$

A natural idea to implement (23) is to consider algorithms based on a sequence of alternating descents (in t) and ascents (in t') steps with or without a "proximal/projection" step and for various choices of "step sizes". A key issue here is that the "marginal" functions $t \to T_{K,\lambda}(t_0', t)$ and $t' \to T_{K,\lambda}(t', t_0)$, for some given (t_0, t_0'), may not be convex. Nevertheless, one can still locally compute the steepest descent by assuming that the index in $[K]$ of the block achieving the median in $\text{MOM}_K\big(\ell_{t_0} - \ell_{t_0'}\big)$ remains constant on a convex open set containing (t_0, t_0').

Assumption 4. *Almost surely (with respect to $(X_i, Y_i)_{i=1}^N$) for almost all $(t_0, t_0') \in \mathbb{R}^d \times \mathbb{R}^d$ (with respect to the Lebesgue measure on $\mathbb{R}^d \times \mathbb{R}^d$), there exists a convex open set B containing (t_0, t_0') and $k \in [K]$ such that for all $(t, t') \in B$, $P_{B_k}(\ell_t - \ell_{t'}) \in MOM_K\big(\ell_t - \ell_{t'}\big)$.*

Under Assumption 4, for $\lambda\otimes\lambda$-almost all couples $(t_0, t_0') \in \mathbb{R}^d \times \mathbb{R}^d$ (where λ is the Lebesgue measure on $\mathbb{R}^d$), $t \to T_{K,\lambda}(t_0', t)$ is "locally strictly convex" and $t' \to T_{K,\lambda}(t', t_0)$ is "locally strictly concave". Therefore, for the choice of index k such that $P_{B_k}(\ell_{t_0} - \ell_{t_0'}) \in \text{MOM}_K\big(\ell_{t_0} - \ell_{t_0'}\big)$, we have
$$\nabla_t \text{MOM}_K\big(\ell_t - \ell_{t_0'}\big)_{|t=t_0} = -2(X^{(k)})^\top (Y^{(k)} - X^{(k)} t_0) \tag{24}$$
where $Y^{(k)} = (Y_i)_{i\in B_k}$ and $X^{(k)}$ is the $|B_k|\times d$ matrix with rows given by $X_i^\top$ for $i \in B_k$. The integer $k \in [K]$ is the index of the median of K real numbers $P_{B_1}(\ell_t - \ell_{t'}), \cdots, P_{B_K}(\ell_t - \ell_{t'})$, which is straightforward to compute. The gradient $-2(X^{(k)})^\top (Y^{(k)} - X^{(k)} t_0)$ in (24) depends on t_0' only through the index k.

Remark 4 (Block Gradient Descent and map-reduce)**.** *Algorithms developed for the minmax estimator can be interpreted as block gradient descent. The major difference with the classical Block Gradient descent (which takes sequentially all the blocks one after another), is that the index of the block is chosen here at each round according to a "centrality measure": the index $k \in [K]$ of the block B_k chosen to make one more descent / ascent step satisfies $P_{B_k}(\ell_{t_0} - \ell_{t'_0}) \in MOM_K\big(\ell_{t_0} - \ell_{t'_0}\big)$. In particular, we expect blocks corrupted by outliers not to be chosen whereas in the classical BGD all blocks are chosen at each pass. Moreover, by choosing the "descent / ascent" block k using this centrality measure we also expect $P_{B_k}(\ell_{t_0} - \ell_{t'_0})$ to be close to the true expectation $P(\ell_{t_0} - \ell_{t'_0})$ which is ultimately the function we would like to know. This makes every descent (resp. ascent) steps particularly efficient since the right descent (resp. ascent) direction is $-\nabla_t P(\ell_t - \ell_{t'_0})_{|t=t_0}$ (resp. $\nabla_{t'} P(\ell_{t_0} - \ell_{t'})_{|t'=t'_0}$).*

Moreover, our algorithms particularly fits the "big data" framework which is our original motivation for the introduction of robust procedures in machine learning. In this framework, the map-reduce paradigm has emerged as a leading idea to design procedures. In this situation, the data are spread out in a cluster of servers and are therefore naturally split into blocks. Then our procedures simply uses for mapper a mean function and for reducer a median function. This makes our algorithms easily scalable into the big data framework even when some servers have crashed down (making outliers data). MOM algorithm could therefore reduced the maintenance of big clusters.

Remark 5 (Normalization)**.** *In the classical i.i.d. setup, the design matrix $\mathbb{X}$ (i.e., the $N \times d$ matrix with row vectors $X_1, \dots, X_N$) is usually normalized so that the ℓ_2^N-norms of the columns equal to one. In our corrupted setup, we cannot normalize the design matrix this way because one row of $\mathbb{X}$ may be corrupted. In that case, normalizing each column of $\mathbb{X}$ would corrupt the entire matrix $\mathbb{X}$. $\mathbb{X}$ has to be kept as it is in all simulations.*

In the sequel, we use this strategy to transform several algorithms implemented for the LASSO into algorithms for the min-max estimator (23). We first start with the subgradient descent algorithm.

A.3 Subgradient descent algorithm

The LASSO is solution of the minimization problem $\min_{t \in \mathbb{R}^d} F(t)$ where F is defined for all $t \in \mathbb{R}^d$ by $F(t) = \|\mathbb{Y} - \mathbb{X}t\|_2^2 + \lambda \|t\|_1$ with $\mathbb{Y} = (Y_i)_{i=1}^N$ and $\mathbb{X}$ is the $N \times d$ matrix with row vectors $X_1, \dots, X_N$. The LASSO can be approximated using a subgradient descent procedure : given a starting point $t_0 \in \mathbb{R}^d$ and $(\gamma_p)_p$ a sequence of step sizes (i.e. $\gamma_p > 0$ and $(\gamma_p)_p$ decreases), at step p we update

$$t_{p+1} = t_p - \gamma_p \partial F(t_p) \tag{25}$$

where $\partial F(t_p)$ is a subgradient of F at t_p, for instance, $\partial F(t_p) = -2\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}t_p) + \lambda \text{sign}(t_p)$ where $\text{sign}(t_p)$ is the vector of signs of the coordinates of t_p with the convention $\text{sign}(0) = 0$. The sub-gradient descent algorithm (25) can be turned into an alternating subgradient ascent/descent algorithm for the min-max estimator

$$\hat{t}_{K,\lambda} \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \sup_{t' \in \mathbb{R}^d} \big(\text{MOM}_K\big(\ell_t - \ell_{t'}\big) + \lambda \left(\|t\|_1 - \left\|t'\right\|_1\right)\big) \enspace .$$

Let

$$\mathbb{Y}_k = (Y_i)_{i \in B_k} \text{ and } \mathbb{X}_k = (X_i^\top)_{i \in B_k} \in \mathbb{R}^{|B_k| \times d} \enspace . \tag{26}$$

```
  input  : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point
           $\epsilon > 0$ : a stopping criteria
           $(\eta_p)_p, (\beta_p)_p$: two step size sequences
  output: approximated solution to the min-max problem (23)
1 while $\|t_{p+1} - t_p\|_2 \geqslant \epsilon$ or $\|t'_{p+1} - t'_p\|_2 \geqslant \epsilon$ do
2     find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_p}) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$
3         $$t_{p+1} = t_p + 2\eta_p \mathbb{X}_k^\top (\mathbb{Y}_k - \mathbb{X}_k t_p) - \lambda \eta_p \mathrm{sign}(t_p)$$
4     find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_{p+1}}) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p})$
5         $$t'_{p+1} = t'_p + 2\beta_p \mathbb{X}_k^\top (\mathbb{Y}_k - \mathbb{X}_k t'_p) - \lambda \beta_p \mathrm{sign}(t'_p)$$
6 end
7 Return $(t_p, t'_p)$
```

Algorithm 1: An alternating sub-gradient descent algorithm for the minimax MOM estimator (23).

The key insight in Algorithm 1 are step 2 and step 4 where the blocks number have been chosen according to their centrality among the other blocks via the median operator. Those steps are expected 1) to remove outliers from the descent / ascent directions 2) to improve the accuracy of the latter directions.

A classical choice of step size γ_p in (25) is $\gamma_p = 1/L$ where $L = \|\mathbb{X}\|_{S_\infty}^2$ ($\|\mathbb{X}\|_{S_\infty}$ is the operator norm of $\mathbb{X}$). Another possible choice follows from the Armijo-Goldstein condition with the following backtracking line search: γ is decreased geometrically while the Armijo-Goldstein condition is not satisfied

$$\textbf{while} \quad F(t_p + \gamma_\ell \partial F(t_p)) > F(t_p) + \delta \gamma_\ell \|\partial F(t_p)\|_2^2 \quad \textbf{do} \quad \gamma_{\ell+1} = \rho \gamma_\ell \tag{27}$$

for some given $\rho \in (0,1)$, $\delta = 10^{-4}$ and initial point $\gamma_0 = 1$.

Of course, the same choice of step size can be made as well for $(\eta_p)_p$ and $(\beta_p)_p$ in Algorithm 1. In the first case, one can take $\eta_p = 1/\|\mathbb{X}_k\|_{S_\infty}^2$ where $k \in [K]$ is the index defined in line 2 of Algorithm 1 and $\beta_p = 1/\|\mathbb{X}_k\|_{S_\infty}^2$ where $k \in [K]$ is the index defined in line 4 of Algorithm 1. In the other backtracking line search case, the Armijo-Goldstein condition adapted for Algorithm 1 reads like

$$\textbf{while} \quad F_k(t_p + \gamma_\ell \partial F_k(t_p)) > F_k(t_p) + \delta \gamma_\ell \|\partial F_k(t_p)\|_2^2 \quad \textbf{do} \quad \eta_{\ell+1} = \rho \eta_\ell \tag{28}$$

where $F_k(t) = \|\mathbb{Y}_k - \mathbb{X}_k t\|_2^2 + \lambda \|t\|_1$ where $k \in [K]$ is the index defined in line 2 of Algorithm 1 and a similar update follows for β_p with an index $k \in [K]$ as defined in line 4 of Algorithm 1.

A.4 Proximal gradient descent algorithms

In this section, we provide a MOM version of the classical ISTA and FISTA algorithms. Recall that ISTA (and its accelerated version FISTA) are proximal gradient descent that fall in the general class of splitting algorithms where one usually decomposes the objective function $F(t) = f(t) + g(t)$ with $f(t) = \|\mathbb{Y} - \mathbb{X}t\|_2^2$ (convex and differentiable) and $g(t) = \lambda \|t\|_1$ (convex).

ISTA stands for Iterative Shrinkage-Thresholding Algorithm. It is a splitting algorithm that alternates between a descent step in the direction of the gradient and a "projection step" through the proximal operator of g which is the soft-thresholding operator in the case of the ℓ_1-norm : at step p

$$t_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\Big(t_p + 2\gamma_p \mathbb{X}^\top(\mathbb{Y} - \mathbb{X}t_p)\Big) \tag{29}$$

where $\mathrm{prox}_{\lambda\|\cdot\|_1}(t) = (\mathrm{sign}(t_j)\max(|t_j| - \lambda, 0))_{j=1}^d$ for all $t = (t_j)_{j=1}^d \in \mathbb{R}^d$. A natural candidate for (23) is given by the following alternating method.

```
input  : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point
         $\epsilon > 0$ : a stopping criteria
         $(\eta_k)_k, (\beta_k)_k$: two step size sequences
output: approximated solution to the min-max problem (23)
while $\|t_{p+1} - t_p\|_2 \geqslant \epsilon$ or $\|t'_{p+1} - t'_p\|_2 \geqslant \epsilon$ do
    find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_p}) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$
        $t_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\Big(t_p + 2\eta_k \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t_p)\Big)$
    find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_{p+1}}) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p})$
        $t'_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\Big(t'_p + 2\beta_k \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t'_p)\Big)$
end
Return $(t_p, t'_p)$
```

Algorithm 2: An alternating proximal gradient descent for the minimaximization procedure (23).

Note that the step sizes sequences $(\eta_p)_p$ and $(\beta_p)_p$ may be chosen according to the remarks below Algorithm 1.

A.5 Douglas-Racheford / ADMM

In this section, we consider the ADMM algorithm. ADMM stands for Alternating Direction Method of Multipliers. It is also a splitting algorithm which reads as follows in the LASSO case: at step p,

$$\begin{aligned} t_{p+1} &= (\mathbb{X}^\top \mathbb{X} + \rho I_{d\times d})^{-1}(\mathbb{X}^\top \mathbb{Y} + \rho z_p - u_p) \\ z_{p+1} &= \mathrm{prox}_{\lambda\|\cdot\|_1}(t_{p+1} + u_p/\rho) \\ u_{p+1} &= u_p + \rho(t_{p+1} - z_{p+1}) \end{aligned} \tag{30}$$

where ρ is some parameter to be chosen (for instance $\rho = 10$). The ADMM algorithm returns t_p after a stopping criteria is met. In Algorithm 4, we provide a MOM version of this algorithm.

```
input  : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point
         $\epsilon > 0$ : a stopping criteria
         $\rho$: a parameter
output: approximated solution to the min-max problem (23)
while $\|t_{p+1} - t_p\|_2 \geqslant \epsilon$ or $\|t'_{p+1} - t'_p\|_2 \geqslant \epsilon$ do
    find $k \in [K]$ such that $\text{MOM}_K\big(\ell_{t'_p} - \ell_{t_p}\big) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$
        $t_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p - u_p)$
        $z_{p+1} = \text{prox}_{\lambda\|\cdot\|_1}(t_{p+1} + u_p/\rho)$
        $u_{p+1} = u_p + \rho(t_{p+1} - z_{p+1})$
    find $k \in [K]$ such that $\text{MOM}_K\big(\ell_{t'_p} - \ell_{t_{p+1}}\big) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p})$
        $t'_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z'_p - u'_p)$
        $z'_{p+1} = \text{prox}_{\lambda\|\cdot\|_1}(t'_{p+1} + u'_p/\rho)$
        $u'_{p+1} = u'_p + \rho(t'_{p+1} - z'_{p+1})$
end
Return $(t_p, t'_p)$
```

Algorithm 3: An ADMM algorithm for the minimaximization MOM estimator (23).

A.6 Cyclic coordinate descent

Repeatedly minimizing the LASSO objective function $t \to \|\mathbb{Y} - \mathbb{X}t\|_2^2 + \lambda \|t\|_1$ w.r.t. to each coordinate t_j has proved to be an efficient algorithm. The closed form solution (cf. for instance [31, p. 83]) is given by

$$t_j = \frac{R_j}{\|\mathbb{X}_{\cdot j}\|_2^2}\left(1 - \frac{\lambda}{2|R_j|}\right)_+ \quad \text{with } R_j = \mathbb{X}_{\cdot j}^\top\left(\mathbb{Y} - \sum_{k\neq j} t_k \mathbb{X}_{\cdot k}\right).$$

We can adapt this algorithm to approximate the MOM estimator (23). Denote by $\mathbb{X}_{kj}$ the j-th column of $\mathbb{X}_k$ for all $j \in [d]$.

```
input  : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$ : initial point
         $\epsilon > 0$ : a stopping criteria
output: approximated solution to the min-max problem (23)
while $\|t_{p+1} - t_p\|_2 \geqslant \epsilon$ or $\|t'_{p+1} - t'_p\|_2 \geqslant \epsilon$ do
    find $k \in [K]$ such that $\mathrm{MOM}_K\big(\ell_{t'_p} - \ell_{t_p}\big) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$
    for $j$ in $[d]$ do
        $t_{p+1,j} = \frac{R_{kj}}{\|\mathbb{X}_{kj}\|_2^2}\left(1 - \frac{\lambda}{2|R_{kj}|}\right)_+$ with $R_{kj} = \mathbb{X}_{kj}^\top\left(\mathbb{Y}_k - \sum_{q<j} t_{p+1,q}\mathbb{X}_{kq} - \sum_{q>j} t_{p,q}\mathbb{X}_{kq}\right)$
    end
    find $k \in [K]$ such that $\mathrm{MOM}_K\big(\ell_{t'_p} - \ell_{t_{p+1}}\big) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p})$
    for $j$ in $[d]$ do
        $t'_{p+1,j} = \frac{R_{kj}}{\|\mathbb{X}_{kj}\|_2^2}\left(1 - \frac{\lambda}{2|R_{kj}|}\right)_+$ with $R_{kj} = \mathbb{X}_{kj}^\top\left(\mathbb{Y}_k - \sum_{q<j} t'_{p+1,q}\mathbb{X}_{kq} - \sum_{q>j} t'_{p,q}\mathbb{X}_{kq}\right)$
    end
end
Return $(t_p, t'_p)$
```

Algorithm 4: A Cyclic Coordinate Descent algorithm for the minimax MOM estimator (23).

A.7 Adaptive choice of hyper-parameters via MOM V-fold Cross Validation

The choice of hyperparameters in a “data corrupted” environment has to be done carefully. Classical Cross-validation methods cannot be used because of the potential presence of outliers in the dataset that are likely to corrupt the classical CV-criterion. We design new (empirical) criteria that trustfully reveal performance of estimators even in situations where “test datasets” might have been corrupted.

MOM’s principles can be combined with the idea of multiple splitting into training / test datasets in cross-validation. Let us now explain the construction of estimators $\hat{f}_{\hat{K},\hat{\lambda}}$ where the number of blocks $\hat{K}$ and the regularization parameter $\hat{\lambda}$ are hyper-parameters learned via such version of the CV principle.

We are given an integer $V \in [N]$ such that N is divided by V. We are also given two finite grids $\mathcal{G}_K \subset [N]$ and $\mathcal{G}_\lambda \subset (0, 1]$. Our aim is to chose the “best” numbers of blocks and best regularization parameter within both grids. The dataset is splitted into V disjoints blocks $\mathcal{D}_1, \dots, \mathcal{D}_V$. For each $v \in [V]$, $\cup_{u \neq v}\mathcal{D}_u$ is used to train a family of estimators

$$\left(\hat{f}_{K,\lambda}^{(v)} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda\right). \tag{31}$$

The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in the family (31). Using these notations, we can define a MOM version of the cross-validation procedure.

Definition 7. *The* ***Median of Means V-fold Cross Validation procedure*** *associated to the family of estimators* (31) *is* $\hat{f}_{\hat{K},\hat{\lambda}}$ *where* $(\hat{K}, \hat{\lambda})$ *is minimizing the* MomCv_V *criteria*

$$(K, \lambda) \in \mathcal{G}_K \times \mathcal{G}_\lambda \to \mathrm{MomCv}_V(K, \lambda) = Q_{1/2}\left(\mathrm{MOM}_{K'}^{(v)}\left(\ell_{\hat{f}_{K,\lambda}^{(v)}}\right)_{v \in [V]}\right),$$

where, for all $v \in [V]$ *and* $f \in F$,

$$\mathrm{MOM}_{K'}^{(v)}(\ell_f) = \mathrm{MOM}_{K'}\left(P_{B_1^{(v)}}\ell_f, \cdots, P_{B_{K'}^{(v)}}\ell_f\right) \tag{32}$$

and $B_1^{(v)} \cup \cdots, \cup B_{K'}^{(v)}$ *is a partition of the test set* $\mathcal{D}_v$ *into* K' *blocks where* $K' \in [N/V]$ *such that* K' *divides* N/V.

The difference between standard V-fold cross validation procedure and its MOM version in Definition 7 is that empirical means on test sets $\mathcal{D}_v$ in the classical V-fold CV procedure have been replaced by MOM estimators in (32). Moreover, the mean over all V splits in the classical V-fold CV is replaced by a median.

The choice of V raises the same issues for MOM CV as for classical V-fold CV [3, 4]. In the simulations we use $V = 5$. The construction of the MOM-CV requires to choose another parameter: K', the number of blocks used to construct the MOM criteria (32) over the test set. A possible solution is to take $K' = K/V$. This has the advantage to make only one split of the dataset $\mathcal{D}$ into K blocks and then use for each of the V rounds, $(V-1)K/V$ of these blocks to construct the family of estimators (31) and then K/V of these blocks to test them.

In Figures 3 and 4, hyper-parameters K (i.e. the number of blocks) and λ (i.e. the regularization parameter) have been chosen for the MOM LASSO estimator via the MOM V-fold Cross validation procedure introduced above.

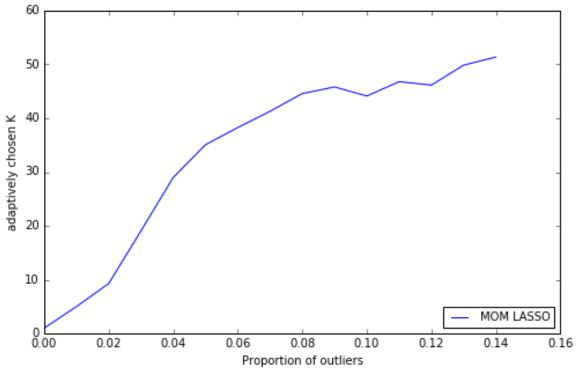

Figure 3: Adaptively chosen number of blocks K for the MOM LASSO.

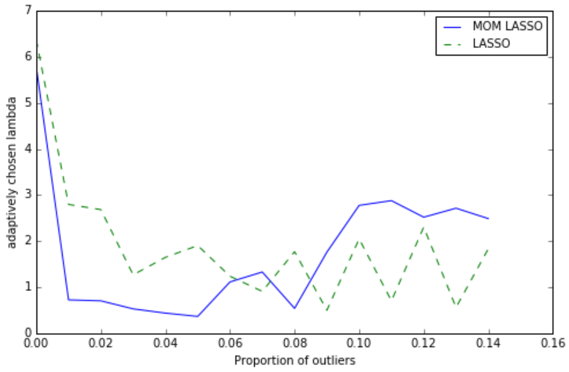

Figure 4: Adaptively chosen λ for the LASSO and MOM LASSO

In Figure 3, the adaptively chosen $\hat{K}$ grows with the number of outliers which is something expected since one needs the number of blocks to be at least twice the number of outliers.

A.8 Maximinimization, saddle-point, random blocks, outliers detection and depth

In the previous sections, we considered a setup that particularly fits a dataset distributed on clusters. This distributed source of data yields some specific partition of the dataset in compliance with the clusters structure – in other words, the blocks of data $B_1, \cdots, B_K$ are imposed and fixed by the specific physical form of the dataset. In a batch setup, there is a priori no structural restriction to choose a fixed partition of the dataset given that it has no predefined organization. It appears that, in this setup, there is a way to improve the performance of the various iterative algorithms constructed in the previous sections by choosing randomly the blocks at every (descent and ascent) steps of the algorithm. The aim of this section is to show some advantages of this choice and how this modified version works on the example of ADMM. Moreover, as a byproduct of this approach, it is possible to construct an outliers detection algorithm. This algorithm outputs a score to each data in the dataset measuring its centrality. In particular, data with a low score should be considered as outliers.

Note that the way we introduced the minimaximization estimator in Section 1 was based on the observation that the oracle is solution to the minmaximization problem $f^* \in \operatorname{argmin}_{f\in F} \sup_{g\in F} P(\ell_f - \ell_g)$. But it appears that f^* is also solution to a maxminimization problem: $f^* \in \operatorname{argmax}_{g\in F} \inf_{f\in F} P(\ell_f - \ell_g)$. This observation naturally yields another estimator: the maxmin estimator

$$\hat{g}_{K,\lambda} \in \underset{g\in F}{\operatorname{argmax}} \inf_{f\in F} T_{K,\lambda}(g, f). \tag{33}$$

Following the same strategy as in Section 4, we can show that $\hat{g}_{K,\lambda}$ has the very same statistical performance as the estimator $\hat{f}_{K,\lambda}$ from Section 3.4 (cf. Section C in the case where there is no regularization for a proof). Nevertheless, there is a priori no reason that $\hat{g}_{K,\lambda}$ and $\hat{f}_{K,\lambda}$ are the same estimator. That is we don't know in advance that

$$\operatorname*{argmin}_{f\in F}\sup_{g\in F} T_{K,\lambda}(g,f) = \operatorname*{argmax}_{g\in F}\inf_{f\in F} T_{K,\lambda}(g,f). \tag{34}$$

In other words there is no evidence that the duality gap is null. Since $T_{K,\lambda}(g,f) = -T_{K,\lambda}(f,g)$, (34) holds if and only if

$$\inf_{f\in F}\sup_{g\in F} T_{K,\lambda}(f,g) = 0.$$

In that case, $\hat{f}$ is by definition a **saddle-point** estimator and therefore the two sets of minmax and maxmin estimators are equal.

Remark 6 (stopping criteria)**.** *Along iterations of the previous algorithms it is possible to track down the values of the objective function* $MOM_K\big(\ell_{t_p} - \ell_{t'_p}\big) + \lambda\left(\|t_p\|_1 - \|t'_p\|_1\right)$. *When this one is close to zero this means that it has achieved a saddle-point and therefore the output of the algorithm is close to the solution of the minmax problem (as well as the maxmin problem).*

When $\hat{f}$ is a saddle-point, the choice of fixed blocks $B_1,\dots,B_K$ may result in a problem of "**local saddle points**": iterations of our original algorithms remain close to a potentially suboptimal local saddle point.

To see this, let us consider the vector case (that is for $f(\cdot) = \langle\cdot,t\rangle$ for $t\in\mathbb{R}^d$). We introduce the following sets: for all $k\in[K]$,

$$\mathcal{C}_k = \left\{(t,t')\in\mathbb{R}^d\times\mathbb{R}^d : \mathrm{MOM}_K\big(\ell_t-\ell_{t'}\big) = P_{B_k}(\ell_t-\ell_{t'})\right\}. \tag{35}$$

It is clear that $\cup_{k\in[K]}\mathcal{C}_k = \mathbb{R}^d\times\mathbb{R}^d$. The key idea behind all the previous algorithms is that at each step the current iteration (t_p,t'_p) lies in one of those cells $\mathcal{C}_k$ and that, if Assumption 4 holds, for almost all iterations there will be an open ball around (t_p,t'_p) contained in the cell $\mathcal{C}_k$ so that the objective function is locally equal to $(t,t')\to P_{B_k}(\ell_t-\ell_{t'}) + \lambda(\|t\|_1 - \|t'\|_1)$ which is a convex-concave function.

The problem here is that our algorithms are looking for saddle-points so that if there are several cells $\mathcal{C}_k$ containing saddle-points the previous algorithms may be stuck in one of them whereas a "better saddle-point" (that is a saddle point closer to t^*) may be in an other cell.

To overcome this issue, we choose at every (descent and ascent) steps of the previous algorithms a random partition of the dataset into K blocks. The decomposition of the space $\mathbb{R}^d\times\mathbb{R}^d$ into cells $\mathcal{C}_1,\cdots,\mathcal{C}_K$ does not exist anymore since the cells are different (randomly chosen) at every steps. As an example, we develop the ADMM procedure with a random choice of blocks In Algorithm 5.

```
input  : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$: initial point
         $\epsilon > 0$: a stopping criteria
         $\rho$: parameter
output: approximated solution to the min-max problem (23)
while $\|t_{p+1} - t_p\|_2 \geqslant \epsilon$ or $\|t'_{p+1} - t'_p\|_2 \geqslant \epsilon$ do
    Partition the datasets into $K$ blocks $B_1, \dots, B_K$ of equal size at random.
    Find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_p}) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$
        $t_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p - u_p)$
        $z_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}(t_{p+1} + u_p/\rho)$
        $u_{p+1} = u_p + \rho(t_{p+1} - z_{p+1})$
    Partition the datasets into $K$ blocks $B_1, \dots, B_K$ of equal size at random.
    Find $k \in [K]$ such that $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_{p+1}}) = P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p})$
        $t'_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z'_p - u'_p)$
        $z'_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}(t'_{p+1} + u'_p/\rho)$
        $u'_{p+1} = u'_p + \rho(t'_{p+1} - z'_{p+1})$
end
Return $(t_p, t'_p)$
```

Algorithm 5: The ADMM algorithm for the minimax MOM estimator (23) with a random choice of blocks at each steps.

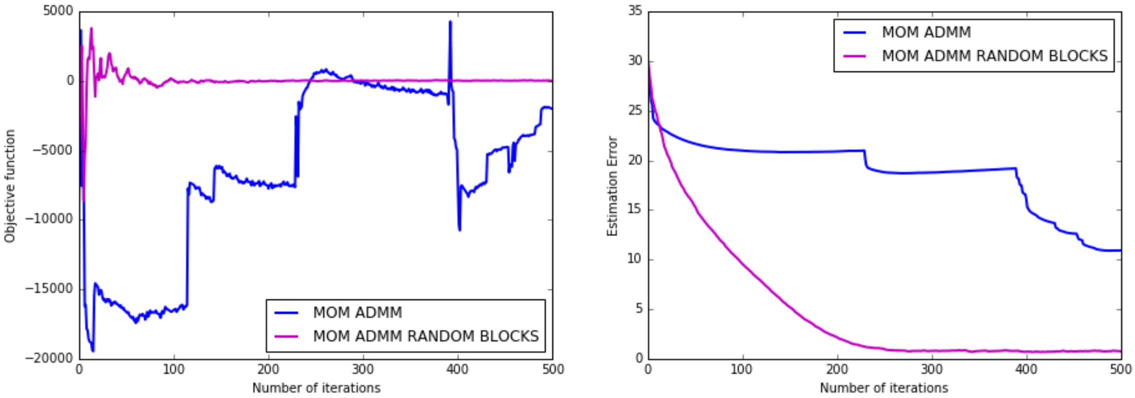

Figure 5: Fixed blocks against random blocks.

In Figure 5, we ran both MOM LASSO procedures via ADMM with fixed and random blocks. Both the objective function and the estimation error of MOM LASSO jump in the case of fixed blocks. These jumps correspond to a change of cell number for these iterations. The algorithm converge to local saddle-points before jumping to other cells. This slows down its convergence. On the other hand, the algorithms with random blocks do not suffer this drawback. Figure 5 shows that the estimation error converges much faster and smoothly for random blocks than for fixed blocks. As a conclusion choosing blocks at random improve both stability and speed of the algorithm, avoiding the issue of "local saddle points" in the cells $\mathcal{C}_k$. Note also that the objective function of the MOM ADMM with random blocks tends to zero so that the duality gap tends to zero (meaning that we do have a natural stopping criterium and that the MOM LASSO is a saddle point).

A byproduct of this approach is that one can construct an **outliers detection procedure**. To that end one simply has to count the number of times each data is selected at steps 2 and 4 of Algorithm 5. Every data starts with a null score. Then, every time a block B is selected as median block, every data it contains increases its score by one. At the end, every data ends up with a score revealing its centrality for our learning task.

It is expected that aggressive outliers corrupt their respective blocks. These "corrupted blocks" will not be median blocks and will not be chosen. In the case of fixed blocks, informative data cannot be distinguished from outliers lying in the same block, therefore, this outliers detection algorithm only makes sense when blocks are chosen at random. Figure 6 shows performance of this strategy on synthetic data (cf. Section A.9 for more details on the simulations). On this example, outliers (data number $1, 32, 170$ and 194) end up with a null score meaning that they have never been selected along the descent / ascent iterations of the algorithm. Rearranging the scores, one can observe a jump in the score function between outliers and informative data. The gap between these scores can be enlarged by running more iterations of the algorithm. Finally, the score may be seen as a **depth** of a data point (X_i, Y_i), the most selected data are the most central.

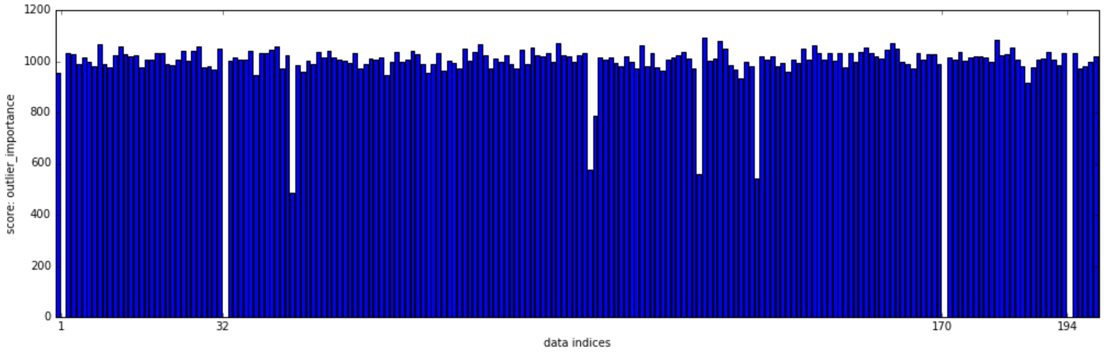

Figure 6: Outliers detection algorithm. The dataset has been corrupted by 4 outliers at number $1, 32, 170$ and 194. The score of the outliers is 0: they haven't been selected even once.

A.9 Simulations setup for the figures

All codes used to obtain the figures in this work are available at [1] and can therefore be used to reproduce all the figures. Let us now precise the parameters of the simulations in Figure 1, Figure 3 and Figure 4: the number of observations is $N = 200$, the number of features is $d = 500$, we construct a sparse vector $t^* \in \mathbb{R}^d$ with sparsity $s = 10$ and support chosen at random and non-zero coordinates t_j^* being either equal to 10, -10 or decreasing according to $\exp(-j/10)$. We end up with a set of informative data $\mathcal{D}_1$ as described in Section A with noise variance $\sigma = 1$. Then this dataset is increasingly corrupted with outliers as in $\mathcal{D}_4$ (cf. Section A): outliers are data with output $Y = 10000$ and input $X = (1)_{j=1}^d$.

The proportion of outliers are $0, 1/100, 2/100, \dots, 15/100$. On each dataset $\mathcal{D}_1 \cup \mathcal{D}_4$ obtained after corruption of $\mathcal{D}_1$ for various proportion of outliers, we run the ADMM algorithm with adaptively chosen regularization parameter λ via V-fold CV with $V = 5$ for the LASSO. Then we run the MOM ADMM with adaptively chosen number of blocks K and regularization parameter λ via the MOM CV procedure as introduced in Section A.7 with $V = 5$ and $K' = \max(\text{grid}_K)/V$ where $\text{grid}_K = \{1, 4, \cdots, 115/4\}$ and $\text{grid}_\lambda = \{0, 10, 20 \cdots, 100\} / \sqrt{100}$ are the search grids used to select the best K and λ during the CV and MOM CV steps. The number of iterations of ADMM and MOM ADMM was taken equal to 200. Those simulations have been run 70 times and the averaged values of the estimation error, adaptively chosen K and λ have been reported in Figure 1, Figure 3 and Figure 4. The ℓ_2 estimation error of the LASSO increases roughly from 0 when there is no outlier and stabilize at 550 right after a single outlier enter the dataset. The ℓ_2 estimation error value 550 can be explained by the fact that the outliers which was added is $Y = 10000$ and $X = (1)_{j=1}^{500}$ which satisfies $Y = \langle X, t^{**} \rangle$ for $t^{**} = (20)_{j=1}^{500}$ which is the solution with minimal ℓ_1^d norm among all the solutions $t \in \mathbb{R}^d$ such that $Y = \langle X, t \rangle$. It appears that $\|t^{**} - t^*\|_2$ is of the order of 550. It means that the LASSO is actually solving the solution to the linear program associated with the outlier instead of solving the linear problem associated with the 200 other informative data. A single outliers is therefore completely misleading the LASSO.

For Figure 5, we have ran similar experiments with $N = 200$, $d = 300$, $s = 20$, $\sigma = 1$, $K = 10$, the number of iterations was 500 and the regularization parameter was $1/\sqrt{N}$.

For Figure 6, we took $N = 200$, $d = 500$, $s = 20$, $\sigma = 1$, the number of outliers is $|\mathcal{O}| = 4$ and the outliers are of the form $Y = 10000$ and $X = (1)_{j=1}^{d}$, $K = 10$, the number of iterations is 5.000 and $\lambda = 1/\sqrt{200}$.

B Examples of applications

This section is dedicated to some applications of Theorem 2. We present two classical examples of regularization in high-dimensional statistics: the ℓ_1-norm and the SLOPE norm. The associated RERM have been extensively studied in the literature under strong assumptions. The aim of this section is to show that these assumptions can be strongly relaxed when dealing with MOM versions. For instance, as observed in Figure 1, the performance of the LASSO are drastically deteriorated in the presence of a single outlier while its MOM version remains informative.

B.1 The LASSO

The LASSO is obtained when $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ and the regularization function is the ℓ_1-norm :

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\langle t, X_i \rangle - Y_i \right)^2 + \lambda \|t\|_1 \right), \quad \text{where} \quad \|t\|_1 = \sum_{i=1}^{d} |t_i| \ .$$

Even if recent advances [98, 91, 74] have shown some limitations of LASSO, it remains the benchmark estimator in high-dimensional statistics because a high dimensional parameter space does not significantly affect its performance as long as t^* is sparse. This was shown for example, in [14, 58, 93, 94, 65, 72, 90] for estimation and sparse oracle inequalities, in [64, 99, 6] for support recovery results; more results and references on LASSO can be found in the books [19, 44].

B.2 SLOPE

SLOPE is an extension of the LASSO that was introduced in [16, 84]. The class F is still $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ and the regularization function is defined for parameters $\beta_1 \geqslant \beta_2 \geqslant ... \geqslant \beta_d > 0$ by

$$\|t\|_{SLOPE} = \sum_{i=1}^{d} \beta_i t_i^\sharp,$$

where $(t_i^\sharp)_{i=1}^d$ denotes the non-increasing re-arrangement of $(|t_i|)_{i=1}^d$. SLOPE norm is a weighted ℓ_1-norm that coincide with ℓ_1-norm when $(\beta_1, ..., \beta_d) = (1, ..., 1)$.

B.3 Classical results for LASSO and SLOPE

Typical results for LASSO and SLOPE have been obtained in the i.i.d. setup under a subgaussian assumption on the design X and, most of the time, on the noise ζ as well. Let us for the moment provide a definition of this assumption and recall the one of isotropicity.

Definition 8. *Let ℓ_2^d be a d-dimensional inner product space and let X be random variable with values in ℓ_2^d. We say that X is isotropic when for every $t \in \ell_2^d$, $\|\langle X, t\rangle\|_{L^2} = \|t\|_{\ell_2^d}^2$ and it is L-subgaussian if for every $p \geqslant 2$ and every $t \in \ell_2^d$, $\|\langle X, t\rangle\|_{L^p} \leqslant L\sqrt{p}\|\langle X, t\rangle\|_{L^2}$.*

In other words, the covariance structure of an isotropic random variable coincides with the inner product in ℓ_2^d, and if X is an L-subgaussian random vector then the L^p norm of all linear forms does not grow faster than the L^p norm of the corresponding Gaussian variable. When dealing with the LASSO and SLOPE, the natural Euclidean structure is used in $\mathbb{R}^d$.

In the following assumption, we recall a setup where both estimators have been studied in [51].

Assumption 5.

1. *the data are i.i.d. (in particular, $|\mathcal{I}| = N$ and $|\mathcal{O}| = 0$, i.e. there is no outlier),*
2. *X is isotropic and L-subgaussian,*
3. *for $f^* = \langle t^*, \cdot\rangle$, $\zeta = Y - f^*(X) \in L^{q_0}$ for some $q_0 > 2$.*

Unlike many results on these estimators, Assumption 5 only requires a "minimal" L^{q_0} for $q_0 > 2$ moment on the noise. It appears that LASSO and SLOPE still achieve optimal rates of convergence under this weak stochastic assumption but the price to pay is a severely deteriorated probability estimate.

Theorem 3 (Theorem 1.4 in [51])**.** *Consider the LASSO under Assumption 5. Let $s \in [d]$. Assume that $N \geqslant c_1 s \log(ed/s)$ and that there is some $v \in \mathbb{R}^d$ supported on at most s coordinates for which $\|t^* - v\|_1 \leqslant c_2\|\xi\|_{L^{q_0}} s\sqrt{\log(ed)/N}$. The Lasso estimator $\hat{t}$ with regularization parameter $\lambda = c_3\|\xi\|_{L^{q_0}}\sqrt{\log(ed)/N}$ is such that with probability at least*

$$1 - \frac{c_4 \log^{q_0} N}{N^{q_0/2-1}} - 2\exp\left(-c_5 s \log(ed/s)\right) \tag{36}$$

for every $1 \leqslant p \leqslant 2$

$$\left\|\hat{t} - t^*\right\|_p \leqslant c_6\|\xi\|_{L_q} s^{1/p}\sqrt{\frac{\log(ed)}{N}}.$$

The constants $(c_j)_{j=1}^6$ depend only on L and q_0.

The error rate in Theorem 3 coincides with the standard estimate on the LASSO (cf. [14]), but in a broader context: t^* does not need to be sparse but should be approximated by a sparse vector; the target Y is arbitrary (there is no need for a statistical model) and the noise ζ may be heavy tailed and does not need to be independent from X. But there is no room for outliers, the design matrix X still needs to be subgaussian and the data are assumed to be i.i.d.. We will see below that the MOM version of the LASSO can go further, achieving minimax optimal error bounds with a much better probability estimate.

Turning to SLOPE, recall the following result for the regularization norm $\Psi(t) = \sum_{j=1}^d \beta_j t_j^\sharp$ when $\beta_j = C\sqrt{\log(ed/j)}$.

Theorem 4 (Theorem 1.6 in [51])**.** *Consider the SLOPE under Assumption 5. Assume that $N \geqslant c_1 s \log(ed/s)$ and that there is $v \in \mathbb{R}^d$ such that $|\mathrm{supp}(v)| \leqslant s$ and $\Psi(t^* - v) \leqslant c_2\|\xi\|_{L_q} s\log(ed/s)/\sqrt{N}$. The SLOPE estimator with regularization parameter $\lambda = c_3\|\xi\|_{L_q}/\sqrt{N}$, satisfies with the same probability as in (36) that*

$$\Psi(\hat{t} - t^*) \leqslant c_4\|\xi\|_{L_q}\frac{s}{\sqrt{N}}\log\left(\frac{ed}{s}\right) \quad \text{and} \quad \left\|\hat{t} - t^*\right\|_2^2 \leqslant c_5\|\xi\|_{L_q}^2\frac{s}{N}\log\left(\frac{ed}{s}\right).$$

The constants $(c_j)_{j=1}^5$ depend only on L and q_0.

B.4 Statistical analysis of MOM LASSO and MOM SLOPE

In this section, Theorem 2 is applied to the set F of linear functionals indexed by $\mathbb{R}^d$ with regularization functions being either the ℓ_1-norm or the SLOPE norm. The aim is to show that the results from Section B.3 are still satisfied (and sometimes even improved) by their MOM version under much weaker assumptions and with a much better probability deviation. Start with the new set of assumptions.

Assumption 6. *Denote by $(e_j)_{j=1}^d$ the canonical basis of $\mathbb{R}^d$. We assume that*

1. $|\mathcal{I}| \geqslant N/2$ *and* $|\mathcal{O}| \leqslant c_1 s \log(ed/s)$,
2. X *is isotropic and for every* $t \in \mathbb{R}^d$, $p \in [C_0 \log(ed)]$ *and* $j \in [d]$, $\left\|\left\langle X, e_j\right\rangle\right\|_{L^p} \leqslant L\sqrt{p}\left\|\left\langle X, e_j\right\rangle\right\|_{L^2}$,
3. *for* $f^* = \left\langle t^*, \cdot\right\rangle$, $\zeta = Y - f^*(X) \in L^{q_0}$ *for some* $q_0 > 2$.
4. *there exists* θ_0 *such that for all* $t \in \mathbb{R}^d$, $\left\|\left\langle X, t\right\rangle\right\|_{L^2} \leqslant \theta_0 \left\|\left\langle X, t\right\rangle\right\|_{L^1}$,
5. *there exists* θ_m *such that* $\operatorname{var}(\zeta\left\langle X, t\right\rangle) \leqslant \theta_m \left\|\left\langle X, t\right\rangle\right\|_{L^2}$.

In order to apply Theorem 2 we need to compute the fixed point functions $r_Q(\cdot)$, $r_M(\cdot)$ and solve the sparsity equation in both cases. We start with the fixed point functions. To that end we recall the definition of Gaussian mean widths: for a set $V \subset \mathbb{R}^d$, the Gaussian mean width of V is defined as

$$\ell^*(V) = \mathbb{E}\left\{\sup_{(v_j)\in V} \sum_{j=1}^d g_j v_j\right\}, \quad \text{where} \quad (g_1, \dots, g_d) \sim \mathcal{N}_d(0, I_d) \ . \tag{37}$$

The dual norm of the ℓ_1^d-norm is the ℓ_∞^d-norm which is 1-unconditional with respect to the canonical basis of $\mathbb{R}^d$ [70, Definition 1.4]. Therefore, [70, Theorem 1.6] applies under the following assumption.

Assumption 7. *There exist constants $q_0 > 2$, C_0 and L such that $\zeta \in L^{q_0}$, X is isotropic and for every $j \in [d]$ and $1 \leqslant p \leqslant C_0 \log d$, $\left\|\left\langle X, e_j\right\rangle\right\|_{L^p} \leqslant L\sqrt{p}\left\|\left\langle X, e_j\right\rangle\right\|_{L^2}$.*

Under Assumption 7, if $\sigma = \|\zeta\|_{L^{q_0}}$, [70, Theorem 1.6] shows that, for every $\rho > 0$,

$$\mathbb{E} \sup_{v \in \rho B_1^d \cap r B_2^d} \left|\sum_{i\in[N]} \epsilon_i \left\langle v, X_i\right\rangle\right| \leqslant c_2 \sqrt{N} \ell^*(\rho B_1^d \cap r B_2^d) \ ,$$

$$\mathbb{E} \sup_{v \in \rho B_1^d \cap r B_2^d} \left|\sum_{i\in[N]} \epsilon_i \zeta_i \left\langle v, X_i\right\rangle\right| \leqslant c_2 \sigma \sqrt{N} \ell^*(\rho B_1^d \cap r B_2^d) \ .$$

Local Gaussian mean widths $\ell^*(\rho B_1^d \cap r B_2^d)$ are bounded from above in [56, Lemma 5.3] and computations of $r_M(\cdot)$ and $r_Q(\cdot)$ follow

$$r_M^2(\rho) \lesssim_{L, q_0, \gamma_M} \begin{cases} \sigma^2 \frac{d}{N} & \text{if } \rho^2 N \geqslant \sigma^2 d^2 \\ \rho\sigma \sqrt{\frac{1}{N} \log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)} & \text{otherwise} \end{cases} ,$$

$$r_Q^2(\rho) \begin{cases} = 0 & \text{if } N \gtrsim_{L, \gamma_Q} d \\ \lesssim_{L, \gamma_Q} \frac{\rho^2}{N} \log\left(\frac{c(L, \gamma_Q) d}{N}\right) & \text{otherwise} \end{cases} .$$

Therefore, one can take

$$r^2(\rho) \sim_{L, q_0, \gamma_Q, \gamma_M} \begin{cases} \max\left(\rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\sigma^2 d}{N}\right) & \text{if } N \gtrsim_L d \\ \max\left(\rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\rho^2}{N}\log\left(\frac{d}{N}\right)\right) & \text{otherwise} \end{cases} . \tag{38}$$

Now we turn to a solution of the sparsity equation for the ℓ_1^d-norm. This equation has been solved in [56, Lemma 4.2], we recall this result.

Lemma 7. *If there exists $v \in \mathbb{R}^d$ such that $v \in t^* + (\rho/20)B_1^d$ and $|\mathrm{supp}(v)| \leqslant c\rho^2/r^2(\rho)$ then*

$$\Delta(\rho) = \inf_{h \in \rho S_1^{d-1} \cap r(\rho)B_2^d} \sup_{g \in \Gamma_{t^*}(\rho)} \left\langle h, g - t^* \right\rangle \geqslant \frac{4\rho}{5} \ .$$

where S_1^{d-1} is the unit sphere of the ℓ_1^d-norm and B_2^d is the unit Euclidean ball in $\mathbb{R}^d$.

As a consequence, if $N \gtrsim s\log(ed/s)$ and if there exists a s-sparse vector in $t^* + (\rho/20)B_1^d$, Lemma 7 and the choice of $r(\cdot)$ in (38) imply that for $\sigma = \|\zeta\|_{L^{q_0}}$,

$$\rho^* \sim_{L,q_0} \sigma s \sqrt{\frac{1}{N}\log\left(\frac{ed}{s}\right)} \text{ and } r^2(\rho^*) \sim \frac{\sigma^2 s}{N}\log\left(\frac{ed}{s}\right)$$

then ρ^* satisfies the sparsity equation and $r^2(\rho^*)$ is the rate of convergence of the LASSO for $\lambda \sim r^2(\rho^*)/\rho^* \sim \|\zeta\|_{L^{q_0}} \sqrt{\log(ed/s)/N}$. But, this choice of λ requires to know the sparsity parameter s which is usually not available. That is the reason why we either need to choose a larger value for the $r(\cdot)$ function as in [51] – this results in the suboptimal $\sqrt{\log(ed)/N}$ rates of convergence from Theorem 3 – or to use an adaptation step as section 3.4.1 – this results in the better minimax rate $\sqrt{\log(ed/s)/N}$ achieved by the MOM LASSO. To get the latter one needs a final ingredient which is the computation of the radii ρ_K and $\lambda \sim r^2(\rho_K)/\rho_K$. Let $K \in [N]$ and $\sigma = \|\zeta\|_{L^{q_0}}$. The equation $K = cr(\rho_K)^2 N$ is solved by

$$\rho_K \sim_{L,q_0} \frac{K}{\sigma}\sqrt{\frac{1}{N}\log^{-1}\left(\frac{\sigma^2 d}{K}\right)} \tag{39}$$

for the $r(\cdot)$ function defined in (38). Therefore,

$$\lambda \sim \frac{r^2(\rho_K)}{\rho_K} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho_K\sqrt{N}}\right)} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma^2 d}{K}\right)} \ . \tag{40}$$

The regularization parameter depends on the “level of noise” σ, the L^{q_0}-norm of ζ. This parameter is unknown in practice. Nevertheless, it can be estimated and replaced by this estimator in the regularization parameter as in [31, Sections 5.4 and 5.6.2].

The following result follows from Theorem 2 together with the computation of ρ^*, $r_Q(\cdot)$, $r_M(\cdot)$ and $r(\cdot)$ from the previous sections.

Theorem 5. *Grant Assumption 6. The MOM-LASSO estimator $\hat{t}$ satisfies, with probability at least $1 - c_1\exp(-c_2 s\log(ed/s))$, for every $1 \leqslant p \leqslant 2$,*

$$\left\|\hat{t} - t^*\right\|_p \leqslant c_3 \left\|\zeta\right\|_{L_{q_0}} s^{1/p}\sqrt{\frac{1}{N}\log\left(\frac{ed}{s}\right)},$$

where $(c_j)_{j=1}^3$ depends only on θ_0, θ_m and q_0.

In particular, Theorem 5 shows that, for our estimator contrary to the one in [60], the sparsity parameter s does not have to be known in advance in the LASSO case.

Proof. It follows from Theorem 2, the computation of $r(\rho_K)$ from (38) and ρ_K in (39) that with probability at least $1 - c_0\exp(-cr(\rho_K)^2 N/\overline{C})$, $\left\|\hat{t} - t^*\right\|_1 \leqslant \rho_{K^*}$ and $\left\|\hat{t} - t^*\right\|_2 \lesssim r(\rho_K)$. The result follows since $\rho_{K^*} \sim \rho^* \sim_{L,q_0} \sigma s\sqrt{\frac{1}{N}\log\left(\frac{ed}{s}\right)}$ and $\|v\|_p \leqslant \|v\|_1^{-1+2/p}\|v\|_2^{2-2/p}$ for all $v \in \mathbb{R}^d$ and $1 \leqslant p \leqslant 2$. ■

Theoretical properties of MOM LASSO (cf. Theorem 5) outperform those of LASSO (cf. Theorem 3) in several ways:

- Estimation rates achieved by MOM-LASSO are the actual minimax rates $s\log(ed/s)/N$, see [13], while classical LASSO estimators achieve the rate $s\log(ed)/N$. This improvement is possible thanks to the adaptation step in MOM-LASSO.
- the probability deviation in (36) is polynomial – $1/N^{(q_0/2-1)}$ – whereas it is exponentially small for MOM LASSO. Exponential rates for LASSO hold only if ζ is subgaussian ($\|\zeta\|_{L_p}\leqslant C\sqrt{p}\,\|\zeta\|_{L_2}$ for all $p\geqslant 2$).
- MOM LASSO is insensitive to data corruption by up to $s\log(ed/s)$ outliers while only one outlier can be responsible of a dramatic breakdown of the performance of LASSO (cf. Figure 1). Moreover, the informative data are only asked to have equivalent L^2 moments to the one of P for the MOM LASSO whereas the properties of the LASSO are only known in the i.i.d. setup.
- Assumptions on X are weaker for MOM LASSO than for LASSO. In the LASSO case, we assume that X is subgaussian whereas for the MOM LASSO we assumed that the coordinates of X have $C_0\log(ed)$ moments and that it satisfies a L^2/L^1 equivalence assumption.

Let us now turn to the SLOPE case. The computation of the fixed point functions $r_Q(\cdot)$ and $r_M(\cdot)$ rely on [70, Theorem 1.6] and the computation from [51]. Again, the SLOPE norm has a dual norm which is 1-unconditional with respect to the canonical basis of $\mathbb{R}^d$, [70, Definition 1.4]. Therefore, it follows from [70, Theorem 1.6] that under Assumption 7, one has

$$\mathbb{E}\sup_{v\in\rho\mathcal{B}\cap rB_2^d}\left|\sum_{i\in[N]}\epsilon_i\big\langle v,X_i\big\rangle\right|\leqslant c_2\sqrt{N}\ell^*(\rho\mathcal{B}\cap rB_2^d)\ ,$$

$$\mathbb{E}\sup_{v\in\rho\mathcal{B}\cap rB_2^d}\left|\sum_{i\in[N]}\epsilon_i\zeta_i\big\langle v,X_i\big\rangle\right|\leqslant c_2\sigma\sqrt{N}\ell^*(\rho\mathcal{B}\cap rB_2^d)\ ,$$

where $\mathcal{B}$ is the unit ball of the SLOPE norm. Local Gaussian mean widths $\ell^*(\rho\mathcal{B}\cap rB_2^d)$ are bounded from above in [56, Lemma 5.3]: $\ell^*(\rho\mathcal{B}\cap rB_2^d)\lesssim\min\{C\rho,\sqrt{d}r\}$ when $\beta_j=C\sqrt{\log(ed)/j}$ for all $j\in[d]$ and computations of $r_M(\cdot)$ and $r_Q(\cdot)$ follow:

$$r_Q^2(\rho)\lesssim_L\begin{cases}0 & \text{if } N\gtrsim_L d\\ \frac{\rho^2}{N} & \text{otherwise,}\end{cases}\quad\text{and}\quad r_M^2(\rho)\lesssim_{L,q,\delta}\begin{cases}\|\xi\|_{L_q}^2\frac{d}{N} & \text{if } \rho^2N\gtrsim_{L,q,\delta}\|\xi\|_{L_q}^2d^2\\ \|\xi\|_{L_q}\frac{\rho}{\sqrt{N}} & \text{otherwise.}\end{cases}$$

The sparsity equation relative to the SLOPE norm has been solved in Lemma 4.3 from [51].

Lemma 8. *Let $1\leqslant s\leqslant d$ and set $\mathcal{B}_s=\sum_{j\leqslant s}\beta_j/\sqrt{j}$. If t^* is $\rho/20$ approximated (relative to the SLOPE norm) by an s-sparse vector and if $40\mathcal{B}_s\leqslant\rho/r(\rho)$ then $\Delta(\rho)\geqslant 4\rho/5$.*

For $\beta_j\leqslant C\sqrt{\log(ed/j)}$, one may verify that $\mathcal{B}_s=\sum_{j\leqslant s}\beta_j/\sqrt{j}\lesssim C\sqrt{s\log(ed/s)}$. Hence, the condition $\mathcal{B}_s\lesssim\rho/r(\rho)$ holds when $N\gtrsim_{L,q_0}s\log(ed/s)$ and $\rho\gtrsim_{L,q_0}\|\xi\|_{L_q}\frac{s}{\sqrt{N}}\log\left(\frac{ed}{s}\right)$. Hence, it follows from Lemma 8 that $\Delta(\rho)\geqslant 4\rho/5$ when there is an s-sparse vector in $t^*+(\rho/20)B_\Psi$; therefore, one may apply Theorem 1 for the choice of the regularization parameter: $\lambda\sim r^2(\rho)/\rho\sim_{L,q,\delta}\|\xi\|_{L_q}/\sqrt{N}$.

Now, the final ingredient is to compute the ρ_K solution to $K=cr(\rho_K)^2N$. It is straightforward to check that $\rho_K\sim K/(\sigma\sqrt{N})$ and still $\lambda\sim r^2(\rho_K)/\rho_K\sim_{L,q,\delta}\|\xi\|_{L_q}/\sqrt{N}$.

The following result follows from Theorem 2 together with the computation of $\rho^*,\rho_K,r_Q(\cdot),r_M(\cdot)$ and $r(\cdot)$ above. Its proof is similar to the one of Theorem 5 and is therefore omitted.

Theorem 6. *Grant Assumption 6. The MOM-SLOPE estimator $\hat{t}$ satisfies, with probability at least $1 - c_1 \exp(-c_2 s \log(ed/s))$,*

$$\left\|\hat{t} - t^*\right\|_2^2 \leqslant c_3 \left\|\zeta\right\|_{L_{q_0}}^2 \frac{s}{N} \log\left(\frac{ed}{s}\right),$$

where $(c_j)_{j=1}^3$ depends only on θ_0, θ_m and q_0.

MOM-SLOPE has the same advantages upon SLOPE as MOM-LASSO upon LASSO. Those improvements are listed below Theorem 5 and will not be repeated. The only difference is that SLOPE, unlike LASSO, already achieves the minimax rate $s\log(ed/s)/N$.

C Learning without regularization

All the results from the previous sections also apply in the setup of learning with no regularization which is the framework one should consider when there is no a priori known structure on the oracle.

We consider the learning problem with no regularization. In this setup, we may use both minmaximization or maxminimization estimators

$$\widehat{f}_K \in \operatorname*{argmin}_{f\in F} \sup_{g\in F} T_K(g,f) \text{ and } \widehat{g}_K \in \operatorname*{argmax}_{g\in F} \inf_{f\in F} T_K(g,f) \tag{41}$$

where $T_K(g,f) = \mathrm{MOM}_K\big(\ell_f - \ell_g\big)$.

We show below that $\widehat{f}_K$ and $\widehat{g}_K$ are efficient procedures even in situations where the dataset is corrupted by outliers. The case $K = 1$ corresponds to the classical ERM: $\widehat{f}_1 = \widehat{g}_1 \in \operatorname{argmin}_{f\in F} P_N \ell_f$ which can only be trusted when used with a "clean dataset".

Indeed, the ideal setup for ERM is the subgaussian (and convex) framework: that is for a convex class F of functions, i.i.d. data $(X_i, Y_i)_{i=1}^N$ having the same distribution as (X, Y) and such that for some $L > 0$ and all $f, g \in F$,

$$\left\|Y\right\|_{\psi_2} < \infty \text{ and } \left\|g(X) - f(X)\right\|_{\psi_2} \leqslant L \left\|g(X) - f(X)\right\|_{L_2}. \tag{42}$$

When F satisfies the right-hand side of (42), we say that F is a L-subgaussian class. It is proved in [53] that in this setup the ERM is an optimal minimax procedure (cf. Theorem A′ from [53] recalled in Theorem 9 below).

But first, we need a version of the two theorems 1 and 2 valid for $\widehat{f}_K$ and $\widehat{g}_K$ (that is for the learning problem with no regularization). Let us first introduce the set of assumptions we use and for the sake of shortness we consider the simplification introduced in Remark 2. Then, we will introduce the two fixed points driving the statistical properties of $\widehat{f}_K$ and $\widehat{g}_K$.

Assumption 8. *For all $i \in \mathcal{I}$ and $f \in F$, $\left\|f(X_i) - f^*(X_i)\right\|_{L^2} = \left\|f(X) - f^*(X)\right\|_{L_2}$,*

$$\left\|Y_i - f(X_i)\right\|_{L^2} = \left\|Y - f(X)\right\|_{L_2}, \mathrm{var}((Y - f^*(X))(f(X) - f^*(X))) \leqslant \theta_m^2 \left\|f(X) - f^*(X)\right\|_{L^2}^2$$

and $\left\|f(X_i) - f^*(X_i)\right\|_{L^2} \leqslant \theta_0 \left\|f(X_i) - f^*(X_i)\right\|_{L^1}$.

The two fixed points associated to this problem are $r_Q(\rho, \gamma_Q)$ and $r_M(\rho, \gamma_M)$ as in Definition 5 for $\rho = \infty$:

$$r_Q(\gamma_Q) = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geqslant \frac{N}{2},\ \mathbb{E} \sup_{f\in F: \|f-f^*\|_{L_P^2} \leqslant r} \left|\sum_{i\in J} \epsilon_i (f - f^*)(X_i)\right| \leqslant \gamma_Q |J| r \right\},$$

$$r_M(\gamma_M) = \inf\left\{ r > 0 : \forall J \subset \mathcal{I}, |J| \geqslant \frac{N}{2},\ \mathbb{E} \sup_{f\in F: \|f-f^*\|_{L_P^2} \leqslant r} \left|\sum_{i\in J} \epsilon_i \zeta_i (f - f^*)(X_i)\right| \leqslant \gamma_M |J| r^2 \right\},$$

and let $r^* = r^*(\gamma_Q, \gamma_M) = \max\{r_Q(\gamma_Q), r_M(\gamma_M)\}$.

Theorem 7. *Grant Assumptions 8 and let $r_Q(\gamma_Q)$, $r_M(\gamma_M)$ and r^* be defined as above for $\gamma_Q = (384\theta_0)^{-1}$, $\gamma_M = \epsilon/192$ and $\epsilon = 1/(32\theta_0^2)$. Assume that $N \geqslant 384\theta_0^2$ and $|\mathcal{O}| \leqslant N/(768\theta_0^2)$. Let K^* denote the smallest integer such that $K^* \geqslant N\epsilon^2(r^*)^2/(384\theta_m^2)$. Then, for all $K \in [\max(K^*, 8|\mathcal{O}|), N/(96\theta_0^2)]$, with probability larger than $1 - 2\exp(-7K/9216)$, the estimators $\widehat{f}_K$ and $\widehat{g}_K$ defined in (41) satisfy*

$$\|\widehat{g}_K - f^*\|_{L_P^2}, \left\|\widehat{f}_K - f^*\right\|_{L_P^2} \leqslant \frac{\theta_m}{\epsilon}\sqrt{\frac{384K}{N}} \quad \textit{and} \quad R(\widehat{g}_K), R(\widehat{f}_K) \leqslant R(f^*) + (1+2\epsilon)\frac{384\theta_m^2 K}{\epsilon^2 N} \ .$$

Moreover, one can choose adaptively K via Lepski's method. We will do it only for the maxmin estimators $\widehat{g}_K$. Similar result hold for the minmax estimators $\widehat{f}_K$ from straightforward modifications (the same as in Section 3.4.1). Define the confidence regions: for all $J \in [K]$ and $g \in F$,

$$\hat{R}_J = \left\{ g \in F : \mathfrak{C}_J(g) \geqslant \frac{-384\theta_m^2 J}{\epsilon N} \right\} \text{ where } \mathfrak{C}_J(g) = \inf_{f \in F} T_J(g, f)$$

and $T_J(g, f) = \mathrm{MOM}_J\big(\ell_f - \ell_g\big)$ for all $f, g \in F$. Next, for all $J \in [\max(K^*, 8|\mathcal{O}|), N/(96\theta_0^2)]$, let

$$\hat{K} = \inf\left\{ K \in \left[\max(K^*, 8|\mathcal{O}|), \frac{N}{96\theta_0^2}\right] : \bigcap_{J=K}^{K_2} \hat{R}_J \neq \emptyset \right\} \text{ and } \widehat{g} \in \bigcap_{J=\hat{K}}^{K_2} \hat{R}_J \ .$$

The following theorem shows the performance of the resulting estimator.

Theorem 8. *Grant Assumption 8. For $\epsilon = 1/(32\theta_0^2)$ and all $K \in [\max(K^*, 8|\mathcal{O}|), N/(96\theta_0^2)]$, with probability larger than $1 - 2\exp(-K/2304)$,*

$$\|\widehat{g} - f^*\|_{L_P^2} \leqslant \frac{\theta_m}{\epsilon}\sqrt{\frac{384K}{N}}, \qquad R(\widehat{g}) \leqslant R(f^*) + (1+2\epsilon)\frac{384\theta_m^2 K}{\epsilon^2 N} \ .$$

The proofs of Theorem 7 and 8 essentially follow the one of Theorem 1 and 2. We will only sketch the proof for the maxmin estimator $\widehat{g}_K$ given that we already studied the minmax estimators in the regularized setup in Section 4.

Proof of Theorem 7. It follows from Lemma 1 and Lemma 2 for $\rho = \infty$ that there exists an event $\Omega(K)$ such that $\mathbb{P}(\Omega(K)) \geqslant 1 - 2\exp\left(-7K/9216\right)$ and, on $\Omega(K)$, for all $f \in F$,

1. if $\|f - f^*\|_{L_P^2} \geqslant r_Q(\gamma_Q)$ then

$$Q_{1/4,K}((f - f^*)^2) \geqslant \frac{1}{(4\theta_0)^2}\|f - f^*\|_{L_P^2}^2 \ , \tag{43}$$

2. there exists $3K/4$ block B_k with $k \in \mathcal{K}$, for which

$$|(P_{B_k} - \overline{P}_{B_k})[2\zeta(f - f^*)]| \leqslant \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2 \right) \ . \tag{44}$$

Moreover, it follows from Assumption 8 that for all $k \in \mathcal{K}$, $\overline{P}_{B_k}[\zeta(f - f^*)] = P[\zeta(f - f^*)]$ and $P[2\zeta(f - f^*)] \leqslant 0$ because of the convexity of F and the nearest point theorem. Therefore, on the event $\Omega(K)$, for all $f \in F$,

$$Q_{3/4,K}(2\zeta(f - f^*)) \leqslant \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2 \right) \tag{45}$$

and

$$\begin{aligned} P[-2\zeta(f - f^*)] &\leqslant P_{B_k}[-2\zeta(f - f^*)] + \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2 \right) \\ &\leqslant Q_{1/4,K}[(f - f^*)^2 - 2\zeta(f - f^*)] + \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2 \right) \\ &\leqslant T_K(f^*, f) + \epsilon \max\left(r_M^2(\gamma_M), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2 \right) \ . \end{aligned} \tag{46}$$

Let us place ourself on the event $\Omega(K)$ and let r_K be such that $r_K^2 = 384\theta_m^2 K/(\epsilon^2 N)$. Given that $r_K \geqslant r^*$, it follows from (43) and (45) that if $f \in F$ is such that $\|f - f^*\|_{L_P^2} \geqslant r_K$ then

$$T_K(f, f^*) \leqslant Q_{3/4,K}(2\zeta(f - f^*)) - Q_{1/4}((f - f^*)) \leqslant \left(\epsilon - \frac{1}{16\theta_0^2}\right) \|f - f^*\|_{L_P^2}^2 \leqslant \left(\frac{-1}{32\theta_0^2}\right) \|f - f^*\|_{L_P^2}^2 \quad (47)$$

for $\epsilon = 1/(32\theta_0^2)$ and if $\|f - f^*\|_{L_P^2} \leqslant r_K$ then $T_K(f, f^*) \leqslant Q_{3/4,K}(2\zeta(f - f^*)) \leqslant \epsilon r_K^2$. In particular,

$$\mathfrak{C}_K(f^*) = \inf_{f \in F} T_K(f^*, f) = -\sup_{f \in F} T_K(f, f^*) \geqslant -\epsilon r_K^2$$

and since $\mathfrak{C}_K(\hat{g}_K) \geqslant \mathfrak{C}_K(f^*)$ one has $\mathfrak{C}_K(\hat{g}_K) \geqslant -\epsilon r_K^2$. On the other hand, we have $\mathfrak{C}_K(\hat{g}_K) = \inf_{f \in F} T_K(\hat{g}_K, f) \leqslant T_K(\hat{g}_K, f^*)$. Therefore, $T_K(\hat{g}_K, f^*) \geqslant -\epsilon r_K^2$. But, we know from (47) that if $g \in F$ is such that $\|g - f^*\|_{L_P^2} > \sqrt{32\epsilon}\theta_0 r_K$ then $T_K(g, f^*) \leqslant (-1/(32\theta_0^2)) \|g - f^*\|_{L_P^2}^2 < -\epsilon r_K^2$. Therefore, one necessarily have $\|\hat{g}_K - f^*\|_{L_P^2} \leqslant \sqrt{32\epsilon}\theta_0 r_K = r_K$.

The oracle inequality now follows from (46):

$$R(\hat{g}_K) - R(f^*) = \|\hat{g}_K - f^*\|_{L_P^2}^2 + P[-2\zeta(\hat{g}_K - f^*)] \leqslant r_K^2 + T_K(f^*, \hat{g}_K) + \epsilon r_K^2 \leqslant (1 + 2\epsilon) r_K^2 \ .$$

■

Proof of Theorem 8. Consider the same notations as in the proof of Theorem 7 and denote $K_2 = N/(96\theta_0^2)$. It follows from the proof of Theorem 7, that with probability larger than $1 - 2\sum_{J=K}^{K_2} \exp(-7J/9216)$, for all $J \in [K, K_2]$, $\mathfrak{C}_J(f^*) \geqslant -\epsilon r_J^2$ therefore, $f^* \in \hat{R}_J$ and so $\hat{K} \leqslant K$. The latter implies that $\widehat{g} \in \hat{R}_K$which, by using the same argument as in the end of the proof of Theorem 7 implies that $\|\widehat{g} - f^*\|_{L_P^2} \leqslant r_K$ and then $R(\widehat{g}) - R(f^*) \leqslant (1 + 2\epsilon) r_K$. ■

Example: Ordinary least squares. Let us consider the case where $F = \{\langle \cdot, t \rangle : t \in \mathbb{R}^d\}$ is the set of all linear functionals indexed by $\mathbb{R}^d$. We assume that for all $i \in \mathcal{I}$ and $t \in \mathbb{R}^d$,

1. $\mathbb{E}\langle X_i, t \rangle^2 = \mathbb{E}\langle X, t \rangle^2$,
2. $\mathbb{E}(Y_i - \langle X_i, t \rangle)^2 = \mathbb{E}(Y - \langle X, t \rangle)^2$,
3. $\mathbb{E}(Y - \langle X, t^* \rangle)^2 \langle X, t \rangle^2 \leqslant \theta_m^2 \mathbb{E}\langle X, t \rangle^2$,
4. $\sqrt{\mathbb{E}\langle X, t \rangle^2} \leqslant \theta_0 \mathbb{E}|\langle X, t \rangle|$.

Let us now compute the fixed points $r_Q(\gamma_Q)$ and $r_M(\gamma_M)$. The proof essentially follows from Example 1 in [45]. Let $J \subset \mathcal{I}$ be such that $|J| \geqslant N/2$. Denote by $V \subset \mathbb{R}^d$ the smallest linear span containing almost surely X. Let $\varphi_1, \cdots, \varphi_D$ be an orthonormal basis of V with respect to the Hilbert norm $\|t\| = \mathbb{E}\langle X, t \rangle^2$. It follows from Cauchy-Schwartz inequality that

$$\mathbb{E} \sup_{f \in F : \|f - f^*\|_{L_P^2} \leqslant r} \left| \sum_{i \in J} \epsilon_i (f - f^*)(X_i) \right| = \mathbb{E} \sup_{\sum_{j=1}^D \theta_j^2 \leqslant r^2} \left| \sum_{j=1}^D \theta_j \sum_{i \in J} \epsilon_i \langle X_i, \varphi_j \rangle \right| \leqslant r \mathbb{E} \left(\sum_{j=1}^D \left(\sum_{i \in J} \epsilon_i \langle X_i, \varphi_j \rangle \right)^2 \right)^{1/2}$$

$$\leqslant r \sqrt{\sum_{j=1}^D \sum_{i \in J} \mathbb{E}\langle X_i, \varphi_j \rangle^2} = r\sqrt{D|J|}.$$

As a consequence, $r_Q(\gamma_Q) = 0$ if $\gamma_Q |J| \geqslant \sqrt{D|J|}$, i.e. if $\gamma_Q \geqslant \sqrt{D/|J|}$. Using the same arguments as above, we have

$$\mathbb{E} \sup_{f \in F : \|f - f^*\|_{L_P^2} \leqslant r} \left| \sum_{i \in J} \epsilon_i \zeta_i (f - f^*)(X_i) \right| \leqslant r \sqrt{\sum_{j=1}^D \sum_{i \in J} \mathbb{E}\zeta_i \langle X_i, \varphi_j \rangle^2} \leqslant r\theta_m \sqrt{D|J|}.$$

Therefore, $r_M(\gamma_M) \leqslant (\theta_m/\gamma_M)\sqrt{D/|J|} \leqslant (\theta_m/\gamma_M)\sqrt{2D/N}$ and $K^* = D$.

Now, it follows from Theorem 8, that if $N \geqslant 2(384\theta_0)^2 D$ and $|\mathcal{O}| \leqslant N/(768\theta_0^2)$ then the MOM OLS with adaptively chosen number of blocks K is such that for all $K \in \left[\max\left(D, 8|\mathcal{O}|\right), N/(96\theta_0^2)\right]$, with probability at least $1 - 2\exp(-K/2304)$,

$$\sqrt{\mathbb{E}\left\langle \hat{t} - t^*, X\right\rangle^2} \leqslant \frac{\theta_m}{\epsilon}\sqrt{\frac{384K}{N}}. \tag{48}$$

A consequence of (48), is that if the number of outliers is less than $D/8$ then the MOM OLS recovers the classical D/N rate of convergence for the means square error. This happens with probability at least $1-2\exp(-D/2304)$, that is with an exponentially large probability. This is a remarkable fact given that we only made assumptions on the L^2 moments of the design X. Moreover, this result is obtained under the only assumption on the informative data that they have equivalent L^2 moments to the one of the distribution of interest P. Therefore, only very little information on P needs to be brought to the statistician via the data; moreover those data can be corrupted up to $D/8$ complete outliers. Finally, note that we did not assume isotropicity of the design X to obtain (48). Therefore, (48) holds even for very degenerate design X and the price we pay is the true dimension of X that is of the dimension of the smallest linear span containing almost surely X not the one of the whole space $\mathbb{R}^d$.

D Minimax optimality of Theorem 1, 2, 7 and 8

The aim of this section is to show that the rates obtained in Theorems 1, 2, 7 and 8 are optimal in a minimax sense. To that end we recall a minimax lower bound result from [53].

Theorem 9 (Theorem A′ in [53]). *There exists an absolute constant c_0 for which the following holds. Let X be a random variable taking values in $\mathcal{X}$. Let F be a class of functions such that $\mathbb{E}f^2(X) < \infty$. Assume that F is star-shaped around one of its point (i.e. there exists $f_0 \in F$ such that for all $f \in F$ the segment $[f_0, f]$ belongs to F). Let ζ be a centered real-valued Gaussian variable with variance σ independent of X and for all $f^* \in F$ denote by Y^{f^*} the target variable*

$$Y^{f^*} = f^*(X) + \zeta. \tag{49}$$

Let $0 < \delta_N < 1$ and $r_N^2 > 0$. Let $\hat{f}_N$ be a statistics (i.e. a measurable function from $(\mathcal{X} \times \mathbb{R})^N$ to $L^2(P_X)$ where P_X is the probability distribution of X). Assume that $\hat{f}_N$ is such that for all $f^* \in F$, with probability at least $1 - \delta_N$,

$$\left\|\hat{f}_N(\mathcal{D}) - f^*\right\|_{L_P^2}^2 = R(\hat{f}_N(\mathcal{D})) - R(f^*) \leqslant r_N^2$$

where $\mathcal{D} = \{(X_i, Y_i) : i \in [N]\}$ is a set of N i.i.d. copies of (X, Y^{f^*}). Then, necessarily, one has

$$r_N^2 \geqslant \min\left(c_0\sigma^2\frac{\log(1/\delta_N)}{N}, \frac{1}{4}\mathrm{diam}(F, L^2(P_X))\right)$$

where $\mathrm{diam}(F, L^2(P_X))$ denotes the $L^2(P_X)$ diameter of F.

Theorem 9 proves that if the statistical model (49) holds then there is a strong connexion between the deviation parameter δ_N and the uniform rate of convergence r_N^2 over F : the smaller δ_N, the larger r_N^2. We now use this result to prove that Theorems 1, 2, 7 and 8 are essentially optimal.

In Theorems 7 and 8, the deviation bounds are $1 - c_1\exp(-c_2K)$ and the residual terms in the L_P^2 (to the square) estimation rates are like c_3K/N. Therefore, setting $\delta_N = c_1\exp(-c_2K)$ then Theorem 9 proves that no procedure can do better than

$$\min\left(c_0\sigma^2\frac{\log(1/\delta_N)}{N}, \frac{1}{4}\mathrm{diam}(F, L^2(P_X))\right) = \min\left(c_4\sigma^2\frac{K}{N}, \frac{1}{4}\mathrm{diam}(F, L^2(P_X))\right).$$

Given that one can obviously bound from above the performance of $\widehat{f}_K$ and $\widehat{g}_K$ as well as those of $\widehat{f}$ and $\widehat{g}$ in Theorems 7 and 8 by the L_P^2-diameter of F (because f^* and those estimators are in F), then the result of Theorem 7 and 8 are optimal even in the very strong Gaussian setup with i.i.d. data satisfying a Gaussian regression model like (49). The remarkable point is that Theorem 7 and 8 have been obtained under much weaker assumptions than those considered in Theorem 9 since outliers may corrupt the dataset, the noise and the design do not have to be independent, the informative data are only assumed to have a L^2 norm equivalent to the one of P and may therefore be heavy tailed.

Given the form of the deviation bounds in Theorems 1 and 2 and given that $r(\rho_K) \sim K/N$ and that $r(2\rho_K) \sim K/N$ (if one assumes a weak regularity assumption on the class F) then the same conclusions hold for Theorems 1 and 2: there is no procedure doing better than the MOM estimators even in the very good framework of Theorem 9.

References

[1] Notebook available at https://github.com/lecueguillaume/MOMpower.

[2] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1, part 2):137–147, 1999. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

[3] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[4] Sylvain Arlot and Matthieu Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *J. Mach. Learn. Res.*, 17:Paper No. 208, 50, 2016.

[5] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[6] Francis R. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 118–126, 2010.

[7] Krishnakumar Balasubramanian and Ming Yuan. Discussion of "Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation" [MR3466172]. *Electron. J. Stat.*, 10(1):71–73, 2016.

[8] Y. Baraud and L. Birgé. Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.*, 126(12):3888–3912, 2016.

[9] Y. Baraud and L. Birgé. Rho-estimators revisited: General theory and applications. *Preprint available in arXiv:1605.05051*, 2017.

[10] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017.

[11] Yannick Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401, 2011.

[12] Vic Barnett and Toby Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[13] Pierre Bellec, Guillaume Lecué, and Alexandre Tsybakov. Slope meets lasso: Improved oracle bounds and optimality. Technical report, CREST, CNRS, Université Paris Saclay, 2016.

[14] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[15] Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[16] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015.

[17] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[18] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 2015.

[19] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[20] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012.

[21] P. Laurie Davies and Ursula Gather. Breakdown and groups. *The Annals of Statistics*, 33(3):977–1035, 2005.

[22] Patrick L Davies. Aspects of robust linear regression. *The Annals of statistics*, pages 1843–1899, 1993.

[23] Víctor H. de la Peña and Evarist Giné. *Decoupling.* Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[24] Luc Devroye, Lászlό Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York).* Springer-Verlag, New York, 1996.

[25] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016.

[26] Yadolah Dodge. An introduction to statistical data analysis l1-norm based. *Statistical data analysis based on the L1-norm and related methods*, pages 1–22, 1987.

[27] David L Donoho. Breakdown properties of multivariate location estimators. *Ph. D. qualifying paper, Dept. Statistics, Harvard University, Boston.*, 1982.

[28] David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827, 1992.

[29] David L Donoho and Peter J Huber. The notion of breakdown point. *A festschrift for erich l. lehmann*, 157184, 1983.

[30] J. Fan, Q. Li, and Y. Wang. Estimation of high-dimensional mean regression in absence of symmetry and light-tail assumptions. *Journal of Royal Statistical Society B*, 79:247–265, 2017.

[31] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, FL, 2015.

[32] Frank R Hampel. Contribution to the theory of robust estimation. *Ph. D. Thesis, University of California, Berkeley*, 1968.

[33] Frank R Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

[34] Frank R Hampel. Robust estimation: A condensed partial survey. *Probability Theory and Related Fields*, 27(2):87–104, 1973.

[35] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[36] Frank R Hampel. Beyond location parameters: Robust concepts and methods. *Bulletin of the International statistical Institute*, 46(1):375–382, 1975.

[37] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. Linear models: robust estimation. *Robust Statistics: The Approach Based on Influence Functions*, pages 307–341, 1986.

[38] J. L. Hodges, Jr. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 163–186. Univ. California Press, Berkeley, Calif., 1967.

[39] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

[40] Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA, 1967.

[41] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[42] Peter J Huber and Elvezio M Ronchetti. Robust statistics. hoboken. *NJ: Wiley. doi*, 10(1002):9780470434697, 2009.

[43] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

[44] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics.* Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[45] Vladimir Koltchinskii. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6):1–50, December 2006. 2004 IMS Medallion Lecture.

[46] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008, 2015.

[47] Vladimir Koltchinskii and Shahar Mendelson. Bounding the Smallest Singular Value of a Random Matrix Without Concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008, 2015.

[48] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[49] Lucien Le Cam. *Asymptotic methods in statistical decision theory.* Springer Series in Statistics. Springer-Verlag, New York, 1986.

[50] G. Lecué and M. Lerasle. Learning from mom's principle : Le cam's approach. Technical report, CNRS, ENSAE, Paris-sud, 2017.

[51] G. Lecué and S. Mendelson. Regularization and the small-ball method i: sparse recovery. Technical report, CNRS, ENSAE, Technion, MSI ANU, 2016. To appear in the Annals of Statistics.

[52] G. Lecué and S. Mendelson. Regularization and the small-ball method ii: complexity dependent error rates. Technical report, CNRS, ENSAE, Technion, MSI ANU, 2017. To appear in Journal of machine learning research.

[53] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

[54] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. Technical report, CNRS, Ecole Polytechnique and Technion, 2014. To appear in Bernoulli journal.

[55] Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. Technical report, CNRS, Ecole Polytechnique and Technion, 2014. To appear in Journal of the European Mathematical Society.

[56] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. Technical report, CNRS, ENSAE and Technion, I.I.T., 2016.

[57] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[58] Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[59] Gabor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Preprint available on arXiv:1701.04112.*

[60] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Preprint available on ArXive:1608.00757.*

[61] Gabor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Preprint available on arXiv:1702.00482.*

[62] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust statistics. Theory and Methods.* John Wiley & Sons, Chichester. ISBN, 2006.

[63] Desire L Massart, Leonard Kaufman, Peter J Rousseeuw, and Annick Leroy. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, 187:171–179, 1986.

[64] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[65] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

[66] Shahar Mendelson. Learning without concentration for general loss function. Technical report, Technion, I.I.T., 2013. arXiv:1410.3192.

[67] Shahar Mendelson. Learning without concentration. In *Proceedings of the 27th annual conference on Learning Theory COLT14*, pages pp 25–39. 2014.

[68] Shahar Mendelson. A remark on the diameter of random sections of convex bodies. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 395–404. Springer, Cham, 2014.

[69] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

[70] Shahar Mendelson. On multiplier processes under weak moment assumptions. Technical report, Technion, 2016.

[71] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[72] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[73] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[74] Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41(6):2852–2876, 2013.

[75] Wen-Xin Zhou Qiang Sun and Jianqing Fan. Adaptive huber regression: Optimality and phase transition. *Preprint available in Arxive:1706.06991*, 2017.

[76] Peter Rousseeuw and Victor Yohai. Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984.

[77] Peter J Rousseeuw. Regression techniques with high breakdown point. *The Institute of Mathematical Statistics Bulletin*, 12:155, 1983.

[78] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

[79] Mark Rudelson and Roman Vershynin. Small ball probabilities for linear images of high dimensional distributions. Technical report, University of Michigan, 2014. International Mathematics Research Notices, to appear. [arXiv:1402.4492].

[80] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[81] Andrew F Siegel. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, 1982.

[82] Werner A Stahel. *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch., 1981.

[83] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[84] Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016.

[85] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[86] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[87] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[88] J.W. Tukey. Adress to international congress of mathematicians. Technical report, Vancouver, 1974.

[89] J.W. Tukey. T6: Order statistics. Technical report, In mimeographed notes for Statistics 411, Princeton Univ., 1974.

[90] Sara van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scand. J. Stat.*, 41(1):72–86, 2014.

[91] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

[92] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[93] Sara A. van de Geer. The deterministic lasso. Technical report, ETH Zürich, 2007. http://www.stat.math.ethz.ch/ geer/lasso.pdf.

[94] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[95] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

[96] Vladimir N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.

[97] M.-P. Victoria-Feser. Robust logistic regression for binomial responses. Technical report, University of Geneva - HEC, Available at SSRN: https://ssrn.com/abstract=1763301 or http://dx.doi.org/10.2139/ssrn.1763301, 2000.

[98] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014.

[99] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.