Submitted to the Annals of Statistics
arXiv: arXiv:1711.10306

ROBUST MACHINE LEARNING BY MEDIAN-OF-MEANS: THEORY AND PRACTICE

By Guillaume Lecué[*],[†] and Matthieu Lerasle[‡]

CREST, CNRS, ENSAE[†] *and CNRS, Université Paris Sud Orsay* [‡]

Median-of-means (MOM) based procedures have been recently introduced in learning theory [43, 35]. These estimators outperform classical least-squares estimators when data are heavy-tailed and/or are corrupted. None of these procedures can be implemented, which is the major issue of current MOM procedures [44].

In this paper, we introduce minmax MOM estimators and show that they achieve the same subgaussian deviation bounds as the alternatives [43, 35], both in small and high-dimensional statistics. In particular, these estimators are efficient under moments assumptions on data that may have been corrupted by a few outliers.

Besides these theoretical guarantees, the definition of minmax MOM estimators suggests simple and systematic modifications of standard algorithms used to approximate least-squares estimators and their regularized versions. As a proof of concept, we perform an extensive simulation study of these algorithms for robust versions of the LASSO.

1. Introduction.

Consider the least-squares regression problem where, given a dataset $(X_i, Y_i)_{i\in\{1,\dots,N\}}$ of points in $\mathcal{X}\times\mathbb{R}$ and a new input $X\in\mathcal{X}$, one wants to predict the associated real valued output $Y\in\mathbb{R}$. A classical approach is to consider (X, Y) as a random variable with values in $\mathcal{X}\times\mathbb{R}$ and, given a set F of functions $f:\mathcal{X}\to\mathbb{R}$, to look for the oracle in F, which is defined by

$$f^*\in\operatorname*{argmin}_{f\in F} P(Y-f(X))^2\enspace .$$

To estimate f^*, we have a dataset $(X_i, Y_i)_{i\in\{1,\dots,N\}}$ for which there exists a partition $\{1,\dots,N\}=\mathcal{O}\cup\mathcal{I}$ such that data $(X_i, Y_i)_{i\in\mathcal{I}}$ are *inliers* or *informative* and data $(X_i, Y_i)_{i\in\mathcal{O}}$ are "outliers" in the sense that *nothing* is assumed on these data. On inliers, one grants independence and finiteness of some moments, allowing for "heavy-tailed" data. Moreover, departing from the independent and identically distributed (i.i.d.) setup, we also allow

[*] Supported by Labex ECODEC (ANR - 11-LABEX-0047), Maison des Sciences de l'Homme (MSH – 17MA01) and a donation of NVIDIA Corporation.

MSC 2010 subject classifications: Primary 60K35, 62G08; secondary 62C20, 62G05, 62G20

Keywords and phrases: Empirical processes, High-dimensional statistics

inliers to have different distributions than (X, Y). We assume that, for all $i \in \mathcal{I}$ and all $f \in F$,

$$\begin{aligned} \mathbb{E}[(Y_i - f^*(X_i))(f - f^*)(X_i)] &= \mathbb{E}[(Y - f^*(X))(f - f^*)(X)] \\ \mathbb{E}[(f - f^*)^2(X_i)] &= \mathbb{E}[(f - f^*)^2(X)] \ . \end{aligned}$$

These assumptions imply that the distribution P of (X, Y) and the distribution P_i of (X_i, Y_i) for $i \in \mathcal{I}$ induce the same L^2-geometry on $F - f^* = \{f - f^* : f \in F\}$ and therefore, in particular, that the oracles w.r.t. P and P_i for any $i \in \mathcal{I}$ are the same. Of course, the sets $\mathcal{O}$ and $\mathcal{I}$ are unknown to the statistician.

Regression problems with possibly heavy-tailed inliers cannot be handled by classical least-squares estimators, which are particular instances of empirical risk minimizers (ERM) of Vapnik [67]. Least-squares estimators have subgaussian deviations under stronger assumptions, such as boundedness [45] or sub-gaussian [38] assumptions on the noise and the design. In this paper, the main hypothesis is the *small ball assumption* of [32, 49] which says that $L^2(P)$ and $L_1(P)$ norms are equivalent over $F - f^*$ – see Section 3.1 for details. Although sometimes restrictive [58, 26], this assumption does not involve high moment conditions unnecessary for the problem to make sense.

Least-squares estimators and their regularized versions are also useless in corrupted environments. This has been known for a long time and can easily be checked in practice. Figure 1 for example shows estimation bounds of the LASSO [60] on a dataset containing a single outlier in the outputs.

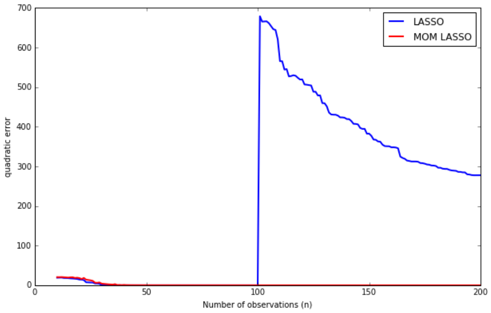

Fig 1. *Estimation error of the LASSO (blue curve) and MOM LASSO (red curve) after one outliers was added at observation* 100.

These restrictions of least-squares estimators gave rise in the 1960's to the theory of robust statistics of John Tukey [61, 62], Peter Huber [27, 28] and Frank Hampel [24, 25]. The most classical alternatives to least-squares estimators are M-estimators, which are ERM based on loss functions $\ell_f(X,Y)$ less sensitive to outliers than the square loss, such as a truncated version of the square loss. The idea is that, while $(Y_i - f(X_i))^2$ can be very large for some outliers data and influence all the empirical mean $N^{-1}\sum_{i=1}^N (Y_i - f(X_i))^2$, the influence of these anomalies will be asymptotically null if $\ell_f(X_i, Y_i)$ is bounded. Recent works study deviation properties of M-estimators: [22, 57, 21, 56] considered the Huber-loss in linear regression with heavy-tailed noise and subgaussian design. They obtain minimax optimal deviation bounds in this setting. The limitation on the design is not surprising: it is well known that M-estimators using loss functions such as Huber or L_1 loss are not robust to outliers in the inputs X_i. This problem is called the"leverage points problem" [29]. In a slightly different approach than M-estimation, [6] proposed a minmax estimator based on losses introduced in [18] in a least-squares regression framework and prove optimal subgaussian bounds under a L_2 assumption on the noise and a L_4/L_2 assumption on the design, which is close to the assumptions we grant on inliers.

This paper focuses on Median-of-means (MOM) [2, 30, 54], which provide alternatives to M-estimators. MOM estimators of the real valued expectation $\mathbb{E}[Z]$ are built as follows, the dataset $Z_1, \dots, Z_N$ is partitioned into blocks $(Z_i)_{i\in B_k}$, $k = 1, \dots, K$ of the same cardinality. The MOM estimator is the median of the K empirical means constructed on each block:

$$\mathrm{MOM}_K\big(Z\big) = \text{median}\left\{ \frac{1}{|B_k|}\sum_{i\in B_k} Z_i,\ k = 1, \dots, K \right\} \ .$$

Subgaussian properties of these estimators can be found in [20, 40].

As in [43, 35], MOM estimators are used to estimate real valued increments of square risks $P[(Y - f(X))^2 - (Y - g(X))^2]$, where $f, g \in F$. This construction does not require a notion of median in dimension larger than 1, contrary to "geometric median-of-means" approach presented in [52, 51]. In [43, 35], each $f \in F$ receives a score which is the $L^2(P)$-diameter $\Delta(f)$ of the set $\mathcal{B}(f)$, where $g \in \mathcal{B}(f)$ if $\mathrm{MOM}_K\big(\ell_f - \ell_g\big) < 0$. The approach of [43, 35] requires therefore an evaluation of the diameter of the sets $\mathcal{B}(f)$ for all $f \in F$, which makes the procedure impossible to implement.

This paper presents an alternative to [43, 35] which relies on the following minmax formulation. By linearity of P, f^* is solution of:

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

Replacing the real valued means $P[(Y-f(X))^2-(Y-g(X))^2]$ in this equation by their MOM estimators produces the minmax MOM estimators of f^* which are rigorously introduced in Section 2.3. Compared with [43, 35], minmax MOM estimators do not require an estimation of L^2-distances between elements in F and are therefore simpler to define. Minmax strategies have also been considered in [6] and [9, 8]. The idea of building estimators of f^* from estimators of increments goes back to seminal works by Le Cam [33, 34] and was further developed by Birgé with the T-estimators [14]. In Le Cam and Birgé's works, the authors used "robust tests" to compare densities f and g and deduce from these an alternative to the non-robust maximum likelihood estimators. Baraud [10] showed that robust tests could be obtained by estimating the difference of Hellinger risks of f and g and used a variational formula to build these new tests. Finally, Baraud, Birgé and Sart [9] used Baraud's estimators of increments in a minmax procedure to build ρ-estimators.

The first aim of this paper is to show that minmax MOM estimators satisfy the same subgaussian deviation bounds as other MOM estimators [42, 35]. The analysis of minmax MOM estimators is conceptually and technically simpler: an adaptation of Lemmas 5.1 and 5.5 in [43] or Lemmas 2 and 3 [35] is sufficient to prove subgaussian bound for minmax MOM estimators while a robust estimation (based on MOM estimates) of the $L^2(P)$-metric was required in [42, 35].

Another advantage of the minmax MOM approach lies in the Lepski-step (see Theorem 2), which selects adaptively the number K of blocks. This step is way easier to implement and to study than the one presented in [35], as only one confidence region is sufficient to grant adaptation with respect to the excess risk, the regularization and L^2 norms. Recall that, in corrupted environments, a data-driven choice of K has to be performed since K must be larger than twice the (unknown) number of outliers. Note that the idea of aggregating estimators built on blocks of data and selecting the number of blocks by Lepski's method was already present in Birgé [13, proof of Theorem 1]. It was also used in [20] to build "multiple-δ" subgaussian estimators of univariate means.

In our opinion, the most interesting feature of the minmax formulation is that it suggests a generic method to modify descent algorithms designed to approximate ERM and their regularized versions and make them efficient even if run on corrupted datasets. Let us give a rough presentation of a "MOM version" of descent algorithms: at each time-step t, all empirical means $P_{B_k}(Y-f_t(X))^2$ for $k=1,\dots,K$ are evaluated and one computes the index $k_{\text{med}} \in [K]$ of the block such that

$$P_{B_{k_{\text{med}}}}(Y - f_t(X))^2 = \text{med}\left\{P_{B_k}(Y - f_t(X))^2,\ k = 1, \ldots, K\right\} \ .$$

The descent direction is the opposite gradient $-\nabla(f \to P_{B_{k_{\text{med}}}}(Y - f(X))^2)_{|f=f_t}$. This descent algorithm can be turned into a descent-ascent algorithm approximating minmax MOM estimators. Section 5 presents several examples of modifications of classical algorithms.

In practice, these basic algorithms perform poorly when applied on a *fixed* partition of the dataset. However, empirical performance are improved when the partition is chosen uniformly at random at each descent step of the algorithm, cf. Section 6.2. In particular, the shuffling step prevents the algorithms to converge to local minimaxima. Besides, randomized algorithms define a notion of depth of data: each time a data belongs to the median block, its "score" is incremented by 1. The higher the final score is, the deeper is the data. This notion of depth is based on the risk function which is natural in a 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. This by-product is presented in Section 6.2.

The paper is organized as follows. Section 2 introduces the framework and presents the minmax MOM estimator, Section 3 details the main theoretical results. These are illustrated in Section 4 on some classical problems of machine learning. Many robust versions of standard optimization algorithms are presented in Section 5. An extensive simulation study illustrating our results is performed in Section 6. Proofs of the main results, complementary theorems showing minimax optimality of our bounds are postponed to the supplementary material.

2. Setting. Let $\mathcal{X}$ denote a measurable space. Let $(X_i, Y_i)_{i\in\{1,\ldots,N\}}$, (X, Y) denote random variables taking values in $\mathcal{X} \times \mathbb{R}$. Let P denote the distribution of (X, Y) and, for $i \in \{1, \ldots, N\}$, 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 $\mathbb{E}[Y^2] < \infty$. For any $Q \in \{P, (P_i)_{i\in[N]}\}$ and any $p \geqslant 1$, let L^p_Q denote the set of functions f such that the norm $\|f\|_{L^p_Q} = (Q|f|^p)^{1/p}$, where $Qg = \mathbb{E}_{Z\sim Q}[g(Z)]$. Assume that $F \subset L^2_P$. For any $(x, y) \in \mathcal{X} \times \mathbb{R}$, let $\ell_f(x, y) = (y - f(x))^2$ denote the square loss and let f^* denote an oracle

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

Let $R(f) = P\ell_f$ denote the risk. The goal is to build estimators $\hat{f}$ satisfying: with probability at least $1-\delta$,

$$R(\hat{f}) \leq \min_{f\in F} R(f) + r_N^{(1)} \text{ and } \left\|\hat{f} - f^*\right\|_{L_P^2} \leq r_N^{(2)} \ .$$

The residue $r_N^{(1)}$ of the oracle inequality, the estimation rate $r_N^{(2)}$ and the confidence level δ should be as small as possible. Oracle inequalities provide risk bounds for the estimation the regression function $\overline{f}(x) = \mathbb{E}[Y|X=x]$: $R(\hat{f}) \leq R(f^*) + r_N^{(1)}$ is equivalent to

$$\|\overline{f} - \hat{f}\|_{L_P^2}^2 \leqslant \|\overline{f} - f^*\|_{L_P^2}^2 + r_N^{(1)} \ .$$

Finally, let $\|\cdot\|$ be a norm defined on the span of F; $\|\cdot\|$ will be used as a regularization norm to induce some low dimensional structure or some regularity, such as the ℓ_1 or SLOPE norm (see Section 4).

2.1. *Minmaximization.* The oracle f^* is solution of the minmax problem:

$$f^* \in \operatorname*{argmin}_{f\in F} P\ell_f = \operatorname*{argmin}_{f\in F} \sup_{g\in F} P(\ell_f - \ell_g) \ . \tag{2}$$

Any estimator of real valued expectations $P\ell_f$ or $P(\ell_f - \ell_g)$ can be plugged in (2) to obtain estimators of f^*. Plugging the empirical means (in both the min and the minmax problems) yields the classical ERM over F for example. In general, plugging non-linear (robust or not) estimators of the mean in the minmax problem or in the min problem in (2) does not yield the same estimator of f^* though. The main advantage of the minmax formulation is that it allows to bound the risk of the estimator using the complexity of F around f^*. This "localization" idea is central to derive optimal (fast) rates for the ERM [31, 45, 15] and cannot be used directly when empirical means are simply replaced by non linear estimators of the mean in a minimization formulation.

2.2. *MOM estimators.* Let K denote an integer smaller than $N/2$ and let $B_1, \dots, B_K$ denote a partition of $[N] = \{1, \dots, N\}$ into blocks of equal size N/K (w.l.o.g. we assume that K divides N). For all functions $\mathcal{L} : \mathcal{X}\times\mathbb{R} \to \mathbb{R}$ and $k \in [K] = \{1, \dots, K\}$, let $P_{B_k}\mathcal{L} = |B_k|^{-1}\sum_{i\in B_k}\mathcal{L}(X_i, Y_i)$.

For all $\alpha \in (0,1)$ and real numbers $x_1, \dots, x_K$, denote by $\mathcal{Q}_\alpha(x_1, \dots, x_K)$ the set of α-quantiles of $\{x_1, \dots, x_K\}$:

$$\{u \in \mathbb{R} : \quad |\{k \in [K] : x_k \geqslant u\}| \geqslant (1-\alpha)\ell, \quad |\{k \in [K] : x_k \leqslant u\}| \geqslant \alpha\ell\}$$

and let $Q_\alpha(x)$ denote any point in $\mathcal{Q}_\alpha(x_1,\dots,x_K)$. For $x=(x_1,\dots,x_K)\in\mathbb{R}^K$ and $t\in\mathbb{R}$, we say that $Q_\alpha(x)\geqslant t$ when there exists $J\subset[K]$ such that $|J|\geqslant(1-\alpha)K$ and for all $k\in J, x_k\geqslant t$; we write $Q_\alpha(x)\leqslant t$ if there exists $J\subset[K]$ such that $|J|\geqslant\alpha K$ and for all $k\in J, x_k\leqslant t$.

Let $y=(y_1,\dots,y_K)\in\mathbb{R}^K$. We write $Q_{1/2}(x-y)\leqslant Q_{3/4}(x)-Q_{1/4}(y)$ when there exist $u,l\in\mathbb{R}$ such that $Q_{1/2}(x-y)\leqslant u-l$, $Q_{3/4}(x)\leqslant u$ and $Q_{1/4}(y)\geqslant l$.

DEFINITION 1. *Let $\alpha\in(0,1)$, $K\in[N]$. For any $\mathcal{L}:\mathcal{X}\times\mathbb{R}\to\mathbb{R}$ 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 $MOM_K\big(\mathcal{L}\big)=Q_{1/2,K}(\mathcal{L})$. For all $f,g\in F$, the* ***MOM estimator on** K **blocks of the loss increment from** g **to** f* *is defined by*

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

and, for a given regularization parameter $\lambda\geqslant0$, its regularized version is

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

2.3. *Minmax MOM estimators.* Minmax MOM estimators are obtained by replacing the unknown expectations $P(\ell_f-\ell_g)$ in (2) by their MOM estimators.

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

$$\hat f_K\in\operatorname*{argmin}_{f\in F}\max_{g\in F}T_K(g,f)\ \textit{and}\ \hat f_{K,\lambda}\in\operatorname*{argmin}_{f\in F}\max_{g\in F}T_{K,\lambda}(g,f). \tag{3}$$

We shall provide results for $\hat f_{K,\lambda}$ only in the main text. The estimators $\hat f_K$ are studied in the supplement in Section 7.

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).*

3. Assumptions and main results.

Denote by $\{\mathcal{O},\mathcal{I}\}$ a partition of $[N]$ and by $|\mathcal{O}|$ the cardinality of $\mathcal{O}$. On $(X_i,Y_i)_{i\in\mathcal{O}}$, **no assumptions** is granted, these data are *outliers*. They may not be independent, nor independent from the remaining data (not even random). $(X_i,Y_i), i\in\mathcal{I}$ are called *inliers* or *informative* data. They are hereafter assumed *independent*. The sets $\mathcal{O},\mathcal{I}$ are unknown.

3.1. *Assumptions.* The main assumptions involve first and second moments of the functions in F and Y under the distributions $P, (P_i)_{i\in\mathcal{I}}$.

Assumption 1. *For all $f \in F$ and all $i \in \mathcal{I}$,*

$$P_i(f-f^*)^2 = P(f-f^*)^2 \text{ and } P_i[\zeta(f-f^*)] = P[\zeta(f-f^*)] \enspace .$$

where $\zeta(x,y) = (y - f^*(x))$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}$.

Assumption 1 holds in the i.i.d. framework, with $\mathcal{I} = [N]$ but it covers also other cases where inliers follow different distributions (see, for instance, multimodal datasets such as in [46] or heteroscedastic noise [4]). It is also possible to weaken Assumption 1 such as in [35]. The second assumption bounds the correlation between $\zeta_i = Y_i - f^*(X_i)$ and the shifted class $F - f^*$.

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

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

Assumption 2 holds when data are i.i.d. and $Y - f^*(X)$ has uniformly bounded L^2-moments conditionally to X. This last assumption holds when $Y - f^*(X)$ is independent of X and has a L^2-moment bounded by θ_m. 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 $f \in F$, $\|f - f^*\|_{L_{P_i}^4} \leqslant \theta_1 \|f-f^*\|_{L_P^2}$. Actually, in this case,

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

so Assumption 2 holds for $\theta_m = \theta_1\theta_2$. The third assumption states that the norms L_P^2 and L_P^1 are equivalent over $F - f^*$.

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

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

Under Assumption 1, $\|f-f^*\|_{L_{P_i}^1} \leqslant \|f-f^*\|_{L_{P_i}^2} = \|f-f^*\|_{L_P^2}$ for all $f \in F$ and $i \in \mathcal{I}$, hence, Assumptions 1 and 3 imply that the norms $L_P^1, L_P^2, L_{P_i}^2, L_{P_i}^1, i \in \mathcal{I}$ are equivalent over $F - f^*$. Assumption 3 is equivalent to the small ball property (cf. [32, 49]), see Proposition 1 in [35].

3.2. *Complexity measures.* For all $\rho, r \geqslant 0$, let

$$B(f^*, \rho) = \{f \in F : \|f - f^*\| \leqslant \rho\},\ B_2(f^*, r) = \{f \in F : \|f - f^*\|_{L_P^2} \leqslant r\}\ .$$

Definition 3. *Let $(\epsilon_i)_{i\in[N]}$ be independent random variables uniformly distributed in $\{-1, 1\}$, independent from $(X_i, Y_i)_{i=1}^N$. For all $f \in F$, $r > 0$ and $\rho \in (0, +\infty]$, let*

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

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

$$\inf\left\{r > 0 : \sup_{J\subset\mathcal{I}, |J|\geqslant N/2} \mathbb{E} \sup_{f\in B_{reg}(f^*,\rho,r)} \left|\frac{1}{|J|}\sum_{i\in J} \epsilon_i (f - f^*)(X_i)\right| \leqslant \gamma_Q r\right\}\ ,$$

and $r_M(\rho, \gamma_M)$ as

$$\inf\left\{r > 0 : \sup_{J\subset\mathcal{I}, |J|\geqslant N/2} \mathbb{E} \sup_{f\in B_{reg}(f^*,\rho,r)} \left|\frac{1}{|J|}\sum_{i\in J} \epsilon_i \zeta_i (f - f^*)(X_i)\right| \leqslant \gamma_M r^2\right\}\ .$$

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 [38] that r_M and r_Q are continuous and non decreasing functions, that depend on f^*. According to [38], for appropriate choice of γ_Q, γ_M, $r(\rho) = \max(r_M(\rho, \gamma_M), r_Q(\rho, \gamma_Q))$ is the minimax rate of convergence over $B(f^*, \rho)$. Note also that r_Q and r_M are well defined when $|\mathcal{I}| \geqslant N/2$, meaning that at least half data should be informative.

3.3. *The sparsity equation.* Risk bounds follow from upper bounds on $T_{K,\lambda}(f, f^*)$ for functions f far from f^* either in L_P^2-norm or for the regularization norm $\|\cdot\|$. Let $f \in F$ and let $\rho = \|f - f^*\|$. When $\|f - f^*\|_{L_P^2}$ is small, $T_{K,\lambda}$ has to be bounded from above by $\lambda(\|f^*\| - \|f\|)$. To bound $\|f^*\| - \|f\|$ from bellow, introduce the subdifferentials of $\|\cdot\|$. Let $(E^*, \|\cdot\|^*)$ be the dual normed space of $(E, \|\cdot\|)$ and for all $f \in F$, let

$$(\partial\|\cdot\|)_f = \{z^* \in E^* : \forall h \in E, \|f + h\| \geqslant \|f\| + z^*(h)\}\ .$$

For any $\rho > 0$, let H_ρ denote the set of functions "close" to f^* in L_P^2 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^*:

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

and $H_\rho = \{f \in F : \|f - f^*\| = \rho \text{ and } \|f - f^*\|_{L_P^2} \leqslant r(\rho)\}$. If there exists f^{**} such that $\|f^* - f^{**}\| \leqslant \rho/20$ and $(\partial \|\cdot\|)_{f^{**}}$ 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. Formally, for all $\rho > 0$, let

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

The sparsity equation, introduced in [39], quantifies these notions of "large".

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

If ρ^* satisfies the sparsity equation, so do all $\rho \geqslant \rho^*$. Let

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

If $\rho \geqslant 20\|f^*\|$, then $0 \in \Gamma_{f^*}(\rho)$. Moreover, $(\partial \|\cdot\|)_0$ is the unit ball of $(E^*, \|\cdot\|^*)$, so $\Delta(\rho) = \rho$. This implies that any $\rho \geqslant 20\|f^*\|$ satisfies the sparsity equation. This simple observation can be used to get "complexity-dependent rates of convergence" [37].

3.4. *Main results.* The first results give risk bounds for $\widehat{f}_{K,\lambda}$. Similar bounds have been obtained for other MOM estimators [42, 35].

Theorem 1. *Grant Assumptions 1, 2 and 3 and let r_Q, r_M denote the complexity functions introduced in Definition 3. Assume that $N \geqslant 384\theta_0^2$ and $|\mathcal{O}| \leqslant N/(768\theta_0^2)$. Let ρ^* be solution to the sparsity equation from Definition 4. Let $\epsilon = 1/(833\theta_0^2)$ and $r^2(\cdot)$ is defined in Definition 3 for $\gamma_Q = (384\theta_0)^{-1}$ and $\gamma_M = \epsilon/192$. Let K^* denote the smallest integer such that*

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

For any $K \geqslant K^*$, define the radius ρ_K and the regularization parameter as

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

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

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

The function r is used to define the regularization parameter in Theorem 1, so it cannot depend on f^*. When r_M, r_Q depend on f^*, r should be a computable upper bound independent from f^*. The best rates of estimation and prediction that follow from Theorem 1 are obtained for $K = K^*$ when $|\mathcal{O}| \leq K^*/8 \sim Nr^2(\rho^*)$. In that case, it is proved in Section 4 on two examples that the rate ρ_{K^*} and the residue $r(2\rho_{K^*})$ are minimax optimal. In a setup where data only induce the same L^2 metric as P and may have been corrupted by up to $K^*/8 \sim Nr^2(\rho^*)$ outliers, Theorem 1 shows that our estimators achieve the subgaussian deviations bounds of the ERM when data are i.i.d. with a noise ζ independent of X and both X and ζ have Gaussian distributions (see Section 8 in the supplement).

3.4.1. *Adaptive choice of K.* In Theorem 1, all rates depend on K, which has to be larger than the number of outliers and $Nr^2(\rho^*)$. These quantities are unknown in general, for instance, $Nr^2(\rho^*) \sim s\log(ed/s)$ in high-dimensional statistics where s is the unknown sparsity parameter. This section presents an adaptive choice of K inspired from Lepski's method that allows to bypass this issue. However, this construction requires the knowledge of constants θ_0 and θ_m (see Section 6.1 for a fully data driven choice of K in practice).

For all $J \in [K]$, $\lambda > 0$, $f \in F$ and $c_{ad} > 0$, let

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

Let $\hat{K}_{c_{ad}} = \inf\left\{ K \in [1, N/(96\theta_0^2)] : \cap_{J=K}^{N/(96\theta_0^2)} \hat{R}_{J,c_{ad}} \neq \emptyset \right\}$ and

$$\widehat{f}_{c_{ad}} \in \bigcap_{J=\hat{K}_{c_{ad}}}^{N/(96\theta_0^2)} \hat{R}_{J,c_{ad}} \ . \tag{4}$$

The following theorem gives risk bounds for these estimators. Bounds in regularization and L^2_P norms have been proved for Le Cam test estimators in [35]. To the best of our knowledge, adaptive bounds in excess risk have never been proved before.

THEOREM 2. *Grant the assumptions of Theorem 1. Choose $c_{ad} = 18/833$ in (4) and let $\epsilon = (833\theta_0^2)^{-1}$. For any $K \in [\max(K^*, 8|\mathcal{O}|), N/(96\theta_0^2)]$, with probability larger than*

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

one has $\left\|\widehat{f}_{c_{ad}} - f^*\right\| \leqslant 2\rho_K$, $\quad \left\|\widehat{f}_{c_{ad}} - f^*\right\|_{L_P^2} \leqslant r(2\rho_K)$ *and*

$$R(\widehat{f}_{c_{ad}}) \leqslant R(f^*) + (1+52\epsilon) r^2(2\rho_K) \ ,$$

In particular, for $K = K^*$, *we have* $r(2\rho_{K^*}) = \max\left(r(2\rho^*), \sqrt{|\mathcal{O}|/N}\right)$.

Theorem 2 shows that $\widehat{f}_{c_{ad}}$ achieves similar performance as $\widehat{f}_{K,\lambda}$ simultaneously for all K from K^* to $O(N)$. For $K = K^*$, these rates match the optimal minimax rates of convergence, see Section 4. The main difference with Theorem 1 is that the knowledge of K^* and $|\mathcal{O}|$ is not necessary to design $\widehat{f}_{c_{ad}}$. This is very useful in applications where these quantities are typically unknown. Moreover, both the construction and the analysis are much simpler for $\widehat{f}_{c_{ad}}$ than the adaptive estimator in [35] since they are based on the analysis of confidence regions for $\mathcal{C}_{J,\lambda}$ only, instead of multiple criteria in [35].

Remark 2 (deviation parameter). *Note that* $r(\cdot)$ *can be any continuous, non decreasing function such that* $r(\rho) \geqslant \max\left(r_Q(\rho,\gamma_Q), r_M(\rho,\gamma_M)\right)$. *In particular, if* $r_* : \rho \to \max\left(r_Q(\rho,\gamma_Q), r_M(\rho,\gamma_M)\right)$ *is continuous, as 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 Theorem 2 but with an extra confidence parameter: for all* $x > 0$, *with probability at least* $1 - 4\exp(-c_0 N r_*^2(\rho_{K^*}) + c_0 x)$,

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

In that case, $\widehat{f}_{c_{ad}}$ *depends on* x *since* $\lambda = 16\epsilon(r_*(\rho_K) + x/N)/\rho_K$.

4. Examples of applications. This section presents two examples of regularization in high-dimensional statistics: the ℓ_1 and the SLOPE norms.

4.1. *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 show some limitations of LASSO [69, 64, 55], 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. One can refer to [12, 41, 65, 66, 48, 53, 63] for estimation and sparse oracle inequalities, [47, 70, 7] for support recovery results; more results and references on LASSO can be found in [17, 31].

4.2. *SLOPE.* SLOPE is an estimator introduced in [16, 59]. 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 $(\beta_1, ..., \beta_d) = (1, ..., 1)$.

4.3. *Classical results for LASSO and SLOPE.* Typical results for LASSO and SLOPE have been obtained when data are i.i.d. with subgaussian design X and, most of the time, subgaussian noise ζ as well.

Definition 5. *Let ℓ_2^d be a d-dimensional inner product space and let X be a random variable with values in ℓ_2^d. We say that X is isotropic when for every $t \in \ell_2^d$, $\|\langle X, t\rangle\|_{L_P^2} = \|t\|_{\ell_2^d}^2$ and it is L-subgaussian if for every $t \in \ell_2^d$ and every $p \geqslant 2$, $\|\langle X, t\rangle\|_{L_P^p} \leqslant L\sqrt{p}\|\langle X, t\rangle\|_{L_P^2}$.*

The covariance structure of an isotropic random variable coincides with the inner product in ℓ_2^d. If X is a L-subgaussian random vector, the L_P^p norms of all linear forms do not grow faster than the L_P^p norm of a Gaussian variable. When dealing with the LASSO and SLOPE, the natural Euclidean structure is used in $\mathbb{R}^d$.

Assumption 4.
1. *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 $f^* = \langle t^*, \cdot\rangle$, $\xi = Y - f^*(X) \in L_P^{q_0}$ for some $q_0 > 2$.*

Assumption 4 requires a L^{q_0} for $q_0 > 2$ moment on the noise. LASSO and SLOPE still achieve optimal rates of convergence under this assumption but with a severely deteriorated probability estimate.

Theorem 3 (Theorem 1.4 in [36]). *Grant Assumption 4. 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 $\|t^* - v\|_1 \leqslant c_2\|\xi\|_{L_P^{q_0}} s\sqrt{\log(ed)/N}$. The Lasso estimator $\hat{t}$ with regularization parameter $\lambda = c_3\|\xi\|_{L_P^{q_0}}\sqrt{\log(ed)/N}$ is such that with probability at least*

$$1 - \frac{c_4 \log^{q_0} N}{N^{q_0/2-1}} - 2\exp\left(-c_5 s \log(ed/s)\right) \tag{5}$$

for every $1 \leqslant p \leqslant 2$

$$\left\|\hat{t} - t^*\right\|_p \leqslant c_6 \|\xi\|_{L_P^{q_0}} s^{1/p} \sqrt{\frac{\log(ed)}{N}}.$$

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

Theorem 3 shows that LASSO achieves its optimal rate (cf. [12]) if t^* is close to a sparse vector and the noise ζ may be heavy tailed and may not be independent from X. On the other hand, the dataset cannot contain outliers and the data should be i.i.d. with subgaussian design matrix X.

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 [36]). *Consider the SLOPE under Assumption 4. 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* $\Psi(t^* - v) \leqslant c_2 \|\xi\|_{L_P^{q_0}} s \log(ed/s)/\sqrt{N}$. *The SLOPE estimator with* $\lambda = c_3 \|\xi\|_{L_P^{q_0}}/\sqrt{N}$ *satisfies, with probability at least* (5),

$$\Psi(\hat{t} - t^*) \leqslant c_4 \|\xi\|_{L_P^{q_0}} \frac{s}{\sqrt{N}} \log\left(\frac{ed}{s}\right), \qquad \left\|\hat{t} - t^*\right\|_2^2 \leqslant c_5 \|\xi\|_{L_P^{q_0}}^2 \frac{s}{N} \log\left(\frac{ed}{s}\right) .$$

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

4.4. *Minmax MOM LASSO and 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 or the SLOPE norm. The aim is to show that the results from Section 4.3 hold and are sometimes even improved by MOM versions of LASSO and SLOPE under weaker assumptions and with a better probability deviation. Start with the new set of assumptions.

Assumption 5. *Denote by* $(e_j)_{j=1}^d$ *the canonical basis of* $\mathbb{R}^d$ *and assume*

1. $(X, Y), (X_i, Y_i)_{i \in \mathcal{I}}$ *are i.i.d.*
2. X *is isotropic and for every* $t \in \mathbb{R}^d$, $p \in [C_0 \log(ed)]$ *and* $j \in [d]$, $\left\|\left\langle X, e_j\right\rangle\right\|_{L_P^p} \leqslant L\sqrt{p} \left\|\left\langle X, e_j\right\rangle\right\|_{L_P^2}$,
3. $\xi = Y - \left\langle t^*, X\right\rangle \in L_P^{q_0}$ *for some* $q_0 > 2$.
4. *there exists* θ_0 *such that for all* $t \in \mathbb{R}^d$, $\left\|\left\langle X, t\right\rangle\right\|_{L_P^2} \leqslant \theta_0 \left\|\left\langle X, t\right\rangle\right\|_{L_P^1}$,
5. *there exists* θ_m *such that* $\mathrm{var}(\xi\left\langle X, t\right\rangle) \leqslant \theta_m \left\|\left\langle X, t\right\rangle\right\|_{L_P^2}$.

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

$$\ell^*(V) = \mathbb{E} \sup_{v \in V} \big\langle G, v \big\rangle, \quad \text{where} \quad G \sim \mathcal{N}_d(0, I_d) \ . \tag{6}$$

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 $\mathbb{R}^d$ [50, Definition 1.4]. Therefore, [50, Theorem 1.6] applies under the following assumption.

Assumption 6. *There exist constants $q_0 > 2$, C_0 and L such that $\xi \in L_P^{q_0}$, X is isotropic and for every $j \in [d]$ and $1 \leqslant p \leqslant C_0 \log d$, $\big\| \big\langle X, e_j \big\rangle \big\|_{L_P^p} \leqslant L\sqrt{p} \big\| \big\langle X, e_j \big\rangle \big\|_{L_P^2}$.*

Under Assumption 6, if $\sigma = \|\xi\|_{L_P^{q_0}}$, [50, Theorem 1.6] shows that for $\zeta_i = Y_i - \big\langle X_i, t^* \big\rangle$ and for every $\rho > 0$,

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

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

Local Gaussian mean widths $\ell^*(\rho B_1^d \cap r B_2^d)$ are bounded from above in [39, 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

$$r^2(\rho) \sim_{L, q_0, \gamma_Q, \gamma_M} \begin{cases} \max\left(\rho\sigma \sqrt{\frac{1}{N} \log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\sigma^2 d}{N}\right) & \text{if } N \gtrsim_L d \\ \max\left(\rho\sigma \sqrt{\frac{1}{N} \log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\rho^2}{N} \log\left(\frac{d}{N}\right)\right) & \text{otherwise} \end{cases} \ . \tag{7}$$

Now we turn to a solution of the sparsity equation for the ℓ_1^d-norm. This equation has been solved in [39, Lemma 4.2], we recall this result.

Lemma 1. *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 1 and the choice of $r(\cdot)$ in (7) imply that for $\sigma = \|\xi\|_{L^{q_0}}$,

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

then ρ^* satisfies the sparsity equation and $r^2(\rho^*)$ is the rate of convergence of the LASSO for $\lambda \sim r^2(\rho^*)/\rho^* \sim \|\xi\|_{L_P^{q_0}} \sqrt{\log(ed/s)/N}$. This choice of λ requires to know the sparsity parameter s. That is the reason why we either need to choose a larger value for the $r(\cdot)$ function as in [36] – this results in the suboptimal $\sqrt{\log(ed)/N}$ rates of convergence from Theorem 3 – or to use an adaptation step as in Section 3.4.1 – this results in the better minimax rate $\sqrt{\log(ed/s)/N}$. Finally, one needs to compute the radii ρ_K and $\lambda \sim r^2(\rho_K)/\rho_K$. Let $K \in [N]$ and $\sigma = \|\xi\|_{L^{q_0}}$. The equation $K = cr(\rho_K)^2 N$ is solved by

$$\rho_K \sim_{L,q_0} \frac{K}{\sigma}\sqrt{\frac{1}{N}\log^{-1}\left(\frac{\sigma^2 d}{K}\right)} \tag{8}$$

for the $r(\cdot)$ function defined in (7). Therefore,

$$\lambda \sim \frac{r^2(\rho_K)}{\rho_K} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho_K\sqrt{N}}\right)} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma^2 d}{K}\right)} \ . \tag{9}$$

The regularization parameter depends on the $L_P^{q_0}$-norm of ξ. This parameter is unknown in practice. Nevertheless, it can be replaced by an estimator in the regularization parameter as in [23, Sections 5.4 and 5.6.2].

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

THEOREM 5. *Grant 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 $\|t^* - v\|_1 \leqslant c_2 \|\xi\|_{L^{q_0}} s \sqrt{\log(ed)/N}$. Assume that $|\mathcal{I}| \geqslant N/2$ and $|\mathcal{O}| \leqslant c_3 s \log(ed/s)$. The MOM-LASSO estimator $\hat{t}$ with the adaptively chosen number of blocks K (and λ) from Section 3.4.1 satisfies, with probability at least $1 - c_4 \exp(-c_5 s \log(ed/s))$, for every $1 \leqslant p \leqslant 2$,*

$$\left\|\hat{t} - t^*\right\|_p \leqslant c_6 \left\|\xi\right\|_{L_{q_0}} s^{1/p} \sqrt{\frac{1}{N} \log\left(\frac{ed}{s}\right)},$$

where $(c_j)_{j=1}^6$ depends only on θ_0, θ_m and q_0.

PROOF. It follows from Theorem 2, the computation of $r(\rho_K)$ from (7) and ρ_K in (8) 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 $\left\|v\right\|_p \leqslant \left\|v\right\|_1^{-1+2/p} \left\|v\right\|_2^{2-2/p}$ 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 [11], 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 (5) is polynomial – $1/N^{(q_0/2-1)}$ – whereas it is exponentially small for MOM LASSO. Exponential rates for LASSO hold only if ξ is subgaussian ($\left\|\xi\right\|_{L_p} \leqslant C\sqrt{p} \left\|\xi\right\|_{L_2}$ 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).
- 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 assume that the coordinates of X have $C_0 \log(ed)$ subgaussian moments and that X satisfies a L^2/L^1 equivalence assumption.

Let us now turn to the study of a "minmax MOM version" of the SLOPE estimator. The computation of the fixed point functions $r_Q(\cdot)$ and $r_M(\cdot)$ rely on [50, Theorem 1.6] and the computation from [36]. Again, the SLOPE norm has a dual norm which is 1-unconditional with respect to the canonical basis of $\mathbb{R}^d$, [50, Definition 1.4]. Therefore, it follows from [50, Theorem 1.6] that under Assumption 6, one has

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

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

where $\mathcal{B}$ is the unit ball of the SLOPE norm and $\zeta_i = Y_i - \left\langle X_i, t^*\right\rangle$. Local Gaussian mean widths $\ell^*(\rho\mathcal{B}\cap rB_2^d)$ are bounded from above in [39, 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 has been solved in [36, Lemma 4.3].

Lemma 2. *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 $\beta_j\leqslant C\sqrt{\log(ed/j)}$, $\mathcal{B}_s = \sum_{j\leqslant s}\beta_j/\sqrt{j}\lesssim C\sqrt{s\log(ed/s)}$. The condition $\mathcal{B}_s\lesssim\rho/r(\rho)$ holds when $N\gtrsim_{L,q_0} s\log(ed/s)$, $\rho\gtrsim_{L,q_0}\|\xi\|_{L_q}\frac{s}{\sqrt{N}}\log\left(\frac{ed}{s}\right)$. Lemma 2 implies that $\Delta(\rho)\geqslant 4\rho/5$ when there is an s-sparse vector in $t^* + (\rho/20)B_\Psi$. Therefore, Theorem 1 applies for $\lambda\sim r^2(\rho)/\rho\sim_{L,q,\delta}\|\xi\|_{L_q}/\sqrt{N}$.

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

The following result follows from Theorem 2 together with the previous computations of ρ^*,ρ_K, $r_Q(\cdot)$, $r_M(\cdot)$ and $r(\cdot)$. The proof, similar to Theorem 5, is omitted.

Theorem 6. *Grant Assumption 5. Let $s\in[d]$. Assume that $N\geqslant c_1 s\log(ed/s)$ and that there is $v\in\mathbb{R}^d$ such that $|\text{supp}(v)|\leqslant s$ and $\Psi(t^*-v)\leqslant c_2\|\xi\|_{L_q}s\log(ed/s)/\sqrt{N}$. Assume that $|\mathcal{I}|\geqslant N/2$ and $|\mathcal{O}|\leqslant c_3 s\log(ed/s)$. The MOM-SLOPE estimator $\hat{t}$ with the adaptive number of blocks K from Section 3.4.1 satisfies, with probability at least $1-c_4\exp(-c_5 s\log(ed/s))$,*

$$\left\|\hat{t}-t^*\right\|_2^2\leqslant c_6\left\|\xi\right\|_{L_{q_0}}^2\frac{s}{N}\log\left(\frac{ed}{s}\right),$$

where $(c_j)_{j=1}^6$ depends only on θ_0, θ_m and q_0.

MOM-SLOPE has the same advantages upon SLOPE as MOM-LASSO upon LASSO. These improvements, listed below Theorem 5 are not repeated. The only difference is that SLOPE, unlike LASSO, already achieves the minimax rate $s\log(ed/s)/N$ whereas, without an extra adaptation step as in [11], the LASSO is not known to achieve a rate better than $s\log(ed)/N$.

5. Algorithms for minmax MOM LASSO.

The aim of this section is to show that there is a systematic way to transform classical descent based algorithms (such as Newton or gradient descent algorithm, or proximal gradient descent algorithms, etc.) into robust ones using MOM approach. This section provides several examples of such modifications.

These algorithms are tested in high-dimensional frameworks. In this setup, there exists an important number of algorithms approximating LASSO. The aim of this section is to show that there is a natural modification of these algorithms that makes them more robust to outliers. The choice of hyperparameters like the number of blocks or the regularization parameter cannot be done via classical Cross-Validation (CV) because of possible outliers in the test sets. CV procedures are also adapted using MOM's principle in Section 6. We also advocate for using random blocks at every iterations of the algorithms, to bypasses a problem of "local saddle points" we have identified. A byproduct of the latter approach is a definition of depth adapted to the learning task and therefore of an outliers detection algorithm. This material and a simulation study are given in Section 6 of Supplement A.

5.1. *From algorithms for LASSO to MOM LASSO.*

Each algorithm designed for the LASSO can be transformed into a robust algorithm for the minmax MOM estimator. Recall that minmax MOM LASSO estimator is

$$\hat{t}_{K,\lambda} \in \operatorname*{argmin}_{t\in\mathbb{R}^d} \sup_{t'\in\mathbb{R}^d} T_{K,\lambda}(t', t) \tag{10}$$

where $T_{K,\lambda}(t',t) = \mathrm{MOM}_K\big(\ell_t - \ell_{t'}\big) + \lambda\left(\|t\|_1 - \|t'\|_1\right)$, $\mathrm{MOM}_K\big(\ell_t - \ell_{t'}\big)$ 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 $k \in [K]$,

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

A natural idea to implement (10) is to consider algorithms based on a sequence of alternating descents (in t) and ascents (in t') steps with possible proximal/projection steps and for various choices of step sizes. A key issue here is that $t \to T_{K,\lambda}(t'_0, t)$ (resp. $t' \to T_{K,\lambda}(t', t_0)$), for some given $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$, may not be convex (resp. concave). Nevertheless, one can still compute the steepest descent by assuming that the index in $[K]$ of the block achieving the median in $\mathrm{MOM}_K\big(\ell_{t_0} - \ell_{t'_0}\big)$ remains constant on a convex open set containing (t_0, t'_0), for almost all (t_0, t'_0). The median is set as the minimal value of the median interval.

Assumption 7. *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 7, for almost all couples $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$, $t \to T_{K,\lambda}(t'_0, t)$ is "locally convex" and $t' \to T_{K,\lambda}(t', t_0)$ is "locally concave". Therefore, for k such that $P_{B_k}(\ell_{t_0} - \ell_{t'_0}) \in \mathrm{MOM}_K\big(\ell_{t_0} - \ell_{t'_0}\big)$,

$$\nabla_t \mathrm{MOM}_K\big(\ell_t - \ell_{t'_0}\big)_{|t=t_0} = -2(X^{(k)})^\top(Y^{(k)} - X^{(k)} t_0) \tag{11}$$

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 (11) depends on t'_0 only through the index k.

Remark 3 (Block Gradient Descent). *Algorithms developed for the minmax estimator using steepest descent steps such as (11) are special instances of Block Gradient Descent (BGD). The major difference with standard BGD (which takes sequentially all blocks), is that the index of the block is chosen here as $P_{B_k}(\ell_{t_0} - \ell_{t'_0}) \in MOM_K\big(\ell_{t_0} - \ell_{t'_0}\big)$. In particular, we expect blocks corrupted by outliers to be avoided which is not the case in the classical BGD. Moreover, choosing the "descent / ascent" block k using its centrality, we also expect $P_{B_k}(\ell_{t_0} - \ell_{t'_0})$ to be close to the objective function $P(\ell_{t_0} - \ell_{t'_0})$. This should make every descent (resp. ascent) steps particularly efficient.*

Remark 4 (map-reduce). *The algorithms presented in this section particularly fits the map-reduce paradigm [19], where data are spread out in a cluster of servers and are therefore naturally split into blocks. Our procedures use for mapper a mean and for reducer a median. This makes our algorithms easily scalable into the big data framework even when some servers have crashed down (making blocks of outliers data). The median identifies* *the correct block of data onto which one should make a descent or an ascent and leaves aside servers which have crashed down.*

Remark 5 (Normalization). *In the 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 normalized to make ℓ_2^N-norms of the columns equal to one. In a corrupted setup, one row of $\mathbb{X}$ may be corrupted and normalizing each column of $\mathbb{X}$ would corrupt the entire matrix $\mathbb{X}$. We therefore do not normalize the design matrix in the following.*

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

$$t_{p+1} = t_p - \gamma_p \partial\psi(t_p) \tag{12}$$

where $\partial\psi(t_p)$ is a subgradient of ψ at t_p like $\partial\psi(t_p) = -2\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}t_p) + \lambda \mathrm{sign}(t_p)$ where $\mathrm{sign}(t_p)$ is the vector of signs of the coordinates of t_p with the convention $\mathrm{sign}(0) = 0$. The sub-gradient descent algorithm (12) can be turned into an alternating subgradient ascent/descent algorithm for the min-max estimator (10): let

$$\mathbb{Y}_k = (Y_i)_{i\in B_k} \text{ and } \mathbb{X}_k = (X_i^\top)_{i\in B_k} \in \mathbb{R}^{|B_k|\times d} \quad . \tag{13}$$

```
input  : $(t_0, t_0') \in \mathbb{R}^d \times \mathbb{R}^d$: initial point, $\epsilon > 0$: a stopping parameter,
         $(\eta_p)_p, (\beta_p)_p$: two step size sequences
output: approximated solution to the min-max problem (10)
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 $P_{B_k}(\ell_{t_p} - \ell_{t'_p}) = \mathrm{MOM}_K(\ell_{t_p} - \ell_{t'_p})$
    $$t_{p+1} = t_p + 2\eta_p \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t_p) - \lambda\eta_p \mathrm{sign}(t_p)$$
    find $k \in [K]$ such that $P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p}) = \mathrm{MOM}_K(\ell_{t_{p+1}} - \ell_{t'_p})$
    $$t'_{p+1} = t'_p + 2\beta_p \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t'_p) - \lambda\beta_p \mathrm{sign}(t'_p)$$
end
Return $(t_p, t'_p)$
```

Algorithm 1: A "minmax MOM version" of the sub-gradient descent.

The key insight in Algorithm 1 are steps 2 and 4 where the blocks number have been chosen by 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 (12) 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

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

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

Of course, the same choices of step size cannot be made for $(\eta_p)_p$ and $(\beta_p)_p$ in Algorithm 1 because $\mathbb{X}$ may be corrupted but it can be adapted. In the first case, one can take $\eta_p = 1/\left\|\mathbb{X}_k\right\|_{S_\infty}^2$ where $k \in [K]$ is the index defined in line 2 of Algorithm 1 and $\beta_p = 1/\left\|\mathbb{X}_k\right\|_{S_\infty}^2$ 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

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

where $\psi_k(t) = \|\mathbb{Y}_k - \mathbb{X}_k t\|_2^2 + \lambda \left\|t\right\|_1$ where $k \in [K]$ is defined in line 2 of Algorithm 1 and, for β_p, with $k \in [K]$ defined in line 4 of Algorithm 1.

5.3. *Proximal gradient descent algorithms.* This section provides MOM versions of ISTA (Iterative Shrinkage-Thresholding Algorithm) and its accelerated version FISTA. ISTA and FISTA are proximal gradient descent where the objective function $\psi(t) = f(t) + g(t)$ with $f(t) = \|\mathbb{Y} - \mathbb{X}t\|_2^2$ (convex and differentiable) and $g(t) = \lambda \left\|t\right\|_1$ (convex). ISTA alternates between a descent in the direction of the gradient of f and a projection through the proximal operator of g, which, for the ℓ_1-norm, is the soft-thresholding:

$$t_{p+1} = \operatorname{prox}_{\lambda\|\cdot\|_1}\left(t_p + 2\gamma_p \mathbb{X}^\top(\mathbb{Y} - \mathbb{X}t_p)\right) \tag{14}$$

where $\operatorname{prox}_{\lambda\|\cdot\|_1}(t) = (\operatorname{sign}(t_j)\max(|t_j| - \lambda, 0))_{j=1}^d$ for all $t = (t_j)_{j=1}^d \in \mathbb{R}^d$.

A natural "MOM version" for ISTA is given by the following alternating method where the step sizes sequences $(\eta_p)_p$ and $(\beta_p)_p$ may be chosen according to the remarks below Algorithm 1 or chosen a posteriori.

```
input  : $(t_0, t'_0) \in \mathbb{R}^d \times \mathbb{R}^d$: initial point, $\epsilon > 0$ : a stopping parameter,
         $(\eta_k)_k, (\beta_k)_k$: two step size sequences
output: approximated solution to the min-max problem (10)
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 $P_{B_k}(\ell_{t_p} - \ell_{t'_p}) = \mathrm{MOM}_K\big(\ell_{t_p} - \ell_{t'_p}\big)$
    $t_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\big(t_p + 2\eta_k \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t_p)\big)$
    find $k \in [K]$ such that $P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p}) = \mathrm{MOM}_K\big(\ell_{t_{p+1}} - \ell_{t'_p}\big)$
    $t'_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\big(t'_p + 2\beta_k \mathbb{X}_k^\top(\mathbb{Y}_k - \mathbb{X}_k t'_p)\big)$
end
Return $(t_p, t'_p)$
```

Algorithm 2: A "minmax MOM version" of ISTA.

5.4. *Douglas-Racheford / ADMM.* This section presents the Alternating Direction Method of Multipliers (ADMM) algorithm. It is also a splitting algorithm which reads as follows in the LASSO case: at step p,

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

where ρ is a tuning parameter. ADMM algorithm returns t_p after a stopping criteria is met. In Algorithm 3, 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 parameter,
         $\rho$: a parameter
output : approximated solution to the min-max problem (10)
while $\left\|t_{p+1} - t_p\right\|_2 \geqslant \epsilon$ or $\left\|t'_{p+1} - t'_p\right\|_2 \geqslant \epsilon$ do
    find $k \in [K]$ such that $P_{B_k}(\ell_{t_p} - \ell_{t'_p}) = \mathrm{MOM}_K\big(\ell_{t_p} - \ell_{t'_p}\big)$
        $t_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p - u_p)$
        $z_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\left(t_{p+1} + u_p/\rho\right)$
        $u_{p+1} = u_p + \rho(t_{p+1} - z_{p+1})$
    find $k \in [K]$ such that $P_{B_k}(\ell_{t_{p+1}} - \ell_{t'_p}) = \mathrm{MOM}_K\big(\ell_{t_{p+1}} - \ell_{t'_p}\big)$
        $t'_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z'_p - u'_p)$
        $z'_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}\left(t'_{p+1} + u'_p/\rho\right)$
        $u'_{p+1} = u'_p + \rho(t'_{p+1} - z'_{p+1})$
end
Return $(t_p, t'_p)$
```

Algorithm 3: A "minmax MOM version" of ADMM

6. Simulations study.

This section provides an extensive simulation study based on algorithms of Section 5. In particular, their robustness and their convergence properties are illustrated on simulated data. The algorithms depend on hyper-parameters that need to be tuned. Due to possible corruption, classical approaches relying on test samples can't be trusted. The section starts therefore by introducing a robust CV procedure based on MOM principle.

6.1. *Adaptive choice of hyper-parameters via MOM V-fold CV.*

MOM's principles can be combined with the idea of multiple splitting into training / test datasets in cross-validation.

Let $V \in [N]$ be such that N can be divided by V. Let also $\mathcal{G}_K \subset [N]$ and $\mathcal{G}_\lambda \subset (0, 1]$. The aim is to select an optimal number of blocks and regularization parameter within both grids. The dataset is split 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

$$\left(\hat{f}^{(v)}_{K,\lambda} : K \in \mathcal{G}_K, \lambda \in \mathcal{G}_\lambda\right) . \tag{15}$$

The remaining $\mathcal{D}_v$ of the dataset is used to test the performance of each estimator in the family (15). Using these notations, we define a MOM version of the cross-validation procedure.

Definition 6. *The* ***Median of Means*** V***-fold CV*** *associated to the estimators (15) is* $\hat{f}_{\hat{K},\hat{\lambda}}$ *where* $(\hat{K},\hat{\lambda})$ *is a minimizer of*

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

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

$$\mathrm{MOM}^{(v)}_{K'}(\ell_f) = \mathrm{MOM}_{K'}\left(P_{B^{(v)}_1}\ell_f, \cdots, P_{B^{(v)}_{K'}}\ell_f\right) \tag{16}$$

and $B^{(v)}_1 \cup \cdots, \cup B^{(v)}_{K'}$ *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 with standard V-fold CV is that empirical means in classical V-fold CV are replaced by MOM estimators in (16). 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, 5]. In the simulations, we use $V = 5$. The construction of MOM-CV requires to choose another parameter: K', the number of blocks used to build MOM criteria (16) over the test set. One can choose $K' = K/V$ to make only one split of $\mathcal{D}$ into K blocks and use, for each round, $(V-1)K/V$ blocks to build estimators (15) and K/V blocks to test them.

In Figures 2, hyper-parameters K (i.e. the number of blocks) and λ (i.e. the regularization parameter) have been chosen for MOM LASSO estimators via MOM V-fold CV. Only the evolution of $\hat{K}$ in function of the proportion of outliers has been depicted (the choice of the adaptively chosen regularization parameter is more erratic and may first require a more deeper understanding of CV in the classical i.i.d. before the study of MOM CV in the $\mathcal{O} \cup \mathcal{I}$ framework). The adaptive $\hat{K}$ grows with the number of outliers as expected, since the number of blocks has to be at least twice the number of outliers. In particular, when there are no outliers in the dataset, MOMCV selects $K = 1$ so minmax MOM LASSO is the LASSO. The algorithm learns that splitting the database is useless in the absence of outliers: LASSO is the best choice among all minmax MOM LASSO estimators for $K \in [N/2]$.

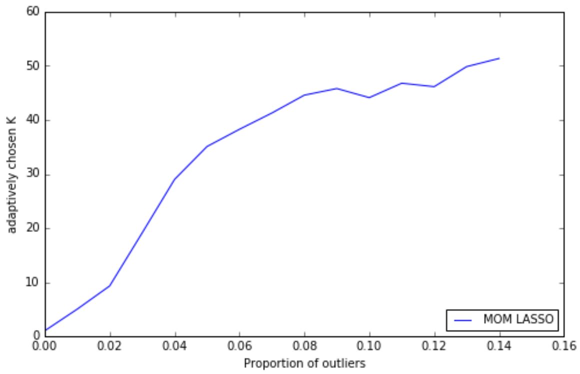

FIG 2. *Adaptively chosen number of blocks K for the minmax MOM LASSO.*

REMARK 6. *Median of Means V-fold CV introduced in Definition 6 aims at testing the performance of estimators on a possibly corrupted test set. This is done by excluding outliers from the test set thanks to the median operator. However, there are situations, for instance in image recognition, where the test set is corrupted but still we expect estimators to perform well even on these corrupted data in the test set. This is a classical robustness issue in Deep Learning [68]. Indeed, deep learning methods are known to fail if a small Gaussian noise is added to images even with a small variance undetectable by human eyes. Even though minmax MOM estimators introduced in this paper have been initially designed to be robust to outliers in the train set, one can use classical tricks to be also robust to corruption in the test set by training minmax MOM estimators onto an augmented database: in practice, given a (clean or not) dataset $(X_i, Y_i)_{i=1}^N$, one can construct an augmented dataset where each data (X_i, Y_i) is replicated m times with an added Gaussian noise: $(X_i + Z_{i1}, Y_i), \cdots, (X_i + Z_{im}, Y_i)$ – where $(Z_{ij} : 1 \leq i \leq N, 1 \leq j \leq m)$ are i.i.d. Gaussian variable – and then a minmax MOM estimator can be trained onto the dataset*

$$(X_i + Z_{ij}, Y_i), i = 1, \dots, n, j = 1, \dots, m.$$

By doing so, we expect the minmax MOM estimator to improve its robustness performance evaluated on a corrupted test set.

6.2. *Saddle-point, random blocks, outliers detection and depth.* The aim of this section is to show some advantages of choosing randomly the blocks at every (descent and ascent) steps of the algorithm and how this modified version works on the example of ADMM. As a byproduct, it is possible to define an outliers detection algorithm.

Let us first explain a problem of "local saddle point" in the case of fixed blocks. Minmax MOM estimators are based on the observation that the oracle f^* is solution to the minmax problem $f^* \in \text{argmin}_{f\in F} \sup_{g\in F} P(\ell_f - \ell_g)$. Likewise, f^* is solution of the maxmin problem: $f^* \in \text{argmax}_{g\in F} \inf_{f\in F} P(\ell_f - \ell_g)$. One can also define the maxmin MOM estimator

$$\hat{g}_{K,\lambda} \in \underset{g\in F}{\text{argmax}} \inf_{f\in F} T_{K,\lambda}(g,f). \tag{17}$$

Following the proofs of Section 6 in the supplement, one can prove the same results for $\hat{g}_{K,\lambda}$ and $\hat{f}_{K,\lambda}$ (see Section 7 in the supplement for a proof in small dimension). However, $\hat{g}_{K,\lambda}$ and $\hat{f}_{K,\lambda}$ may differ since, in general

$$\underset{f\in F}{\text{argmin}} \sup_{g\in F} T_{K,\lambda}(g,f) \neq \underset{g\in F}{\text{argmax}} \inf_{f\in F} T_{K,\lambda}(g,f). \tag{18}$$

In other words, the duality gap may not be null. Since $T_{K,\lambda}(g,f) = -T_{K,\lambda}(f,g)$ for all $f,g \in F$, (18) 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 a **saddle-point** estimator and minmax and maxmin estimators are equal. The left-hand side of Figure 3 shows a simulation where this happens. The choice of fixed blocks $B_1, \dots, B_K$ may result in a problem of "**local saddle points**" and the algorithms remain close to suboptimal local saddle points. To see this, consider the vector case (that is for $F = \{f(\cdot) = \langle \cdot, t\rangle : t \in \mathbb{R}^d\}$ and introduce, for all $k \in [K]$,

$$\mathcal{C}_k = \left\{(t,t') \in \mathbb{R}^d \times \mathbb{R}^d : \text{MOM}_K\big(\ell_t - \ell_{t'}\big) = P_{B_k}(\ell_t - \ell_{t'})\right\}. \tag{19}$$

The problem is that, if a cell $\mathcal{C}_k$ contains a saddle-point of $(t,t') \to P_{B_k}(\ell_t - \ell_{t'}) + \lambda(\|t\|_1 - \|t'\|_1)$ the algorithms gets stuck in that cell instead of looking for "better saddle-point" in other cells.

To overcome this issue, the partition is chosen at random at every descent and ascent steps of the algorithms, so the decomposition into cells $\mathcal{C}_1, \cdots, \mathcal{C}_K$ changes at every steps. As an example, we develop the ADMM procedure with a random choice of blocks in Algorithm 4.

```
input  : $(t_0, t_0') \in \mathbb{R}^d \times \mathbb{R}^d$: initial point, $\epsilon > 0$: a stopping parameter,
         $\rho$: parameter
output: approximated solution to the min-max problem (10)
while $\|t_{p+1} - t_p\|_2 \geqslant \epsilon$ or $\|t_{p+1}' - t_p'\|_2 \geqslant \epsilon$ do
    Build an equipartition $B_1, \dots, B_K$ of $[N]$ at random.
    Find $k \in [K]$ such that $P_{B_k}(\ell_{t_p} - \ell_{t_p'}) = \mathrm{MOM}_K\big(\ell_{t_p} - \ell_{t_p'}\big)$
    $t_{p+1} = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p - u_p)$
    $z_{p+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}(t_{p+1} + u_p/\rho)$
    $u_{p+1} = u_p + \rho(t_{p+1} - z_{p+1})$
    Build an equipartition $B_1, \dots, B_K$ of $[N]$ at random.
    Find $k \in [K]$ such that $P_{B_k}(\ell_{t_{p+1}} - \ell_{t_p'}) = \mathrm{MOM}_K\big(\ell_{t_{p+1}} - \ell_{t_p'}\big)$
    $t_{p+1}' = (\mathbb{X}_k^\top \mathbb{X}_k + \rho I_{d\times d})^{-1}(\mathbb{X}_k^\top \mathbb{Y}_k + \rho z_p' - u_p')$
    $z_{p+1}' = \mathrm{prox}_{\lambda\|\cdot\|_1}\big(t_{p+1}' + u_p'/\rho\big)$
    $u_{p+1}' = u_p' + \rho(t_{p+1}' - z_{p+1}')$
end
Return $(t_p, t_p')$
```

Algorithm 4: minmax MOM ADMM with changing random blocks.

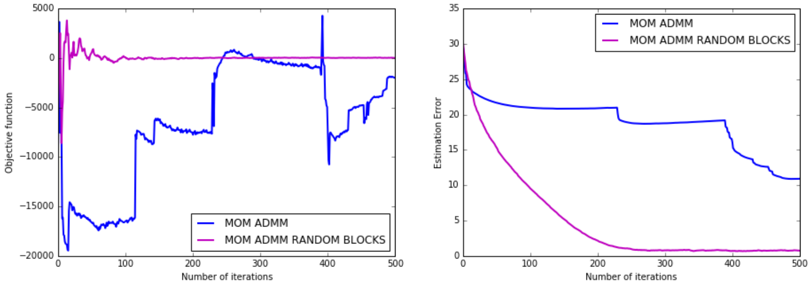

FIG 3. *Fixed blocks against random blocks.*

In Figure 3, both MOM LASSO via ADMM with fixed and changing blocks are run. Both the objective function and the estimation error of MOM LASSO jump with fixed blocks. These jumps correspond to a change of cell number. The algorithm converges to local saddle-points before jumping to other cells, thanks to the regularization of the ℓ_1-norm. On the other hand, the algorithms with changing blocks do not suffer this drawback. Figure 3 shows that the estimation error converges faster and more smoothly for changing blocks. The objective function of MOM ADMM with changing blocks converges to zero so the duality gap converges to zero. This gives a natural stopping criterion and shows that minmax and maxmin MOM LASSO are solution of a saddle point problem even though the objective function is not convex-concave.

A byproduct is an **outliers detection procedure**. Count the number of times each data is selected in the selected median blocks of steps 3 and 7 of Algorithm 4. At the end of the algorithm (for instance, Algorithm 4), every data ends up with a score revealing its centrality for the learning task. Aggressive outliers are likely to corrupt their respective blocks and should therefore not be selected at steps 3 and 7 of Algorithm 4. With 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 changing at every steps. Figure 4 shows performance of this strategy on synthetic data (cf. Section 6.3 for more details on the simulations). Outliers (data $1, 32, 170$ and 194) end up with a null score.

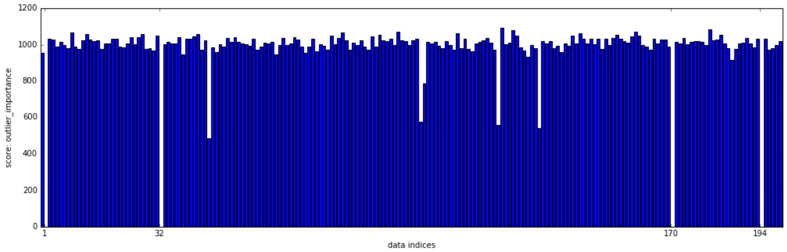

FIG 4. *Outliers detection algorithm. The dataset has been corrupted by* 4 *outliers at number* $1, 32, 170$ *and* 194*. The score of the outliers is* 0*: they haven't been selected even once.*

6.3. *Simulations setup for the figures.* All codes are available at [1] and can be used to reproduce the figures. Many other simulations and algorithms can be found in [1].

6.3.1. *Data generating process and corruption by outliers.* The algorithms introduced in Section 5 are tested on datasets corrupted by outliers of various forms in [1]. The basic set of informative data is called $\mathcal{D}_1$. The outliers are named $\mathcal{D}_2, \mathcal{D}_3$, $\mathcal{D}_4$ and $\mathcal{D}_5$. These data are merged and shuffled in the dataset $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3 \cup \mathcal{D}_4 \cup \mathcal{D}_5$ given to the algorithm.

1. The set $\mathcal{D}_1$ of inliers contains N_{good} i.i.d. data (X_i, Y_i) with common distribution

$$Y = \left\langle X, t^* \right\rangle + \xi \quad , \tag{20}$$

where $t^* \in \mathbb{R}^d$, $X \sim \mathcal{N}(0, I_{d\times d})$ and $\xi \sim \mathcal{N}(0, \sigma^2)$ is independent of X.

2. $\mathcal{D}_2$ is a dataset of N_{bad-1} outliers (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} outliers (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} outliers (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 (20), with the same target vector t^* and a different choice of design X and noise ξ. the design $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = (\rho^{|i-j|})_{1\leqslant i,j\leqslant d}$ and ξ is a heavy-tailed noise distributed according to a Student distribution with various degrees of freedom.

The different types of outliers $\mathcal{D}_j, j = 2, 3, 4, 5$ are useless to learn the oracle t^* some are not independent nor random as in $\mathcal{D}_2$ and $\mathcal{D}_3$.

6.3.2. *Simulations setup for the figures.* Let us now precise the parameters of the simulations in Figure 1 and Figure 2: the number of observations is $N = 200$, the number of features is $d = 500$, $t^* \in \mathbb{R}^d$ has sparsity $s = 10$ and support chosen at random, with non-zero coordinates t_j^* being either equal to 10, -10 or decreasing according to $\exp(-j/10)$. Informative data $\mathcal{D}_1$, described in Section 6.3.1, have variance $\sigma = 1$. This dataset is increasingly corrupted with outliers in $\mathcal{D}_3$.

The proportion of outliers are $0, 1/100, 2/100, \dots, 15/100$. ADMM algorithm is run with adaptive λ chosen by V-fold CV with $V = 5$ for the LASSO. Then MOM ADMM is run with adaptive K and λ chosen by MOM CV 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 is 200. Simulations have been run 70 times and the averaged values of the estimation error and adaptive $\hat{K}$ have been reported in Figure 1, Figure 5 and Figure 2. The ℓ_2 estimation error of LASSO increases roughly from 0 when there is no outliers and stabilize at 550 right after a single outliers enters the dataset. The value 550 comes from the fact that $Y = 10000$ and $X = (1)_{j=1}^{500}$ satisfy that the vector t with minimal ℓ_1^d norm among all the solutions t of $Y = \langle X, t\rangle$ is $t^{**} = (20)_{j=1}^{500}$, and $\|t^{**} - t^*\|_2$ is approximately 550. This means that LASSO is trying to fit a model on the single 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 3, we have run similar experiments with $N = 200$, $d = 300$, $s =$

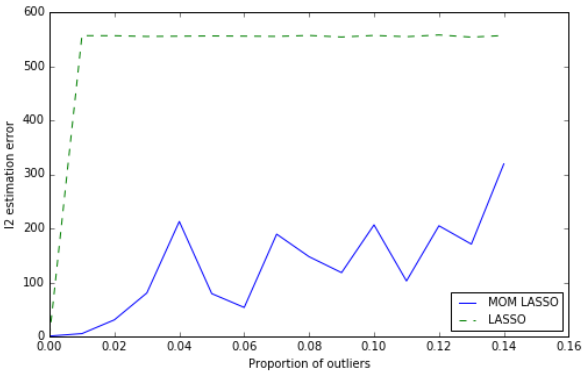

FIG 5. *Estimation error versus proportion of outliers for LASSO and the minmax MOM LASSO.*

20, $\sigma = 1$, $K = 10$, the number of iterations was 500 and the regularization parameter was $1/\sqrt{N}$.

For Figure 4, 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 $\lambda = 1/\sqrt{200}$.

SUPPLEMENTARY MATERIAL

Supplement A: Supplementary material to "Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions"
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). Section 6 gives the proof of the main results. These main results focus on regularized version of the MOM estimates of the increments presented in this introduction that are well suited for high dimensional learning frameworks, we complete these results in Section 7, providing results for the basic estimators without regularization in small dimension. Finally, Section 8 provides minimax optimality results for our procedures.

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 Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.*, 21(4):613–632, 2011.

[5] 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.

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

[7] 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.

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

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

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

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

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

[13] Lucien Birgé. Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.*, 20(3):201–223, 1984.

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

[15] Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36(2):489–531, 2008.

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

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

[19] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.

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

[21] Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *To appear in the Annals of Statistics*, 2016.

[22] 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.

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

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

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

[26] Qiyang Han and Jon A. Wellner. A sharp multiplier inequality with applications to heavy-tailed regression problems. *arXiv:1706.02410*, 2017.

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

[28] 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.

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

[30] 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.

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

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

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

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

[35] G. Lecué and M. Lerasle. Learning from mom's principle : Le cam's approach. Technical report, CNRS, ENSAE, Paris-sud, 2017. To appear in Stochastic Processes and their applications.

[36] 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.

[37] 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.

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

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

[40] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv:1112.3914*, 2011.

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

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

[43] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in JEMS.*

[44] Gabor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *To appear in Ann. Statist. arXiv:1702.00482.*

[45] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[46] Mathurin Massias, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gener-

alized concomitant multi-task lasso for sparse multimodal regression. 2017.

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

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

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

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

[51] S. Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. *Preprint available on arXiv:1704.02658.*

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

[53] 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.

[54] 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.

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

[56] Alquier Pierre, Cottet Vincent, and Lecué Guillaume. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *arXiv preprint arXiv:1702.01402*, 2017.

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

[58] Adrien Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018.

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

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

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

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

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

[64] 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.

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

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

[67] 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.

[68] Timothy E. Wang, Yiming Gu, Dhagash Mehta, Xiaojun Zhao, and Edgar A. Bernal.

Towards robust deep neural networks. *arXiv preprint arXiv:1810.11726*, 2018.

[69] 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.

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

ENSAE
5 avenue Henry Le Chatelier
91120 Palaiseau, France
E-mail: guillaume.lecue@ensae.fr
URL: http://lecueguillaume.github.io

University Paris Sud Orsay
Mathematics department
91405 Orsay
E-mail: matthieu.lerasle@math.u-psud.fr
URL: http://lerasle.perso.math.cnrs.fr