**Statistical Learning**
Article report, first semester 2021–2022

# Introduction to robust estimation

Article: *Robust machine learning by median-of-means : theory and practice (Lecué and Lerasle, 2017)*

*Author:*
Tomas Zouhir
Homer Durand

*Supervisor:*
Gerard Biau

January 17, 2022

# Contents

# 1 Introduction

As explained in Maronna et al. (2006), any statistical method is based on a number of assumptions, whether implicit or explicit. The most widely used framework is based on the idea that observations follow a Gaussian distribution. This hypothesis is essentially based on one of the fundamental theorem of probability, namely the limit central theorem, which states that the empirical mean of independent and identically distributed random variables tends towards a normal distribution. A large number of statistical tools and methods have been studied in this framework and are, thus, based on solid theoretical results but also have the advantage of being generally accessible from a computational point of view. But *"real world problems"* usually come with datasets that fall outside the classical statistical framework of independant and identically distributed (i.i.d) observations with Gaussian or sub-Gaussian behaviours. Indeed, it is common for data from concrete experiments to be corrupted by outliers or to exhibit heavy-tailed distributions undermining many of the methods used by data scientists such as the classical empirical mean or the Maximum Likelihood Estimate. This led to the creation of a new field of statistics in the 1960s with the work of John Tuckey (Tukey (1962)), Peter Huber (Huber (1964)) and Franck Hampel (Hampel (1973)) whose implementation was made possible by the concomitant increase in computing power, which made it possible to envisage heavier computing methods. These methods have become even more necessary with the advent of modern machine learning and the very large data sets over which data-scientists generally have little control.

Through this report, we seek to give an idea of the different issues of robust estimation, whether for the construction of estimators from corrupted datasets (see 2.1) or for the detection of outliers (see 3.3). In this purpose, we describe in a first time the broad outlines of this field of statistics (see 2.2), we then introduce the estimator proposed in Lecué and Lerasle (2017), describing the theoretical framework on which it is based (see 3.1) and its theoretical properties (see 3.4) and practical apects (see 3.2) .

# 2 Robust estimation

## 2.1 Data corruption

Real datasets often contain outliers, that is data that differ more or less greatly from other observations.
Outliers are often a problem in applications of statistics and machine learning because they may break classical estimators performance, whose theoretical guarantees are often derived from assumptions that does not take them into account (like the i.i.d assumption).
Part of those corrupted data are generally removed during the data cleaning step that is part of every data science project.
However, this process can get very complicated in the following cases :

1. High-dimensional statistic : the peculiarity of high-dimensional spaces makes it hard to figure out which data is problematic. Even nice gaussian vectors exhibit

odd properties in this setting (see chapter 1 of Giraud (2021))

2. Big Data : a dataset may simply be too big to perform a thorough examination of each data point.
3. Heavy-tailed distributions : outliers can be part of the data-generating process if the latter is a heavy-tailed distribution (this is common occurence in finance for instance).
4. Outliers detection : last but not least, our main goal might be to detect outliers, like in the case of fraud detection or terrorist activity surveillance. Removing them is then out of question.

Considering the aforementioned obstacles, we would like to build estimatiors able to resist (and if possible, to detect) outliers, that is estimators whose performance is as close as possible as the one we would get without outliers in our dataset. That is the whole point of the field of Robust Estimation.

## 2.2 A quick review of robust estimation

With the formalisation of the robust estimation problem, a variety of estimators have appeared. We focus here particularly on the location estimators of central tendency but many works have been proposed for the scale estimate (with the interquantile range or the median deviation) or the correlation estimate (with for example the Spearman rank correlation).

A very simple idea when trying to estimate the central tendency of a sample that could potentially have been corrupted by outliers is to remove extreme values, i.e. the $alpha \times n$ smallest samples and the $\alpha \times n$ largest samples. This is what is formalised by the $\alpha$-trimmed mean, for $\alpha \in [0, \frac{1}{2})$ :

$$\bar{T}_\alpha = \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} \hat{F}^{-1}(t) dt$$

with $\hat{F}$ being the empirical cumulative distribution function of the sample. On the other hand, Peter Huber proposes in Huber (1964) a generalization of the Maximum Likelihood Estimate, namely the M-estimators. Let us begin by briefly recalling the basic principle of the maximum likelihood estimator. Let $x_1, ..., x_n$ be a sample of observations such that

$$x_i = \mu + u_i \qquad (i = 1, ..., n)$$

where the errors $(u_i)_{i=1}^n$ are random variables with distribution function $F_0$. Then $x_1, ..., x_n$ are i.i.d with common distribution function $F(x) = F_0(x - \mu)$. Then the joint density of the observations is given by

$$L(x_1, ..., x_n; \mu) = \prod_{i=1}^n f_0(x_i - \mu)$$

where $f_0 = F_0'$. And the Maximum Likelihood Estimate $\mu$ is

$$\hat{\mu} = \text{argmax}_\mu L(x_1, ..., x_n; \mu) \qquad (1)$$

If we knew exactly $F_0$ the MLE would be optimal but as we stated earlier, it is generally not the case and for this reason we seek estimators that are nearly optimal in the normal case but also in cases deviating from the Gaussian framework. M-estimators proposed by peter Huber are then defined by

$$\hat{\mu} = \text{argmin}_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu)$$

where $\rho = -\log f_0$. Thus we see that $\rho(x) = x^2$ is optimal in the Gaussian case and it leads to the least squares estimator. The Huber function given by

$$\rho_k(x) = x^2 \mathbb{1}_{|x|<k} + (2k|x| - k^2)\mathbb{1}_{|x|\geq k} \tag{2}$$

brings to important M-estimates because it leads to the limit cases of the mean and the median when $K \to \infty$ and $K \to 0$.

### 2.2.1 Quantifying robustness with breakdown points

In order to quatify the robustness of an estimator $T$ to the corruption of a dataset $\mathscr{D}_{\mathscr{I}}$ by outliers $\mathscr{D}_{\mathscr{O}}$, the Machine Learning Community has introduced the notion of breakdown point. If by an adversarial choice of a corrupted dataset $\mathscr{D}_{\mathscr{O}}$ one can make $T(\mathscr{D}_{\mathscr{I}} \cup \mathscr{D}_{\mathscr{O}}) - T(\mathscr{D}_{\mathscr{I}})$ arbitrarly large, we say that the estimator $T$ breaks down. We call breakdown point the minimal proportion $\frac{|\mathscr{D}_{\mathscr{I}}|}{|\mathscr{D}_{\mathscr{O}}|+|\mathscr{D}_{\mathscr{O}}|}$ under which the estimator breaks down :

$$\epsilon^*(T, \mathscr{D}_{\mathscr{I}}) = \min_{m\in\mathbb{N}} \left\{ \frac{|\mathscr{D}_{\mathscr{O}}|}{|\mathscr{D}_{\mathscr{I}}| + |\mathscr{D}_{\mathscr{O}}|} : \sup_{\mathscr{D}_{\mathscr{O}}:|\mathscr{D}_{\mathscr{O}}|=m} |T(\mathscr{D}_{\mathscr{I}} \cup \mathscr{D}_{\mathscr{O}}) - T(\mathscr{D}_{\mathscr{I}})| = \infty \right\} \tag{3}$$

Thus, in the 1−dimensional case, the empirical mean has a breakdown point of $\frac{1}{|\mathscr{D}_{\mathscr{I}}|+1}$ because by adding a single outlier to the dataset one can make this estimator arbitrarily large. This is therefore the worst breakdown point value that an estimator can take. On the opposite, the empirical median has a breakdown point de 1/2 since it takes half of the observations to make this estimator arbitrarly large. We can see with these two trivial examples that the notion of breakdown point allows us to quantify the robustness of an estimator confirming the idea that the empirical mean is not very robust to outliers when the empirical median is the most robust estimator in a unidimensional framework.

Another point of view is proposed in Lecué and Lerasle (2017), focusing on the risk involved in a certain estimator. They define the breakdown number as the minimum number of outliers needed in a dataset to break the performance of an estimator.

**Definition 2.1.** *Let $\delta \in (0,1), \mathscr{R} > 0, N \geq 1, F$ be a class of functions from $\mathscr{X}$ to $\mathbb{R}$ and $\mathscr{P}$ be a set of distributions on $\mathscr{X} \times \mathbb{R}$. Let $T : \bigcup_{n\geq 1}(\mathscr{X} \times \mathbb{R})^n \mapsto F$ denote an estimator and let $\mathscr{D} = \{(X_i, Y_i)_{i=1}^N\}$ be a dataset made of $N$ i.i.d random variables with a common distribution in $\mathscr{P}$. For any $P \in \mathscr{P}$, let $f_P^* \in \text{argmin}_{f\in F} \mathbb{E}_{(X,Y)\sim P}[(Y - f(X))^2]$. The breakdown number of the estimator $T$ on the class $\mathscr{P}$ at rate $\mathscr{R}$ with confidence $\delta$ is*

$$K^*_{ML}(T,N,\mathscr{R},\delta,\mathscr{P}) = \min\{k\in\mathbb{Z}_+ : \inf_{P\in\mathscr{P}} \mathscr{P}_{\mathscr{D}\sim P^{\otimes n}}(\sup_{|\mathscr{O}|=k} \|T(\mathscr{D}\cup\mathscr{O}) - f^*_P\|_{L^2(P_X)\leq\mathscr{R}}) \geq 1-\delta\}$$

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

Minimax rates of cv as benchmark rates of cv. For any class $\mathscr{P}$ containing the Gaussian model for all $\mathscr{R} < \mathscr{R}(\delta, F)$ (minimax rate) it is clear that $K^* = 0$. On s'intéresse au ratio $\mathscr{R} \geq \mathscr{R}(\delta, F)$ et plus particuliérement au ratio de l'ordre de $\mathscr{R}(\delta, F)$ car estimateur avec $K^* > 0$ sont minimax (statistiquement optimaux) même si corrompu par $K^*$ outliers.

Also we may show the following relation between Relation breakdown point and breakdown number :

$$\epsilon^*(T,\mathscr{D}) \geq \frac{1+K^*}{1+K^*+N} \tag{4}$$

The breakdown number of the MOM estimator is given in theorem (5.1) in Appendix.

# 3 An example of a robust estimator : MOM estimator

## 3.1 Theoretical framework

Let $\mathscr{X}$ denote a measurable space and let $(X,Y),(X_i,Y_i)_{i\in[N]}$ denote random variables taking values in $\mathscr{X}\times\mathbb{R}$. 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 $f:\mathscr{X}\to\mathbb{R}$ and suppose that $F\subset L^2_P, \mathbb{E}\left[Y^2\right]<\infty$. For any $(x,y)\in\mathscr{X}\times\mathbb{R}$, let $\ell_f(x,y)=(y-f(x))^2$ denote the square loss function and let $f^*$ denote an oracle

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

For any $Q\in\left\{P,(P_i)_{i\in[N]}\right\}$ and any $p\geqslant 1$, let $\|f\|_{L^p_Q} = (Q|f|^p)^{1/p}$ the $L^p_Q$-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.

### 3.1.1 Interest

Informally, the MOM-estimator can be defined as followed : divide your data into blocks and estimate the expectation by the median of the empirical means respectively computed over each block.

The MOM-estimator that we will define in the next section have many upsides, the most prominent ones being that it is easy to compute and to understand, and that it has theoretical guarantees under minimal assumptions.

More precisely, we make basically no assumptions over the distribution or the dependance structure of the outliers. Furthermore, the assumptions we make over the distributions

of the informative data $P$ and $(P_i)_{i\in[N]}$ only involve that the latter have first and second moment, which make the MOM estimator particularly adapted to deal with heavy-tailed processes.
We will now expand on the two key ingredients of the MOM procedure:

1. turn the minimization problem into a minimaximization one, using the linearity of the expectation:
$$\arg\min_{f\in F} \mathbb{E}((Y-f(X))^2) = \arg\min_{f\in F} \sup_{g\in F} \mathbb{E}((Y-f(X))^2-(Y-g(X))^2)$$
2. estimate the unknown expectation by the empirical Median-of-Means (MOM) instead of the empirical mean.

Let's start by defining the latter.

### 3.1.2 Definition

Let $K$ denote an integer smaller than $N$ and let $B_1,\dots,B_K$ 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 $\mathscr{L}:\mathscr{X}\times\mathbb{R}\to\mathbb{R}$ and $k\in[K]$, let $P_{B_k}\mathscr{L}=|B_k|^{-1}\sum_{i\in B_k}\mathscr{L}(X_i,Y_i)$. Then $\mathrm{MOM}_K(\mathscr{L})$ is a median of the set of $K$ real numbers $\{P_{B_1}\mathscr{L},\cdots,P_{B_K}\mathscr{L}\}$.

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 $x_1,\dots,x_K$, we denote by

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

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

$$\begin{array}{rcl} Q_\alpha(x)\geqslant t & \text{iff} & \sup Q_\alpha(x)\geqslant t \\ Q_\alpha(x)\leqslant t & \text{iff} & \inf Q_\alpha(x)\leqslant t \\ z=Q_\alpha(x)+Q_\alpha(y) & \text{iff} & z\in Q_\alpha(x)+Q_\alpha(y) \end{array}$$

where in the last inequality we use the Minkowsky sum of two sets. More generally, inequalities involving are always understood in the worst possible case.

We can now give a formal definition of the MOM-estimator :

**Definition 3.1.** *Let $\alpha\in(0,1)$ and $K\in[N]$. For all functions $\mathscr{L}:\mathscr{X}\times\mathbb{R}\to\mathbb{R}$ the $\alpha$-quantile on $K$ blocks of $\mathscr{L}$ is $Q_{\alpha,K}(\mathscr{L})=Q_\alpha\left((P_{B_k}\mathscr{L})_{k\in[K]}\right)$. In particular, the Median-of-Means (MOM) of $\mathscr{L}$ on $K$ blocks is defined as $\mathrm{MOM}_K(\mathscr{L})=Q_{1/2,K}(\mathscr{L})$. For all $f,g\in F$, the MOM test on $K$ blocks of $g$ against $f$ is defined by*

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

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

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

### 3.1.3 From minimization to minimaximization:

One may want plug the estimator directly into the minimization problem instead of the empirical mean. However, doing so lead to suboptimal minimax rate, mostly because the non-linearity of the median that prevent to use similar arguments used for the empirical mean.

Though the minimization and minimaximizations problems are equivalent when the unknown expectation is involved, it is easy to see that it is no longer the case when the latter is replaced by the MOM estimator, due (again) to the non-linearity of the median. We will not expand on this point and invite the reader to dive into a more recent paper of the authors, Lecué et al. (2018), where they study the MOM-estimator for classification problems without minimaximization transformation.

## 3.2 Estimateur MOM en pratique

### 3.2.1 Algo ADMM/LASSO

In this section, we show an implementation of the MOM estimator in the particular case of the linear model with LASSO regularization, called MOM LASSO.

We will see that algorithms used to solve the minimization problem to compute the classical LASSO estimator can easily be adapated to the min-max problem. The MOM LASSO minimaximization problem is formulated as follow :

$$\hat{t}_{K,\lambda} \in \operatorname*{argmin}_{t\in\mathbb{R}^d} \sup_{t'\in\mathbb{R}^d} T_{K,\lambda}\left(t',t\right)$$

where $T_{K,\lambda}\left(t',t\right) = \text{MOM}_K\left(\ell_t - \ell_{t'}\right) + \lambda\left(\|t\|_1 - \|t'\|_1\right), \text{MOM}_K\left(\ell_t - \ell_{t'}\right)$ is a median of the set of real numbers $\left\{P_{B_1}\left(\ell_t - \ell_{t'}\right), \cdots, P_{B_K}\left(\ell_t - \ell_{t'}\right)\right\}$ and for all $k \in [K]$,

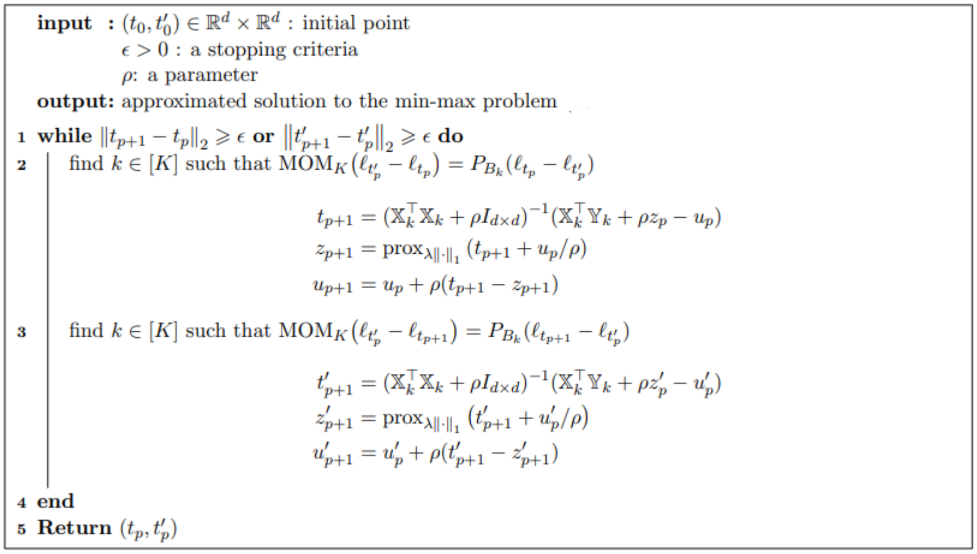
$$P_{B_k}\left(\ell_t - \ell_{t'}\right) = \frac{1}{|B_k|}\sum_{i\in B_k}\left(\left(Y_i - \langle X_i, t\rangle\right)^2 - \left(Y_i - \langle X_i, t'\rangle\right)\right)^2$$

One (among many others) algorithms used to compute the LASSO estimator is the ADMM (Alternating Direction Method of Multipliers) which belong to the class of the Douglas-Ratchford convex optimization methods. In figure 1 is presented the MOM version of the ADMM algorithm. The idea is to turn the original algorithm based on a sequence of sequence of descents into one based on a sequence of alternating descents (in $t$) and ascents (in $t'$).

As we can see in the figure (3.2.1), the addition of a single outlier can significantly degrade the performance of the classical LASSO algorithm. The MOM ADMM algorithm seems to be much more robust to different types of data corruption.

### 3.2.2 Adaptative choice of hyper-parameters

When using the MOM estimator, it is necessary to choose values for the hyper-parameters, namely the number of blocks $K$ and the regularisation parameter $\lambda$. The method that is classically used in Machine Learning to approximate the optimal values of the hyper-parameters is the $V$-fold Cross Validation but, in the context of robust estimation, it does



```
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
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 $\mathrm{MOM}_K(\ell_{t'_p} - \ell_{t_p}) = P_{B_k}(\ell_{t_p} - \ell_{t'_p})$
        $t_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p - u_p)$
        $z_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}(t_{p+1} + u_p/\rho)$
        $u_{p+1} = u_p + \rho(t_{p+1} - z_{p+1})$
    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 $(t_p, t'_p)$
```

Figure 1: An ADMM algorithm for the minimaximization MOM estimator

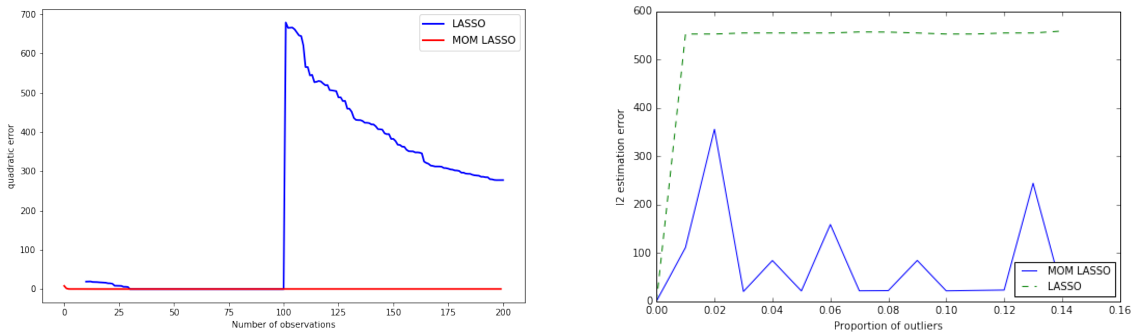


Figure 2: Performance of ADMM and MOM ADMM in term of $l_2$ error. Adding one outlier at index 100 (left), evolution of the $l_2$ error regarding the proportion of outlier (right)

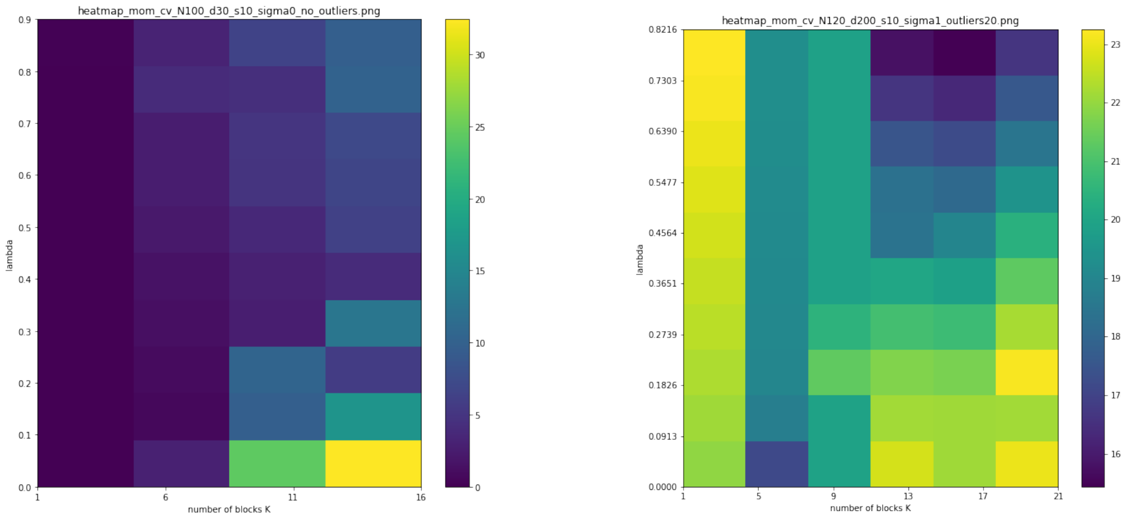


Figure 3: Adaptative choice of $K$ and $\lambda$ with cross validation on sparse data (left). Adaptative choice of $K$ and $\lambda$ with cross validation on sparse corrupted data (right).

not seem to be optimal because the test sets might have been corrupted by outliers. The authors of Lecué and Lerasle (2017) therefore propose a more robust procedure, adapted to the MOM estimator which shows good empirical results. The procedure simply replaces the empirical mean estimator classically used in the classical Cross Validation with the MOM estimator and also replaces the mean over the $V$ partitions with a median. They therefore propose the following hyper-parameter selection procedure: Deux hyperparamètres $\hat{K}$ et $\lambda$

**Definition 3.2** (Median of Mean $V$-fold Cross Validation)**.** *Let* $(\hat{f}^{(v)}_{K,\lambda} : K \in \mathscr{G}_K, \lambda \in \mathscr{G}_\lambda)$ *be a family of estimators with* $\mathscr{G}_K \subset [N]$ *and* $\mathscr{G}_\lambda \subset (0,1]$*. The Median of Mean* $V$*-fold Cross Validation procedure associated to this family of estimators is given* $\hat{f}^{(v)}_{\hat{K},\hat{\lambda}}$ *where* $(\hat{K},\hat{\lambda})$ *is minimizing the* $MomCv_V$ *criteria*

$$(K,\lambda) \in \mathscr{G}_K \times \mathscr{G}_\lambda \to MomCv_V(K,\lambda) = Q_{1/2}(MOM^{(v)}_{K'}(l_{\hat{f}^{(v)}_{K,\lambda}})_{vin[V]})$$

*where* $\forall v \in [V], f \in F$*,*

$$MOM^{(v)}_{K'}(l_f) = MOM_{K'}(P_{B^{(v)}_1} l_f, ..., P_{B^{(v)}_{K'}} l_f)$$

*and* $B^{(v)}_1 \cup ... \cup B^{(v)}_{K'}$ *is a partition of the test set* $\mathscr{D}_v$ *into* $K'$ *blocks where* $K' \in [N/V]$ *such that* $K'$ *divides* $N/V$*.*

We see that the adaptive choice of $K$ and $\lambda$ seems relevant in the case of uncorrupted data since it chooses $\hat{K} = 1$, i.e. that it does not separate the data and thus that it computes a traditional LASSO estimation. It also seems to fit well in the case of corrupted data and we see that the best performances in this case are with $\lambda = 0.822$ and $K = 17$.

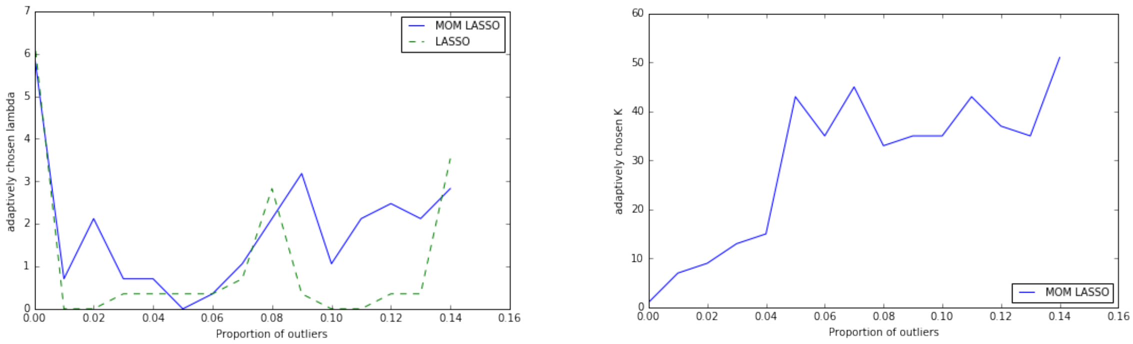


Figure 4: Adaptative choice of $K$ and $\lambda$ with cross validation on corrupted data. Adaptatively chosen $\lambda$ (center) for MOM LASSO and LASSO estimators. Adaptatively chosen $K$ for MOM LASSO estimator (right).

We can observe on figure [3.2.2](left) that the MOM LASSO estimator performs much better in terms of squared error than the LASSO estimator as soon as the data are corrupted by outliers. We also notice that the adaptive $\hat{K}$ selected by Cross Validation increases with the number of outliers in the data set, which is consistent with the fact that $K$ must be at least twice as large as the number of outliers.

## 3.3 Outliers detection

We can improve again the stability and performance of the algorithm by simply shuffling the $K$ blocks at each step, like in the algotithm in figure 5.
In figure 6 we compare the convergence of the ADMM MOM LASSO algorithm with fixed blocks versus random blocks. We can see that not only shuffling the blocks at each step makes our algorithm more stable, but also greatly improves the estimation error.

We can even derive an outlier detection procedure from this. Indeed, provided that the number of blocks is large enough, we expect that the outliers would not be in the median block, given that outlier data tend to yield extreme values of the empirical mean. So, what we can do by shuffling the blocks is to give a score to the data selected in the median block. More precisely every data start with a score of 0, and at each step the data selected in the median block are given one point. At the final step, we compare the score of each data : outliers should have a much lower score than informative data. Figure 7 shows an example of this procedure on synthetic data.

```
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
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 $(t_p, t'_p)$
```

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

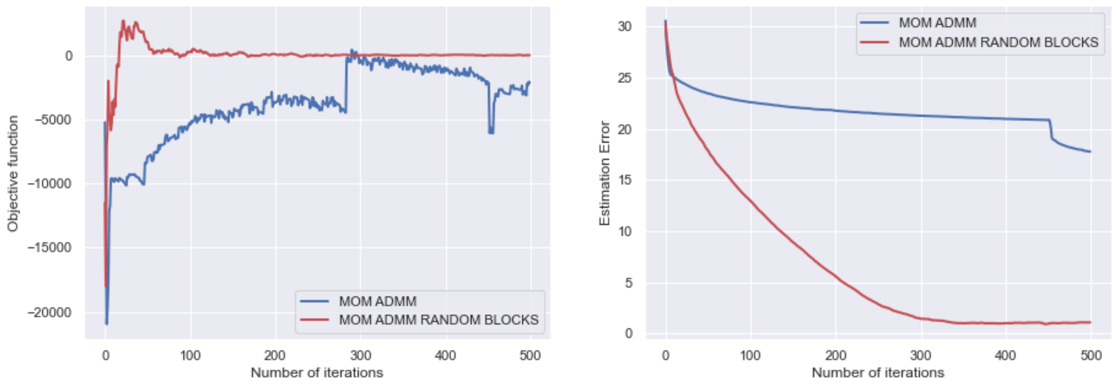


Figure 6: Fixed blocks against random blocks.

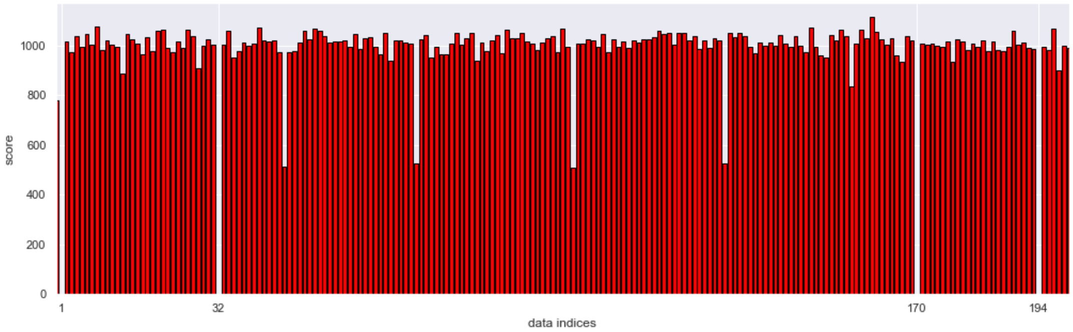


Figure 7: Outliers detection algorithm. The dataset has been corrupted by 4 outliers at number 1, 32, 170 and 194. The score of the outliers is 0: they haven't been selected even once.

## 3.4 Theoretical properties

In this section we study the theorectical performance of the MOM estimator. The main result of this section is given by theorem (3.4) which gives rates of convergence of the estimator wich are optimal in the minimax sense for the regularization norm and the $L_P^2$ norm. The minimax optimality will not be proven in this work but evidence can be found in Lecué and Lerasle (2017).

### 3.4.1 Assumptions

We only need the following simple assumptions on informative data in order to prove (3.4).

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

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

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

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

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

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 $\zeta$ is independent of $X$ and $\zeta$ has a finite $L^2$-moment bounded by $\theta_m$. In non-i.i.d. setups, assumption 2 also holds if for all $i \in I$, $\|\zeta\|_{L_{P_i}^4} \leqslant \theta_2 < \infty-$ where

$\zeta(x,y) = y - f^*(x)$ for all $x \in \mathscr{X}$ and $y \in \mathbb{R}-$ and, for every $f \in F$, $\|f - f^*\|_{L^4_{P_i}} \leqslant \theta_1 \|f - f^*\|_{L^2_P}$, because, in that case,

$$\sqrt{\operatorname{var}_{P_i}\left(\zeta\left(f - f^*\right)\right)} \leqslant \left\|\zeta\left(f - f^*\right)\right\|_{L^2_{P_i}} \leqslant \|\zeta\|_{L^4_{P_i}} \|f - f^*\|_{L^4_{P_i}} \leqslant \theta_1\theta_2 \|f - f^*\|_{L^2_P}$$

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 $i \in I$*

$$\|f - f^*\|_{L^2_P} \leqslant \theta_0 \|f - f^*\|_{L^1_{P_i}} .$$

Note that $\|f - f^*\|_{L^1_{P_i}} \leqslant \|f - f^*\|_{L^2_{P_i}}$ for all $f \in F$ and $i \in I$. Therefore, 1 and 3 are together equivalent to assume that all the norms $L^2_P, L^2_{P_i}, L^1_{P_i}, i \in I$ are equivalent over $F - f^*$.

Before stating the main theorem, we still need to introduce some definitions.

### 3.4.2 Rademacher complexities

We define the balls associated with the regularization norm $\|\cdot\|$ and the $L^2_P$ norm. For all $\rho \geqslant 0$,

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

where $B = \{f \in \operatorname{span}(F), \|f\| \leqslant \rho\}$ and for $r \geqslant 0$,

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

We now introduce the Rademacher complexities of the sets $B\left(f^*, \rho\right) \cap B_2\left(f^*, r\right)$ :

**Definition 3.3.** *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 $\rho$ centered at $f$ by*

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

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

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

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

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

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

It follows from Lemma 2.3 in Lecué and Mendelson (2016) 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 Lecué and Mendelson (2016), 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 $B\left(f^*, \rho\right)$. Note also that $r_Q$ and $r_M$ are well defined when $|I| \geqslant N/2$, which implies that at least half data are informative.

**Theorem 3.4.** *Grant Assumptions 1, 2 and 3 and let $r_Q, r_M$ denote the functions introduced in Definition 5. Assume that $N \geqslant 384\left(\theta_0\theta_{r0}\right)^2$ and $|\mathscr{O}| \leqslant N/\left(768\theta_0^2\right)$. Let $\rho^*$ be solution to the sparsity equation from Definition 6. Let $K^*$ denote the smallest integer such that*

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

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

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

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

$$\left|P_i\zeta\left(f - f^*\right) - P\zeta\left(f - f^*\right)\right| \leqslant \epsilon \max\left(r_M^2\left(\rho, \gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_P^2}^2\right). \quad (5)$$

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

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

### 3.4.3 Quadratic term and multiplier decomposition

In order to control the risk of our estimator, we bound from above $T_{K,\lambda}(f, f^*)$ for all functions $f$ far from $f^*$. For this purpose we recall the quadratic/multiplier decomposition of the difference of the quadratic losses.

$$\begin{aligned} l_f(x,y) - l_g(x,y) &= (y - f(x))^2 - (y - g(x))^2 \\ &= f^2(x) + g^2(x) - 2f(x)g(x) - 2(yf(x) - yg(x)) + 2f(x)g(x) - 2g^2(x) \\ &= (f(x) - g(x))^2 + 2(y - g(x))(g(x) - f(x)) \end{aligned}$$

So we get

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

where $2\zeta(f - f^*)$ is the multiplier term and $(f - f^*)^2$ is the quadratic term. The following two flemmas control the quantiles of the means of those two terms.

**Lemma 3.5** (Granted)**.** *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 $K \in \left[|\mathscr{O}|/(1-\gamma), N\alpha/\left(2\theta_0\theta_{r0}\right)^2\right]$. 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 $\|f-f^*\|_{L_P^2} \geqslant r_Q\left(\rho, \gamma_Q\right)$ then*

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

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

**Lemma 3.6** (Granted)**.** *Grant Assumption 2. Fix $\eta \in (0,1), \rho \in (0,+\infty]$, and let $\alpha, \gamma_M, \gamma, x$ and $\epsilon$ be positive absolute constants such that $\gamma\left(1-\alpha-x-8\gamma_M/\epsilon\right) \geqslant 1-\eta$. Let $K \in [|\mathscr{O}|/(1-\gamma), N]$. There exists an event $\Omega_M(K)$ such that $\mathbb{P}\left(\Omega_M(K)\right) \geqslant 1-\exp\left(-\gamma K x^2/2\right)$ and on the event $\Omega_M(K)$ : if $f \in F$ is such that $\|f-f^*\| \leqslant \rho$ then*

$$\left|\left\{k \in \mathscr{K} : \left|2\left(P_{B_k}-\bar{P}_{B_k}\right)\left(\zeta\left(f-f^*\right)\right)\right| \leqslant \epsilon \max\left(C_K, r_M^2, \left\|f-f^*\right\|_{L_P^2}^2\right)\right\}\right| \geqslant (1-\eta)K,$$

*with $C_K = \frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}$, $\bar{P}_{B_k}\left(\zeta\left(f-f^*\right)\right) := \left|B_k\right|^{-1}\sum_{i \in B_k}\mathbb{E}\left(\zeta_i\left(f(X_i)-f^*(X_i)\right)\right)$*

### 3.4.4 The sparsity equation:

For $\|f-f^*\|_{L_P^2}$ small, the quadratic term $(f-f^*)^2$ will not help to bound from above $T_{K,\lambda}(f,f^*)$ and we then only rely on the regularization term. For this we bound from below $(\|f^*\|-\|f\|)$ using the *saprsity equation*. We need in this purpose to introduce the subdifferiantials of the $L_2$ norm :

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

where $E^*$ is the dual space of $E$ with norm $\|.\|^*$.

**Definition 3.7.** *Let us introduce, for all $\rho > 0$,*

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

*with $H_\rho$, the set of function close to $f^*$ in the $L_P^2$ sens and with distance $\rho$ in the $L^2$ sense, define as*

$$H_\rho = \{f \in F : \|f-f^*\| = \rho \textit{ and } \|f-f^*\|_{L_P^2} \leqslant r(\rho)\Big\}$$

*and $\Gamma_{f^*}(\rho)$ as*

$$\Gamma_{f^*}(\rho) = \bigcup_{f \in F : \|f-f^*\| \leqslant \rho/20} (\partial\|\cdot\|)_f.$$

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

The *sparsity equation* quantifies the largeness of $\|f\|-\|f^{**}\|$ for any $f^{**}\in\{f\in F:\|f^*-f\|\leqslant\rho/20\}$ and $f\in H_\rho$ such that $\|f\|-\|f^{**}\|\geqslant\|f\|-\|f^*\|-\|f^*-f^{**}\|$ is large as well. Then we can bound $\|g\|-\|f^*\|$ for $g\in F$ with the lemma :

**Lemma 3.8.** *Let* $\rho\geqslant 0,\Gamma_{f^*}(\rho)=\cup_{f\in F:\|f-f^*\|\leqslant(\rho/20)}(\partial\|\cdot\|)_f$. *For all* $g\in F$,

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

*Proof.* For $z^*\in(\partial\|.\|)_{f^{**}}$ we have

$$\begin{aligned}\|g\|-\|f^*\|&\geqslant\|g\|-\|f^{**}\|-\|f^{**}-f^*\|\\&\geqslant z^*(g-f^{**})-\frac{\rho}{20}=z^*(g-f^*)-z^*(f^{**}-f^*)-\frac{\rho}{20}\\&\geqslant z^*(g-f^*)-\frac{\rho}{20}\qquad\text{because }z^*(f^{**}-f^*)\leqslant\|f^{**}-f^*\|\end{aligned}$$

□

### 3.4.5 Bounding the empirical criterion

In order to up bound the term $C_{K,\lambda}(f^*)$ we consider a partition of the space $F$ according to the distance between $g$ and $f^*$ in term of $L_2$ and $L_2^P$ norms. We define for $\kappa\geq 1$ :

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

The following flemma gives upper bounds for $C_{K,\lambda}(f^*)$ for each of these partitions.

**Lemma 3.9.** *On the event* $\Omega(K)$*, it holds for all* $\kappa\in\{1,2\}$,

$$\sup_{g\in F_1^{(\kappa)}}T_{K,\lambda}\left(g,f^*\right)\leqslant\left(2+c'\kappa\right)\epsilon r^2\left(\kappa\rho_K\right)\tag{6}$$

$$\sup_{g\in F_2^{(\kappa)}}T_{K,\lambda}\left(g,f^*\right)\leqslant\left(\left(2+c'\kappa\right)\epsilon-\frac{1}{16\theta_0^2}\right)r^2\left(\kappa\rho_K\right)\tag{7}$$

$$\sup_{g\in F_3^{(\kappa)}}T_{K,\lambda}\left(g,f^*\right)\leqslant\kappa\max\left(2\epsilon-\frac{1}{16\theta_0^2}+\frac{11c'\epsilon}{10},2\epsilon-\frac{7c'\epsilon}{10}\right)r^2\left(\rho_K\right)\tag{8}$$

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

*Proof.* First by Lemma (3.6) we get that, there exist $(3/4)K$ block $B_k$ with $k\in\mathscr{K}$, for which,

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

and by assumption (5) from Theorem (3.4), we get for those blocks that for all $f \in F$ such that $\|f - f^*\| \leq \rho$,

$$P_{B_k}\left[2\zeta\left(f - f^*\right)\right] |\leqslant P\left[2\zeta\left(f - f^*\right)\right] + 2\epsilon \max\left(r_M^2\left(\rho, \gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_p^2}^2\right) \tag{9}$$

As $f^*$ is a minimizer of $P(l_f - l_g)$ over $F$ which is convex, it follows from the *nearest point theorem* [1] that $P\left[2\zeta\left(f - f^*\right)\right] \leq 0$ for all $f \in F$. So we get that for all $f \in F$ such that $\|f - f^*\| \leq \rho$,

$$Q_{3/4,K}\left(2\zeta\left(f - f^*\right)\right) \leqslant 2\epsilon \max\left(r_M^2\left(\rho, \gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|{-f^*}\|_{L_p^2}^2\right) \tag{10}$$

Using same arguments as in equation (9) we get for all $f \in F$ such that $\|f - f^*\| \leq \rho$ that,

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

Finally by using that $\|f - f^*\| \leq \rho$ we bound from above $P_{B_k}\left[-2\zeta\left(f - f^*\right)\right]$ by $Q_{1/4,K}\left[\left(f - f^*\right)^2 - 2\zeta\left(f - f^*\right)\right] + \lambda\left(\|f\| - \|f^*\|\right) + \lambda\rho$ and we have,

$$P\left[-2\zeta\left(f - f^*\right)\right] \leqslant T_{K,\lambda}\left(f^*, f\right) + 2\epsilon \max\left(r_M^2\left(\rho, \gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|f - f^*\|_{L_p^2}^2\right) + \lambda\rho \tag{11}$$

We finally decompose the proof for each partition of F.

1. **Bound over $F_1^{(\kappa)}$.** We have by definition of $T_{K,\lambda}$

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

Using equation (10), we immediatly get, for all $g \in F_1^{(\kappa)}$,

$$\begin{aligned} T_{K,\lambda}\left(g, f^*\right) &\leqslant 2\epsilon \max\left(r_M^2\left(\rho, \gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|g - f^*\|_{L_p^2}^2\right) + \lambda\left\|f^* - g\right\| \\ &\leqslant 2\epsilon \max\left(r_M^2\left(\rho, \gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \|g - f^*\|_{L_p^2}^2\right) + \lambda\kappa\rho_K \quad \text{by definition of } F_1^{(\kappa)} \end{aligned}$$

Choosing the radius $\rho_K$ and the regularization parameter as in Theorem (3.4) lead immediately to (6).

[1] Let $S$ be a non-empty closed convex set in $\mathbb{R}^n$ and let $y \notin S$, then $\exists$ a point $\bar{x} \in S$ with minimum distance from y, i.e., $\|y - \bar{x}\| \leq \|y - x\| \forall x \in S$.
Furthermore, $\bar{x}$ is a minimizing point if and only if $(y - \hat{x})^T(x - \hat{x}) \leq 0$

2. **Bound over $F_2^{(\kappa)}$.** Using the fact that $Q_{1/2}(x-y) \leqslant Q_{3/4}(x) - Q_{1/4}(y)$, we have

$$\begin{aligned} T_{K,\lambda}(g,f^*) &\leqslant Q_{3/4,K}\left(2\zeta\left(g-f^*\right)\right) - Q_{1/4,K}\left(\left(g-f^*\right)^2\right) + \lambda\left\|f^*-g\right\| \\ &\leqslant Q_{3/4,K}\left(2\zeta\left(g-f^*\right)\right) - Q_{1/4,K}\left(\left(g-f^*\right)^2\right) + \lambda\kappa\rho_K \end{aligned}$$

We then bound $Q_{1/4,K}\left(\left(g-f^*\right)^2\right)$ using lemma (3.5) and $Q_{3/4,K}\left(2\zeta\left(g-f^*\right)\right)$ using (10) and get

$$\begin{aligned} T_{K,\lambda}(g,f^*) &\leqslant 2\epsilon\max\left(r_M^2\left(\rho,\gamma_M\right), \frac{384\theta_m^2}{\epsilon^2}\frac{K}{N}, \left\|g-f^*\right\|_{L_p^2}^2\right) - \left(4\theta_0\right)^{-2}\left\|g-f^*\right\|_{L_P^2}^2 + \lambda\kappa\rho_K \\ &\leqslant \left(2\epsilon - \frac{1}{\left(4\theta_0\right)^2}\right)\left\|g-f^*\right\|_{L_P^2}^2 + \lambda\kappa\rho_K \qquad \text{using that } 2\epsilon \leq \left(4\theta_0\right)^{-2} \\ &\leqslant \left(2\epsilon - \frac{1}{16\theta_0^2}\right) r^2\left(\kappa\rho_K\right) + \lambda\kappa\rho_K \end{aligned}$$

Choosing the radius $\rho_K$ and the regularization parameter as in Theorem (3.4) and using that $\lambda\kappa\rho_K = c'\kappa\epsilon r^2(\rho_K)$ lead immediately to (7).

3. **Bound over $F_3^{(\kappa)}$** is shown using simple homogeneity argument.

□

**Lemma 3.10.** *Let $\rho \geqslant 0$. Let $g \in F$ be such that $\|g-f^*\| \geqslant \rho$. Define $f = f^* + \rho\left(g-f^*\right)/\left\|g-f^*\right\|$. Then $f \in F$, $\|f-f^*\| = \rho$ and,*

$$\mathrm{MOM}_K\left(\left(g-f^*\right)^2 - 2\zeta\left(g-f^*\right) + \lambda\sup_{z^*\in\Gamma_{f^*}(\rho)} z^*\left(g-f^*\right) \geqslant\right.$$
$$\frac{\left\|g-f^*\right\|_{L_P^2}}{\rho}\left(\mathrm{MOM}_K\left(\left(f-f^*\right)^2 - 2\zeta\left(f-f^*\right)\right) + \lambda\sup_{z^*\in\Gamma_{f^*}(\rho)} z^*\left(f-f^*\right)\right).$$

*Proof.* The two first propositions are respectively a direct consequence of the convexity of $F$ and the definition of $f$.
For the last one, let $\Upsilon = \left\|g-f^*\right\|/\rho$ and note that $\Upsilon \geqslant 1$ and $g-f^* = \Upsilon\left(f-f^*\right)$, so we have

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

□

Now, let us bound $\sup_{g\in F_3^{(\kappa)}} T_{K,\lambda}\left(g,f^*\right)$. Let $g\in F_3^{(\kappa)}$. Applying lemma 3.8 and lemma 3.10 to $\rho=\rho_K$ : there exists $f\in F$ such that $\|f-f^*\|=\rho_K$ and

$$\begin{aligned}&T_{K,\lambda}\left(g,f^*\right)=\mathrm{MOM}_K\left(2\zeta\left(g-f^*\right)-\left(g-f^*\right)^2\right)-\lambda\left(\|g\|-\|f^*\|\right)\\&\leqslant \mathrm{MOM}_K\left(2\zeta\left(g-f^*\right)-\left(g-f^*\right)^2\right)-\lambda\sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*\left(g-f^*\right)+\lambda\frac{\kappa\rho_K}{10}\\&\leqslant\frac{\|g-f^*\|}{\rho_K}\left(\mathrm{MOM}_K\left(2\zeta\left(f-f^*\right)-\left(f-f^*\right)^2\right)-\lambda\sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*\left(f-f^*\right)\right)+\lambda\frac{\kappa\rho_K}{10}\end{aligned}\tag{12}$$

First assume that $\|f-f^*\|_{L_P^2}\leqslant r\left(\rho_K\right)$. In that case, $\|f-f^*\|=\rho_K$ and $\|f-f^*\|_{L_P^2}\leqslant r\left(\rho_K\right)$ 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 $\rho_K$ satisfies the sparsity equation. Therefore, $\sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*\left(f-f^*\right)\geqslant\Delta\left(\rho_K\right)\geqslant 4\rho_K/5$. Now, it follows from fact that $\lambda=\frac{c'\epsilon r^2(\rho_K)}{\rho_K}$ that

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

Moreover, since the quadratic process is non-negative, by 10 applied to $\rho=\rho_K$,

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

Finally, noting that $2\epsilon-4c'\epsilon/5\leqslant 0$ when $c'\geqslant 10/4$, binding all the pieces together in 12 yields

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

Second, assume that $\|f-f^*\|_{L_P^2}\geqslant r\left(\rho_K\right)$. Since $\|f-f^*\|=\rho_K$, it follows from 3.5 and 3.6 for $\rho=\rho_K$ that

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

where we used that $2\epsilon\leqslant\left(16\theta_0\right)^{-2}$ when $c\geqslant 32$ in the last inequality. Plugging the last result in 12 we get

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

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

### 3.4.6 Statistical performance

**Lemma 3.11.** *Let $\hat{f} \in F$ be such that, on $\Omega(K), C_{K,\lambda}(\hat{f}) \leqslant (2+c') \epsilon r^2 (\rho_K)$. Then, on $\Omega(K), \hat{f}$ satisfies $\left\|\hat{f} - f^*\right\| \leqslant 2\rho_K$, $\left\|\hat{f} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K)$ and $R(\hat{f}) \leqslant R(f^*) + (1+(4+3c')\epsilon) r^2 (2\rho_K)$, 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,

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

Thus, on $\Omega(K), \hat{f} \in \{g \in F : T_{K,\lambda}(g, f^*) \geqslant -(2+c')\epsilon r^2(\rho_K)\}$. 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 11 for $\rho = \kappa\rho_K$ that

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

□

We end the proof of theorem (3.4) using that, by definition of $\widehat{f}_{K,\lambda}$,

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

where $\left\{F_1^{(1)}, F_2^{(1)}, F_3^{(1)}\right\}$ is the decomposition of $F$. It follows from 8 (for $\kappa = 1$ ) tha on the event $\Omega(K)$,

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

Therefore, for $c' = 16$ and $c = 833$ the conclusion of the proof of Theorem 1 follows from 3.11.

# 4 Opening

In their article, Lecué and Lerasle (2017) proposed a new estimator for robust machine learning based on median-of-mean. As we shown in section (3.4), this estimator show very interesting theoretical properties. In fact it is shown that it optimal in the minimax sens in both $l_2$ and $L_P^2$ norm. In addition, those optimal rates of convergence are achieved under minimal assuptions on the dataset, knwoing the informative data are independant (not necessarly i.i.d) and the outliers are not assumed independant nor independant to the informative data nor identically distributed, in fact, they can even be adversarial. Also the authors developp a new notion that quantify the robustness of their estimator which is non-asymptotic and takes into acount the statistical performances of the estimators and show that their estimator have a *breakdown number* of order *number of iterations*$\times$*rate of convergence*.

In addition, the estimator developped in the article easily computable in practice and is appliable to, basically, any problem which require to estimate the empirical risk minimizer (ERM). In fact, we focused on the MOM version of the LASSO estimator in this work but it can also be used for classical ERM estimation without regulraization or with other ones, like SLOPE regularization.

Also, the estimator can be used for outliers detection which is an active field of research in the data science and machine learning community. This come from the randomization of the blocks at each step of the descent algorithm, which, in addition of improving a lot the performance of the estimator, gives a measure of singularity of the data which can be used as an outlier dection tool.

# 5 Appendix : Adaptative choice of $K$ with Lepski's method

The choice of $K$ in the MOM estimator is of primary importance as all rates in Theorem (3.4) depends on it. This construction, inspired from the Lepski's method provides an adaptative choice of this parameter (see Goldenshluger and Lepski (2011) for more information about this method). For $\lambda > 0, f \in F$ and a constant $c_{ad} > 0$, the adaptative choice of $K$ is given, for all $J \in [\max(K^*, 8|\mathscr{O}|), N/(96(\theta_0\theta_{r0})^2)]$

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

$$\text{and choose } \widehat{f}_{c_{ad}} \in \cap_{J=\hat{K}_{c_{ad}}}^{N/\left(96(\theta_0\theta_{r0})^2\right)} \hat{R}_{J,c_{ad}}.$$

with

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

For this choice of $K$ we get the following rates

**Theorem 5.1.** *Grant the assumptions of Theorem 1 and assume moreover that and* $|\mathscr{O}| \leqslant N/\left(768\theta_0^2\theta_{r0}^2\right)$. *For any* $K \in \left[\max\left(K^*, 8|\mathscr{O}|\right), N/\left(96\left(\theta_0\theta_{r0}\right)^2\right)\right]$, *with probability larger than*

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

*one has*

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

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

# References

Giraud, C. (2021). Introduction to high-dimensional statistics.

Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608 – 1632.

Hampel, F. (1973). Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 27:87–104.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: Theory and practice. *Annals of Statistics*, 48.

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via mom minimization.

Lecué, G. and Mendelson, S. (2016). Learning subgaussian classes : Upper and minimax bounds.

Maronna, R., Martin, D., and Yohai, V. (2006). Robust statistics: Theory and methods.

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1):1 – 67.