Robust high dimensional learning for Lipschitz and convex losses.

Chinot Geoffrey GEOFFREY.CHINOT@ENSAE.FR
Department of Statistics
ENSAE CREST
5 avenue Henry Le Chatelier 91120 Palaiseau, France

Lecué Guillaume GUILLAUME.LECUE@ENSAE.FR
Department of Statistics
ENSAE CREST
5 avenue Henry Le Chatelier 91120 Palaiseau, France

Lerasle Matthieu LERASLE.MATTHIEU@MATH.U-PSUD.FR
Department of Mathematics
Université Paris Sud
91405 Orsay, France

Editor: ..

Abstract

We establish risk bounds for Regularized Empirical Risk Minimizers (RERM) when the loss is Lipschitz and convex and the regularization function is a norm. We obtain these results in the i.i.d. setup under subgaussian assumptions on the design. In a second part, a more general framework where the design might have heavier tails and data may be corrupted by outliers both in the design and the response variables is considered. In this situation, RERM performs poorly in general. We analyse an alternative procedure based on median-of-means principles and called "minmax MOM". We show optimal subgaussian deviation rates for these estimators in the relaxed setting. The main results are meta-theorems allowing a wide-range of applications to various problems in learning theory. To show a non-exhaustive sample of these potential applications, it is applied to classification problems with logistic loss functions regularized by LASSO and SLOPE, to regression problems with Huber loss regularized by Group LASSO, Total Variation and Fused LASSO and to matrix completion problems with quantile loss regularized by the nuclear norm. A short simulation study concludes the paper, illustrating in particular robustness properties of regularized minmax MOM procedures.

Keywords: Robust Learning, Lipschtiz and convex loss functions, sparsity bounds, Rademacher complexity bounds, LASSO, SLOPE, Group LASSO, Total Variation, Fused LASSO, Nuclear norm matrix completion.

1. Introduction

Regularized empirical risk minimizers (RERM) are standard estimators for high dimensional classification and regression where the estimator is a minimizer of a regularized empirical mean of loss functions. In regression, the quadratic loss of linear functionals regularized by the ℓ_1-norm (LASSO) Tibshirani (1996) is probably the most famous example of RERM, see for example Koltchinskii (2011); Bühlmann and van de Geer (2011) for overviews. Recent results and references, including more general regularization functions can be found, for example in Lecué and

Mendelson (2018); Bellec et al. (2017). RERM based on quadratic loss are highly unstable when data have heavy-tails or when the dataset has been corrupted by even a few outliers. These problems have attracted a lot of attention in robust statistics, see for example Huber and Ronchetti (2011) for an overview. By considering alternative losses, one can efficiently solve these problems when heavy-tails or corruption happen in the output variable Y. There is a growing literature analyzing performance of some of these alternatives in learning theory. In regression problems, among others, one can mention the L_1 absolute loss Shalev-Shwartz and Tewari (2011), the Huber loss Zhou et al. (2018); Elsener and van de Geer (2018) and the quantile loss that is popular in finance and econometrics. In classification, besides the $0/1$ loss function which is known to lead to computationally untractable RERM, the logistic loss and the hinge loss are among the most popular convex surrogates. Quantile, L_1, Huber loss functions for regression and Logistic, Hinge loss functions for classification are all Lipschitz and convex loss functions (in their first variable, see Assumption 2 for a formal definition). This remark motivated Alquier et al. (2017) to study systematically RERM based on Lipschitz loss functions. A remarkable feature of Lipschitz losses proved in Alquier et al. (2017) is that optimal results can be proved with almost no assumption on the response variable Y.

This paper is built on the approach initiated in Chinot et al. (2018). Compared with Alquier et al. (2017), the approach of Chinot et al. (2018) improves the results by deriving risk bounds depending on a localized complexity parameters rather than global ones and by considering a more flexible setting where a global Bernstein condition is relaxed into a local one, see Assumption 5 and the following discussion for details. The paper Chinot et al. (2018) only considers estimators that are not regularized and that can therefore only be efficient in small dimensional settings. The first main result of this paper is a high dimensional extension of these results that is achieved by analysing estimators regularized by sparsity inducing norms Bach et al. (2012). The main result is a meta-theorem allowing to study simultaneously a broad range of applications including in particular LASSO, SLOPE, group LASSO, total variation, fused LASSO or Shatten S_1 norm. Section 6 provides applications of the main results to some examples among these.

While RERM is studied without assumption on the output variables, somehow strong, albeit classical, hypotheses are granted on the design X in our first main result. We assume actually in this analysis subgaussian assumptions on the input variables as in Alquier et al. (2017). The necessity of this assumption to derive optimal exponential deviation bounds for RERM is not surprising as RERM have downgraded performance when the design is heavy tailed (see Mendelson (2014) for instance).

In a second part of this paper, we study an alternative to RERM in a framework with less stringent assumptions on the design. These estimators, based on Median-Of-Means (MOM) Nemirovsky and Yudin (1983); Birgé (1984); Jerrum et al. (1986); Alon et al. (1999) and minmax principles Audibert and Catoni (2011); Baraud et al. (2017) are called minmax MOM estimators as in Lecué and Lerasle (2019). A non-regularized version of these estimators was analysed in Chinot et al. (2018). The second main and most important result of the paper shows that minmax MOM estimators achieve optimal subgaussian deviation bounds in the relaxed setting where RERM perform poorly. This result is obtained under a local Bernstein condition as for the RERM. It allows to derive fast rates of convergence in a large set of applications where typically, subgaussian assumptions on the design X are replaced by moment assumptions. Minmax MOM estimators are then analysed without the local Bernstein condition. Oracle inequalities holding with exponentially large probability are proved in this case. Compared with results under Bernstein's assumption, an extra variance term appears in the convergence rate. This extra term typically would yield to slow rates of convergence in the applications, which are known to be minimax in this case. However, the variance term disappears under the Bernstein's condition, which shows that fast rates can be recovered from the general results. In addition, all results on minmax MOM estimators, both with or without Bernstein condition, are shown in the "$\mathcal{O} \cup \mathcal{I}$" framework – where $\mathcal{O}$ stands for "outliers" and $\mathcal{I}$ for "informative"– see Section 4.1 or Lecué and Lerasle (2017, 2019) for details. In this framework, all assumptions (such as the Bernstein's condition) are granted on "inliers" $(X_i, Y_i)_{i \in \mathcal{I}}$. These inliers may have different distributions but the oracles of these distributions should match. On the other hand, no assumption are granted on outliers $(X_i, Y_i)_{i \in \mathcal{O}}$, which is to the best of our knowledge the strongest form of aggressive outliers. The minmax MOM estimators perform well in this setting, it means that the accuracy of their predictions is not downgraded by the presence of a few outliers in the dataset. Mathematically, this robustness is not surprising as it is a byproduct of the median step. However, in practice, it is an important advantage of MOM estimators compared to RERM.

The main results on minmax MOM estimators are also meta-theorems that can be applied to the same examples as RERM. Each of these examples provide a new (to the best of our knowledge) estimator that reach performance that RERM could not typically. For example, when the class of classifiers/regressors is the class of linear functions on $\mathbb{R}^p$, minmax MOM estimators have a risk bounded by the minimax rate with optimal exponential probability of deviation even if the inputs X only satisfy weak moment assumptions and/or have been corrupted by outliers. These applications are also discussed in Section 6. Finally Section 7 shows heuristics to implement minmax MOM estimators and presents some modifications of these estimators that seem to perform much better in practice. Take-home messages from these simulations are 1) there exist algorithms approximating minmax MOM estimators and so they should not be considered as pure mathematical constructions 2) robustness properties of minmax-MOM estimators are confirmed on practical examples 3) almost any classical algorithm used in high-dimensional statistics can be easily turned into a robust one using a MOM approach.

The remaining of paper is decomposed as follows. Section 2 presents the formal setting. Section 3 presents results for RERM and Section 4 those for minmax MOM estimators with a local Bernstein condition and in Section 5 without this condition. Section 6 details several examples of applications of the main results. A short simulation study illustrating our theoretical findings is presented in Section 7. The proofs are postponed to Sections 9- 11.

2. Mathematical background and notations

Let $(\mathcal{Z}, \mathcal{A}, P)$ denote a probability space, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is a product space such that $\mathcal{X}$ denotes a measurable space of inputs and $\mathcal{Y} \subset \mathbb{R}$ is the subset of outputs. Let $Z = (X, Y)$ denote a random variable taking values in $\mathcal{Z}$ with distribution P and let μ denote the marginal distribution of the design X.

Let $\overline{\mathcal{Y}} \subset \mathbb{R}$ denote a convex set and let F denote a class of functions $f : \mathcal{X} \to \overline{\mathcal{Y}}$. The set $\overline{\mathcal{Y}}$ typically contains all possible values of the predictions $f(x)$ of y, for $f \in F$ and $x \in \mathcal{X}$. As such, it will always contain $\mathcal{Y}$. Let $\ell : \overline{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ denote a loss function such that $\ell(f(x), y)$ measures the error made when predicting y by $f(x)$. For any distribution Q on $\mathcal{Z}$ and any function $g : \mathcal{Z} \to \mathbb{R}$ for which it makes sense, let $Qg = \mathbb{E}_{Z \sim Q}[g(Z)]$ denote the expectation of the function g under the distribution Q and, for any $p \geqslant 1$, let $\|g\|_{L_p(Q)} := (Q[|g|^p])^{1/p}$, $\|g\|_{L_p} := \|g\|_{L_p(P)}$. The risk of any $f \in F$ is given by $P\ell_f$, where $\ell_f(x, y) := \ell(f(x), y)$. The prediction of Y with minimal risk is given by $f^*(X)$, where f^*, called *oracle*, is defined as any function such that

$$f^* \in \underset{f\in F}{\operatorname{argmin}}\, P\ell_f \quad .$$

Hereafter, for simplicity, it is assumed that f^* exists and is uniquely defined. The oracle is unknown to the statistician that has only access to a dataset $(X_i, Y_i)_{i\in\{1,\dots,N\}}$ of random variables taking values in $\mathcal{X}\times\mathcal{Y}$. The goal is to build a data-driven estimator $\hat{f}$ of f^* that predicts almost as well as f^*. The quality of an estimator $\hat{f}$ is measured by the error rate $\|\hat{f}-f^*\|_{L_2}^2$ and the excess risk $P\mathcal{L}_{\hat{f}}$, where, respectively,

$$\|\hat{f}-f^*\|_{L_2}^2 = P[(\hat{f}-f)^2] = \mathbb{E}\left[\left(\hat{f}(X)-f^*(X)\right)^2 |(X_i,Y_i)_{i=1}^N\right] \text{ and } \mathcal{L}_{\hat{f}} := \ell_{\hat{f}} - \ell_{f^*} \quad . \tag{1}$$

Let P_N denote the empirical measure i.e $P_N(A) = (1/N)\sum_{i=1}^N I(X_i \in A)$. A natural candidate for the estimation of f^* is the Empirical Risk Minimizer (ERM) of Vapnik and Červonenkis (1971), see also Vapnik (1998) for an overview, which is defined by

$$\hat{f}^{ERM} \in \underset{f\in F}{\operatorname{argmin}}\, P_N\ell_f \quad . \tag{2}$$

The choice of F is a central issue: enlarging the space F deteriorates the quality of the oracle estimation but improves the prediction of this oracle. It is possible to use large classes F without significantly altering the quality estimation if certain structural properties are granted on the oracle and known to the statistician. In that case, a widely spread approach is to add to the empirical loss a regularization term promoting this structural property. In this paper, we consider this problem when the regularization term is a norm. Formally, let E be a linear space such that $F \subset E \subset L_2(\mu)$ and let $\|\cdot\| : E \mapsto \mathbb{R}^+$ denote a norm on E. For any $\lambda \geq 0$, the regularized ERM (RERM) is defined by

$$\hat{f}_\lambda^{RERM} \in \underset{f\in F}{\operatorname{argmin}}\, P_N\ell_f^\lambda, \quad \text{where} \quad \ell_f^\lambda(x,y) = \ell_f(x,y) + \lambda\|f\| \quad . \tag{3}$$

In regression, one can mention Thikonov regularization which promotes smoothness Golub et al. (1999) and ℓ_1 regularization which promotes sparsity Tibshirani (1996). Likewise, for matrix reconstruction, the 1-Schatten norm S_1 promotes low rank solutions (see Koltchinskii et al. (2011); Cai et al. (2016)).

In the remaining of the paper, the following notations will be used repeatedly: for any $r > 0$, let

$$rB_{L_2} = \{f \in L_2(\mu) : \|f\|_{L_2} \leqslant r\}, \quad rS_{L_2} = \{f \in L_2(\mu) : \|f\|_{L_2} = r\} \quad .$$

Let $rB = \{f \in E : \|f\| \leq r\}$ and $rS = \{f \in E : \|f\| = r\}$. For any set H for which it makes sense, let $H + f^* = \{h + f^* \text{ s.t } h \in H\}$, $H - f^* = \{h - f^* \text{ s.t } h \in H\}$. Let $(e_i)_{i=1}^p$ be the canonical basis of $\mathbb{R}^p$. Let c denote an absolute constant whose value might change from line to line and let $c(A)$ denote a function depending on the parameters A whose value may also change from line to line.

3. Regularized ERM with Lipschitz and convex loss functions

This section presents and improves results from Alquier et al. (2017). A local Bernstein assumption, holding in a neighborhood of the *oracle* f^* is introduced in the spirit of Chinot et al. (2018). This assumption does not imply boundedness of F in L^2-norm unlike the global Bernstein condition considered in Alquier et al. (2017). New rates of convergence are obtained, depending on **localized** complexity parameters improving the global ones from Alquier et al. (2017).

3.1 Main assumptions

We start with a set of assumptions sufficient to prove exponential deviation bounds for the error rate and excess risk of RERM for general convex and Lipschitz loss functions and for any regularization norm. In this section, we consider the classical i.i.d. assumption.

Assumption 1 $(X_i, Y_i)_{i=1}^N$ *are independent and identically distributed with distribution* P.

All along the paper, we consider Lipschitz and convex loss functions.

Assumption 2 *There exists* $L > 0$ *such that, for any* $y \in \mathcal{Y}$, $\ell(\cdot, y)$ *is* L-***Lipschitz*** *i.e for every* f *and* g *in* F, $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$, $|\ell(f(x), y) - \ell(g(x), y)| \leq L|f(x) - g(x)|$ *and* ***convex*** *i.e for all* $\alpha \in [0, 1]$, $\ell(\alpha f(x) + (1-\alpha)g(x), y) \leq \alpha\ell(f(x), y) + (1-\alpha)\ell(g(x), y)$.

There are many examples of loss functions satisfying Assumption 2. The three examples studied in this work (see Section 6) are

- the **logistic loss function** defined for any $u \in \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\ell(u, y) = \log(1 + \exp(-yu))$. It satisfies Assumption 2 for $L = 1$.
- Tte **Huber loss function** with parameter $\delta > 0$ is defined for all $u, y \in \mathbb{R}$, by

$$\ell(u, y) = \begin{cases} \frac{1}{2}(y-u)^2 & \text{if } |u-y| \leq \delta \\ \delta|y-u| - \frac{\delta^2}{2} & \text{if } |u-y| > \delta \end{cases} .$$

  It satisfies Assumption 2 for $L = \delta$.
- The **quantile loss function** with parameter $\tau \in (0, 1)$ is defined for all $u, y \in \mathbb{R}$, by $\ell(u, y) = \rho_\tau(u - y)$ where, for any $z \in \mathbb{R}$, $\rho_\tau(z) = z(\tau - I\{z \leq 0\})$. It satisfies Assumption 2 with $L = \max(\tau, 1-\tau)$. For $\tau = 1/2$, the quantile loss is called the L_1 (or absolute) loss function.

Assumption 3 *The class* F *is convex.*

In particular, Assumption 3 holds in the important case considered in high-dimensional statistics when F is the class of all linear functions indexed by $\mathbb{R}^p$, $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^p\}$. This example is studied in great details in Section 6 as well as the case of linear functions indexed by the set of matrices $\mathbb{R}^{m \times T}$ to handle problems such as matrix completion.

RERM performs well when the empirical excess risk $f \in F \to P_N\mathcal{L}_f$ is uniformly concentrated around the excess risk $f \in F \to P\mathcal{L}_f$. This requires strong concentration properties of the class of random variables $\{\mathcal{L}_f(X) : f \in F\}$, which is implied by concentration properties of $\{(f - f^*)(X) : f \in F\}$ under the Lipschitz assumption. Here, we study RERM under a subgaussian assumption on the design. We first recall the definition of a subgaussian class of functions.

Definition 1 *A class F is called L_0-subgaussian (with respect to X), where $L_0 \geq 1$, when for all f in F and for all $\lambda > 1$, $\mathbb{E} \exp(\lambda |f(X)|/\|f\|_{L_2}) \leq \exp(\lambda^2 L_0^2/2)$.*

Assumption 4 *The class $F - f^*$ is L_0-subgaussian with respect to X.*

Assumptions 1-4 are also granted in Alquier et al. (2017). In this setup, a natural way to measure the statistical complexity of the problem is via Gaussian mean widths. We recall the definition of this measure of complexity.

Definition 2 *Let $H \subset L_2(\mu)$ and $(G_h)_{h \in H}$ be the canonical centered Gaussian process indexed by H, with covariance structure given by $\left(\mathbb{E}(G_{h_1} - G_{h_2})^2\right)^{1/2} = \left(\mathbb{E}(h_1(X) - h_2(X))^2\right)^{1/2}$ for all $h_1, h_2 \in H$. The **Gaussian mean-width** of H is $w(H) = \mathbb{E} \sup_{h \in H} G_h$.*

Gaussian mean widths of various sets have been computed in Amelunxen et al. (2014), Bellec (2017), Chatterjee and Goswami (2019) or Gordon et al. (2007) for example. Risk bounds for $\hat{f}_\lambda^{RERM}$ are driven by fixed point solutions of a Gaussian mean width of regularization balls $(F - f^*) \cap \rho B$, which measure the local complexity of F around f^* .

Definition 3 *For all $A > 0$, the **complexity function** is a non-decreasing function $r(A, \cdot)$, such that for every $\rho \geq 0$,*

$$r(A, \rho) \geq \inf\{r > 0 : 96 A L_0 L w\big(F \cap (f^* + \rho B \cap r B_{L_2})\big) \leq r^2 \sqrt{N}\} \ .$$

Here, L is the Lipschitz constant of Assumption 2 and L_0 is the subgaussian constant from Assumption 4.

The last tool and assumption comes from Lecué and Mendelson (2018). A key observation is that the regularization norm $\|\cdot\|$ promoting some sparsity structure has large subdifferentials at sparse functions (see, for instance, atomic norms in Bhaskar et al. (2013)). The subdifferential of $\|\cdot\|$ in f is defined as

$$(\partial \|.\|)_f = \{z^* \in E^* \ : \ \|f + h\| - \|f\| \geq z^*(h) \ \text{ for every } h \in E\} \ , \tag{4}$$

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

$$\Gamma_{f^*}(\rho) = \bigcup_{f \in f^* + \frac{\rho}{20} B} (\partial \|\cdot\|)_f \ .$$

$\Gamma_{f^*}(\rho)$ is the union of all subdifferentials of the regularization norm $\|\cdot\|$ of functions f close to the oracle f^*.

Definition 4 *For any $A > 0$ and $\rho > 0$, let*

$$H_{\rho,A} = \{f \in F \ : \ \|f^* - f\| = \rho \ \textit{ and } \ \|f^* - f\|_{L_2} \leq r(A, \rho)\} \ .$$

Let

$$\Delta(\rho, A) = \inf_{h \in H_{\rho,A}} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*) \ . \tag{5}$$

*A real number $\rho > 0$ satisfies the (A-)**sparsity equation** if $\Delta(\rho, A) \geq 4\rho/5$.*

Any constant in $(0, 1)$ could replace $4/5$ in Definition 4 as can be seen from a close inspection of the proof of Theorem 1. If the norm $\|\cdot\|$ is "smooth" in f, the subdifferential of $\|\cdot\|$ in f is just the gradient of $\|\cdot\|$ in f. In that case, $(\partial \|\cdot\|)_f$ is not rich and the regularization norm has only a low "sparsity inducing power" unless the variety of gradients of $\|\cdot\|$ at f in the neighborhood $f^* + (\rho/20)B$ is rich enough (the latter case can be seen as $\|\cdot\|$ being "almost not differentiable" in f^*). However, any norm has a subdifferential in 0 equal to the entire unit dual ball. Therefore, when 0 belongs to $f^* + (\rho/20)B$, for example when $\rho \geq 20\|f^*\|$, the sparsity equation is satisfied since, in that case, $\Delta(\rho) = \rho$. We can use this fact to obtain "complexity dependent" rates of convergence – i.e. rates depending on $\|f^*\|$. In high-dimensional setups, we also look for statistical bounds depending on the sparsity of f^* enforced by $\|\cdot\|$ (see Chinot (2019); Lecué and Mendelson (2017, 2018) for details regarding the difference between "complexity and sparsity" dependent bounds). Hereafter, we focus on norms $\|\cdot\|$ promoting some sparsity structure and we establish sparsity dependent rates of convergence and sparse oracle inequalities in Section 6 (except for matrix completion, see Section 6).

Margin assumptions Mammen and Tsybakov (1999); Tsybakov (2004); van de Geer (2016) such as the Bernstein conditions from Bartlett and Mendelson (2006) have been widely used in statistics and learning theory to prove fast convergence rates of RERM. Here, we use a **local Bernstein condition** in the spirit of Chinot et al. (2018).

Assumption 5 *There exist constants $A > 0$ and ρ^* such that ρ^* satisfies the A-sparsity equation and for all $f \in F$ satisfying $\|f - f^*\|_{L_2} = r(A, \rho^*)$ and $\|f - f^*\| \leq \rho^*$, then $\|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f$.*

Hereafter, whenever Assumption 5 is granted, we assume that the constant A is fixed satisfying this assumption and write $r(\rho)$ instead of $r(A, \rho)$. As explained in Chinot et al. (2018), the local Bernstein condition holds in examples where F is not bounded in L_2-norm. It allows to cover the class of all linear functions on $\mathbb{R}^d$ where the global Bernstein condition of Alquier et al. (2017) – $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ for all $f \in F$– does not hold. In Assumption 5, it is important to remark that only functions from F in the intersection $f^* + (r(A, \rho^*)S_{L_2} \cap \rho^* B)$ have to satisfy the condition "$\|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f$" and not all functions from F in $f^* + r(A, \rho^*)S_{L_2}$. There are situations where this extra intersection with the $\rho^* B$ ball is important to verify the Bernstein Assumption. We provide such an example in Section 6.7.

Remark 1 *From Assumption 2 it follows that if the local Bernstein condition is granted as in Assumption 5 that is for all functions f in F such that $\|f - f^*\|_{L_2} = r(A, \rho^*)$ and $\|f - f^*\| \leq \rho^*$ (and if there exists such an f) then we necessary have $r(A, \rho^*) \leq AL$ since*

$$r^2(A, \rho^*) = \|f - f^*\|_{L_2}^2 \leq AP\mathcal{L}_f \leq AL\|f - f^*\|_{L_2} = ALr(A, \rho^*).$$

It will be always verified as soon as N is large enough. For example, for the LASSO regularization we recover the condition $N \geq cs\log(p/s)$ where s is the oracle's sparsity and $c > 0$ is an absolute constant from this restriction.

3.2 Main theorem for the RERM

The following theorem gives the main result on the statistical performance of RERM.

Theorem 1 *Grant Assumptions 1, 2, 3, 4. Suppose that Assumption 5 holds with $\rho = \rho^*$ satisfying the A-sparsity equation from Definition 4. With this value of A, let $r(\cdot) := r(A, \cdot)$ denote the complexity function from Definition 3. Assume that*

$$\frac{10}{21A}\frac{r^2(\rho^*)}{\rho^*} < \lambda < \frac{2}{3A}\frac{r^2(\rho^*)}{\rho^*} \quad . \tag{6}$$

Then, with probability larger than

$$1 - 2\exp\big(- c(A, L, L_0) r^2(\rho^*) N\big) \quad , \tag{7}$$

the following bounds hold

$$\|\hat{f}_\lambda^{RERM} - f^*\| \le \rho^*, \quad \|\hat{f}_\lambda^{RERM} - f^*\|_{L_2} \le r(\rho^*) \text{ and } P\mathcal{L}_{\hat{f}_\lambda^{RERM}} \le \frac{r^2(\rho^*)}{A} \quad .$$

Remark 2 *A remarkable feature of Theorem 1 is that it holds without assumption on Y. This is an important consequence of the Lipschitz property which has been widely used in robust statistics, see for example Huber and Ronchetti (2011) for an overview.*

Remark 3 *Theorem 1 holds for subgaussian classes of functions F. As in Alquier et al. (2017), it is possible to extend this result under boundedness assumptions.*

Theorem 1 improves (Alquier et al., 2017, Theorem 2.1) in two directions: First, the complexity function $r(\cdot)$ measures the (Gaussian mean width) complexity of the **local** set $(F - f^*) \cap \rho B \cap r B_{L_2}$ and not the global gaussian mean width of $(F - f^*) \cap \rho B$ such as in Alquier et al. (2017). Second, Theorem 1 holds in a setting where F can be unbounded in L_2-norm. The proof of Theorem 1 is postponed to Section 9. The proof relies on the convexity of the loss function (and F) which allows to use an homogeneity argument as in Chinot et al. (2018) for Lipshitz and convex loss functions and in Lecué and Mendelson (2013) for the quadratic loss function, simplifying the peeling step of Alquier et al. (2017). Theorem 1 is a general result which is applied in various applications in Section 6.

4. Minmax MOM estimators

Even if the results of Section 3 are interesting on their own because the i.i.d. sub-gaussian framework is one of the most considered setup in Statistics and Learning theory, the setup considered in Section 3 can be restrictive in some applications. It does not cover more realistic situations where data are heavy-tailed and/or corrupted. In this section, we consider a more general setup covering these situations. The results of Section 3 are used as benchmarks: we show that similar bounds can be achieved in a more realistic framework by alternative estimators. These use median-of-means principles instead of empirical means.

4.1 Definition

Recall the definition of MOM estimators of univariate means from Alon et al. (1999); Jerrum et al. (1986); Nemirovsky and Yudin (1983). Let $(B_k)_{k=1,\dots,K}$ denote a partition of $\{1, \dots, N\}$ into blocks B_k of equal size N/K (it is implicitly assumed that K divides N. An extension to blocks with almost equal size is possible. It is not considered here to simplify the presentation of the results, the extension is thus left to the interested reader). For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $k \in \{1, \dots, K\}$, let $P_{B_k} f = (K/N) \sum_{i \in B_k} f(X_i, Y_i)$ denote the empirical mean on the block B_k. The MOM estimator based on this partition is the empirical median of the latter empirical means:

$$\text{MOM}_K [f] = \text{Med}(P_{B_1} f, \cdots, P_{B_K} f) \ . \tag{8}$$

The estimator $\text{MOM}_K [f]$ of Pf achieves subgaussian deviation tails if $(f(X_i, Y_i))_{i=1}^N$ have 2 moments, see Devroye et al. (2016). The number of blocks K is a tuning parameter of the procedure. The larger K, the more outliers are allowed. When $K = 1$, $\text{MOM}_K [f]$ is the empirical mean, when $K = N$, it is the empirical median.

Building on ideas introduced in Audibert and Catoni (2011); Baraud et al. (2017), Lecué and Lerasle (2019) proposed the following strategy to use MOM estimators in learning problems. Since the *oracle* f^* is also solution of the following minmax problem

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

minmax MOM estimators are obtained by plugging MOM estimators of the unknown expectations $P(\ell_f - \ell_g)$ in this minmax formulation. Applying this principle to regularized procedures yields the following "minmax MOM version" of RERM that we study in this paper:

$$\hat{f}_{K,\lambda} \in \operatorname*{argmin}_{f \in F} \sup_{g \in F} \text{MOM}_K [\ell_f - \ell_g] + \lambda \big(\|f\| - \|g\| \big) \ . \tag{9}$$

The linearity of the empirical process P_N is important to use localization techniques in the analysis of RERM to derive fast rates of convergence for these estimators improving upon the slow rates of Vapnik (1998), see Tsybakov (2004); Koltchinskii (2011) for example. The minmax reformulation comes from Audibert and Catoni (2011), it allows to overcome the lack of linearity of robust mean estimators and obtain fast rates of convergence for robust estimators based on nonlinear estimators of univariate expectations.

4.2 Assumptions and main results

To highlight robustness properties of minmax MOM estimators with respect to outliers in the dataset, their analysis is performed in the following framework. Let $\mathcal{I} \cup \mathcal{O}$ denote a partition of $\{1, \cdots N\}$ that is unknown to the statistician. Data $(X_i, Y_i)_{i \in \mathcal{O}}$ are considered as outliers. **No assumption** on the distribution of these data is made, they can be dependent or adversarial. Data $(X_i, Y_i)_{i \in \mathcal{I}}$ bring information on f^* and are called informative or inliers. All assumptions are made on these data. They have to induce the same L_2 geometries on F and the same excess risks.

Assumption 6 *$(X_i, Y_i)_{i \in \mathcal{I}}$ are independent and for all $i \in \mathcal{I}$: $P_i(f - f^*)^2 = P(f - f^*)^2$ and $P_i \mathcal{L}_f = P\mathcal{L}_f$.*

Assumption 6 holds in the i.i.d case, it also covers situations where informative data $(X_i, Y_i)_{i \in I}$ may have different distributions. It implies in particular that f^* is also the oracle in F w.r.t. all the distributions P_i for $i \in \mathcal{I}$.
Several quantities introduced to study RERM have to be modified to state the results for minmax MOM estimators. First, the complexity function is no longer based on Gaussian mean width,

it is now defined as a fixed point of local Rademacher complexities. Let $(\sigma_i)_{i\in\mathcal{I}}$ denote i.i.d. Rademacher random variables (i.e. uniformly distributed on $\{-1,1\}$), independent from $(X_i,Y_i)_{i\in\mathcal{I}}$. The **complexity function** $\rho\to r_2(\gamma,\rho)$ is a non-decreasing function defined for all $\rho>0$ as

$$r_2(\gamma,\rho)\geq \inf\left\{r>0:\forall J\subset\mathcal{I}\text{ s.t } |J|\geqslant N/2,\quad \mathbb{E}\left[\sup_{f\in(F-f^*)\cap\rho B\cap rB_{L_2}}\left|\sum_{i\in J}\sigma_i f(X_i)\right|\right]\leq \gamma r^2|J|\right\} \ . \tag{10}$$

As in Theorem 1, this parameter measures the complexity of $(F-f^*)\cap\rho B$ locally in a L_2-neighborhood of 0. It only involves the distribution of informative data and does not depend on the distribution of the outputs $(Y_i)_{i\in\mathcal{I}}$. The local Bernstein condition, Assumption 5, as well as the sparsity equation have now to be extended to this new definition of complexity. Start with the sparsity equation.

Definition 5 *For any $A>0$ and $\rho>0$, let*

$$C_{K,r}(\rho,A)=\max\left(r_2^2(\gamma,\rho),c(A,L)\frac{K}{N}\right) \ . \tag{11}$$

and $\tilde{H}_{\rho,A}=\left\{f\in F \ : \ \|f^*-f\|=\rho \text{ and } \|f^*-f\|_{L_2}\leq\sqrt{C_{K,r}(\rho,A)}\right\}$. Let

$$\tilde{\Delta}(\rho,A)=\inf_{h\in\tilde{H}_{\rho,A}}\sup_{z^*\in\Gamma_{f^*}(\rho)}z^*(h-f^*) \ . \tag{12}$$

*A real number $\rho>0$ satisfies the A-**sparsity equation** if $\tilde{\Delta}(\rho,A)\geq 4\rho/5$.*

The value of $c(A,L)$ in Definition 5 is made explicit in Section 10. To simplify the presentation we write $c(A,L)$ as it is an absolute constant depending only on A and L. With this definition in mind, one can extend the local Bernstein assumption.

Assumption 7 *There exist a constant $A>0$ and ρ^* such that ρ^* satisfies the A-sparsity equation from Definition 5 and, for all $f\in F$ such that $\|f-f^*\|_{L_2}^2=C_{K,r}(2\rho^*,A)$ and $\|f-f^*\|\leq 2\rho^*$, $\|f-f^*\|_{L_2}^2\leqslant AP\mathcal{L}_f$.*

As in Assumption 5, the link between $\|f-f^*\|_{L_2}$ and the excess risk $P\mathcal{L}_f$ in Assumption 7 is only granted in a $L_2(\mu)$-sphere around the oracle f^* whose radius is proportional to the rate of convergence of the estimators (see Theorems 1 and 2). The local Bernstein assumption is somehow "minimal" since it is only granted on the smallest set of the form $F\cap(f^*+2\rho^*B\cap r_2(\gamma,2\rho^*)B_{L_2})$ centered in f^* that can be proved to contain $\hat{f}_{K,\lambda}$ (when K is such that $\sqrt{C_{K,r}(2\rho^*,A)}=r_2(\gamma,2\rho^*)$).

Remark 4 *As in Remark 1 we necessary have $\sqrt{C_{K,r}(2\rho^*,A)}\leq AL$.*

We are now in position to state our main result on the statistical performances of the regularized minmax MOM estimator.

Theorem 2 *Grant Assumptions 2, 3, 6 and 7 for ρ^* satisfying the A-sparsity equation from Definition 5. Let $K\geq 7|\mathcal{O}|/3$, $\gamma=1/(6528L)$, and define*

$$\lambda=\frac{5}{17A}\frac{C_{K,r}(2\rho^*,A)}{\rho^*} \ .$$

Then, with probability larger than $1 - 2\exp(-cK)$, the minmax MOM estimator $\hat{f}_{K,\lambda}$ defined in (9) *satisfies*

$$\|\hat{f}_{K,\lambda} - f^*\| \leq 2\rho^*, \quad \|\hat{f}_{K,\lambda} - f^*\|_{L_2}^2 \leq C_{K,r}(2\rho^*, A) \quad \textit{and} \quad P\mathcal{L}_{\hat{f}_{K,\lambda}} \leq \frac{1}{A} C_{K,r}(2\rho^*, A) \ .$$

Suppose that $K = c(A, L) r_2^2(\gamma, 2\rho^*) N$, which is possible as long as $|\mathcal{O}| \leq c(A, L) N r_2^2(\gamma, 2\rho^*)$. The L_2-estimation bound is then $r_2^2(\gamma, 2\rho^*)$ and the probability that this bound the L_2 risk is $1 - \exp(-c(A, L) N r_2^2(\gamma, 2\rho^*))$. Up to absolute constants, regularized minmax MOM estimators achieve the same bounds as RERM with the same probability as in the i.i.d. subgaussian framework from Theorem 1. Indeed, in that case, a straighforward chaining argument shows that the Rademacher complexity from (10) is upper bounded by the Gaussian mean width. The difference with Theorem 1 is that the estimator depends on K. On the other hand, the results hold in a setting where the subgaussian assumption on F is relaxed into moment assumptions and the data may not be identically distributed and may have been corrupted by some outliers. Section 6.2 will show an example where rate optimal bounds can be derived from this general result. It is also possible to adapt in a data-driven way to the best K and λ by using a Lepski's adaptation method such as in Devroye et al. (2016); Lecué and Lerasle (2017, 2019); Chinot et al. (2018); Chinot (2019). This step is now well understood, it is not reproduced here. Theorem 2 is "universal" in the sense that it allows to handle many applications, some of these are presented in Section 6

5. Relaxing the Bernstein condition

In this section, we study minmax MOM estimators when the Bernstein assumption 7 is relaxed. The price to pay for the relaxation is that, on one hand, the L_2-risk is not controlled and on the other hand an extra variance term appears in the excess risk $P\mathcal{L}_{\hat{f}_K^\lambda}$. Nevertheless, under a slightly stronger local Bernstein's condition, the extra variance term can be controled and the bounds from Theorem 2 is recovered .

Assumption 8 *$(X_i, Y_i)_{i \in \mathcal{I}}$ are independent and for all $i \in \mathcal{I}$, (X_i, Y_i) has distribution P_i, X_i has distribution μ_i. We assume that, for any $i \in \mathcal{I}$, $F \subset L_1(\mu_i)$ and $P_i \mathcal{L}_f = P\mathcal{L}_f$ for all $f \in F$.*

Since the local Bernstein Assumption 7 does not hold, the localization argument has to be modified. Instead of using the L_2-norm to define neighborhoods of f^*, we use the excess loss $f \in F \to P\mathcal{L}_f$. The new fixed point is defined for all $\gamma, \rho > 0$ and $K \in \{1, \cdots, N\}$:

$$\bar{r}(\gamma, \rho) = \inf \left\{ r > 0 : \max \left(\frac{E(r, \rho)}{\gamma}, \sqrt{c} V_K(r, \rho) \right) \leq r^2 \right\}, \quad \text{where} \tag{13}$$

$$E(r, \rho) = \sup_{J \subset \mathcal{I} : |J| \geq N/2} \mathbb{E} \sup_{f \in F : P\mathcal{L}_f \leq r^2, \, \|f - f^*\| \leq \rho} \left| \frac{1}{|J|} \sum_{i \in J} \sigma_i (f - f^*)(X_i) \right| \ ,$$

$$V_K(r, \rho) = \max_{i \in \mathcal{I}} \sup_{f \in F : P\mathcal{L}_f \leq r^2, \, \|f - f^*\| \leq \rho} \left(\sqrt{\mathbb{V}ar_{P_i}(\mathcal{L}_f)} \right) \sqrt{\frac{K}{N}} \ ,$$

and $(\sigma_i)_{i \in \mathcal{I}}$ are i.i.d. Rademacher random variables independent from $(X_i, Y_i)_{i \in \mathcal{I}}$. The value of c in Equation (13) can be found in Section 11. The main differences between $r_2(\gamma, \rho)$ in (13) and

$\tilde{r}(\gamma, \rho)$ in (10) are the extra variance V_K term and the L_2 localization which is replaced by an "excess of risk" localization. Under the local Bernstein Assumption 9, this extra term disappears and this complexity matches the one of Theorem 2. As in Section 4, the sparsity equation has to be extended.

Definition 6 *For any $\rho > 0$, let*

$$\bar{H}_\rho = \left\{ f \in F \;\; : \;\; \|f^* - f\| = \rho \;\; and \;\; P\mathcal{L}_f \leq \bar{r}^2(\gamma, \rho) \right\} \;\; . \tag{14}$$

Let

$$\bar{\Delta}(\rho) = \inf_{h \in \bar{H}_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*) \;\; . \tag{15}$$

*A real number $\rho > 0$ satisfies the **sparsity equation** if $\bar{\Delta}(\rho) \geq 4\rho/5$.*

We are now in position to state the main result of this section.

Theorem 3 *Grant Assumptions 2, 3, 8 and assume that $|\mathcal{O}| \leq 3N/7$. Let ρ^* satisfying the sparsity equation from Definition 6. Let $\gamma = 1/(3840L)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. Define*

$$\lambda = \frac{11}{40} \frac{\bar{r}^2(\gamma, 2\rho^*)}{\rho^*}$$

The minmax MOM estimator $\hat{f}_K^\lambda$ defined in (9) satisfies, with probability at least $1 - 2\exp(-cK)$,

$$P\mathcal{L}_{\hat{f}_K^\lambda} \leq \bar{r}^2(\gamma, 2\rho^*) \quad and \quad \|\hat{f}_K^\lambda - f^*\| \leq 2\rho^* \;\; .$$

In Theorem 3, the only stochastic assumption is Assumption 8 which says that the inliers data are independent and define the same excess risk as (X, Y) over F. In particular, Theorem 3 does not assume anything on the outliers $(X_i, Y_i)_{i \in \mathcal{O}}$ nor on the outputs of the inliers $(Y_i)_{i \in \mathcal{I}}$. The difficulty of the problem is contained in the computation of the local Rademacher complexities $E(r, \rho)$ that are computed in examples using stronger stochastic assumptions.

To conclude the section, let us show that Theorem 2 can be recovered from Theorem 3 under the following slightly stronger local Bernstein assumption.

Assumption 9 *There exist a constant $\bar{A} > 0$ and ρ^* satisfying the sparsity equation from Definition 6 such that, for all $f \in F$, if $P\mathcal{L}_f \leqslant \bar{C}_{K,r}(\rho^*, \bar{A})$ and $\|f - f^*\| \leq 2\rho^*$, then $\|f - f^*\|_{L_2}^2 \leqslant \bar{A} P\mathcal{L}_f$, where*

$$\bar{C}_{K,r}(\rho, A) = \max\left(\frac{r_2^2(\gamma/A, 2\rho)}{\sqrt{A}}, c(A, L)\frac{K}{N} \right) \quad and \quad \gamma = 1/(3840L) \;\; . \tag{16}$$

Up to constants, $\bar{C}_{K,r}$ is equivalent to $C_{K,r}$ given in Definition 5. Assumption 9 holds for any function $f \in F$ such that $P\mathcal{L}_f \leq \bar{C}_{K,r}(\rho^*, \bar{A})$ which is slightly stronger than being in the L_2-sphere as in Assumption 7.

Theorem 4 *Grant Assumptions 2, 3, 6 and assume that $|\mathcal{O}| \leq 3N/7$. Assume that the local Bernstein condition Assumption 9 holds with ρ^* satisfying the $\bar{A}$-sparsity equation from Definition 6. Let $\gamma = 1/(3840L)$ and $K \in \big[7|\mathcal{O}|/3, N\big]$. Define*

$$\lambda = \frac{11}{40}\frac{\bar{r}^2(\gamma, 2\rho^*)}{\rho^*} \quad .$$

The minmax MOM estimator $\hat{f}_K^\lambda$ defined in (9) satisfies, with probability at least $1 - 2\exp(-cK)$,

$$||\hat{f}_K^\lambda - f^*||_{L_2}^2 \leq \bar{C}_{K,r}(\rho^*, \bar{A}), \quad P\mathcal{L}_{\hat{f}_K^\lambda} \leq \bar{C}_{K,r}(\rho^*, \bar{A}) \quad \textit{and} \quad \|\hat{f}_K^\lambda - f^*\| \leq 2\rho^* \quad .$$

Theorem 4 is proved in Section 11.1.

Remark 5 *Under Assumption 9 and a slight modification in the constants, ρ^* satisfies the sparsity equation of Definition 6 if it verifies the sparsity equation of Definition 5.*

6. Applications

This section presents some applications of Theorem 2. To check the assumptions of the Theorem 2, the following routine is applied:

1. Check Assumptions 2, 3, 6.
2. Compute the local rademacher complexity $r_2(\gamma, \rho)$.
3. Solve the sparsity equation from Definition 5: find ρ^* such that $\Delta(\rho^*, A) \geq 4\rho^*/5$.
4. Check the local Bernstein condition from Assumption 7.

In this section, we focus on high dimensional statistical problems with sparsity inducing regularization norms Bach et al. (2012) given by the ℓ_1 norm Tibshirani (1996), the SLOPE norm Bogdan et al. (2015), the group LASSO norm Simon et al. (2013), the Total Variation norm Osher et al. (2005), the Fused Lasso norm Tibshirani et al. (2005) and the nuclear norm Koltchinskii et al. (2011). In the vector case, we consider the class of linear functions $F = \{\langle t, \cdot\rangle : t \in \mathbb{R}^p\}$ indexed by $\mathbb{R}^p$. We denote by $t^* \in \mathbb{R}^p$ the vector such that $f^*(\cdot) = \langle t^*, \cdot\rangle$. We consider the logistic loss function for the LASSO and the SLOPE, with data $(X_i, Y_i)_{i=1}^N$ taking values in $\mathbb{R}^p \times \{-1, 1\}$ and the Huber loss function for the Group LASSO, the Total Variation and Fused Lasso, with data $(X_i, Y_i)_{i=1}^N$ taking values in $\mathbb{R}^p \times \mathbb{R}$. In particular, the results of this section extend results on the logistic LASSO and logistic SLOPE from Alquier et al. (2017) and present new results for the Group Lasso, the Total Variation and the Fused Lasso. Similar notation are used in the matrix case for the quantile loss and the trace-norm regularization in Section 6.7.

6.1 Preliminary results

In this section, we recall some tools to check the Local Bernstein condition, compute the local Rademacher complexity and verify the sparsity equation.

6.1.1 Local Bernstein conditions for the logistic, Huber and quantile losses

In this section, we recall results from Chinot et al. (2018) on the local Bernstein condition for the logistic, Huber and quantile loss functions. For the logistic loss function (i.e. $\ell_f : (x, y) \in \mathbb{R}^p \times \{\pm 1\} \to \log(1 + \exp(-yf(x)))$), we first introduce the following assumption. Note that we do not use the localization with resect to the regularization norm in this Section.

Assumption 10 *Let $\varepsilon > 0$, there are constants C' and $c_0 > 0$ such that*

- *a) for all f in F, $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$*
- *b) $\mathbb{P}(|f^*(X)| \leq c_0) \geq 1 - 1/(2C')^{(4+2\varepsilon)/\varepsilon}$*

It turns out that the localization with respect to the regularization norm becomes useful when the constant C' from point a) in assumption 10 depends on the dimension of the problem (see Section 6.7 for a clear example). For the sake of simplicity we only show that the Bernstein condition is verified on a L_2-ball ignoring the regularization localization.

Proposition 1 (Chinot et al. (2018), Theorem 9) *Grant Assumption 10. Let $r > 0$. The local Bernstein condition holds for the logistic loss function: for all $f \in F$ such that $\|f - f^*\|_{L_2} \leqslant r$, $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ for*

$$A = \frac{\exp\big(-c_0 - r(2C')^{(2+\varepsilon)/\varepsilon}\big)}{2\bigg(1 + \exp\big(c_0 + r(2C')^{(2+\varepsilon)/\varepsilon}\big)\bigg)^2} \ .$$

Note that if r is larger than the order of a constant then A is no longer a constant and the convergence rates are deteriorated (see the link with Remark 1).

For the Huber loss function with parameter $\delta > 0$ (i.e. $\ell_f(x, y) = \rho_\delta(y - f(x))$ where $\rho_\delta(t) = t^2/2$ if $|t| \leq \delta$ and $\rho_\delta(t) = \delta|t| - \delta^2/2$ if $|t| \geq \delta$), we use the following result also borrowed from Chinot et al. (2018). Let us introduce the following assumption.

Assumption 11 *Let $\varepsilon > 0$ and let $F_{Y|X=x}$ be the conditional cumulative function of Y given $X = x$.*

- *a) There exists a constant C' such that, for all f in F, $\|f - f^*\|_{L_{2+\varepsilon}} \leq C'\|f - f^*\|_{L_2}$.*
- *b) Let C' be the constant defined in a). There exist $r > 0$ and $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \leq r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon}$, $F_{Y|X=x}(z + \delta) - F_{Y|X=x}(z - \delta) \geqslant \alpha$.*

Note that if r is larger than the order of a constant the point b) can be verified only if δ, the Lipschitz constant, is large enough. It would degrade the convergence rates. It is related, one more time, with Remark 1.

Proposition 2 (Chinot et al. (2018), Theorem 7) *Grant Assumption 11 for $r > 0$. The Huber loss function with parameter $\delta > 0$ satisfies the Bernstein condition: for all $f \in F$, if $\|f - f^*\|_{L_2} \leq r$ then $(4/\alpha)P\mathcal{L}_f \geq \|f - f^*\|_{L_2}^2$.*

Je pense qu'on peut le virer cet exemple car tu vas le refaire a la main. Il faudra bien que tu insiste dans la partie matrix completion qu'on verifie bersntein sur un sous espace de la space et que la localisation par rapport a la norme de regularization nous aide lorsque la constante C' de L_4/L_2 depend de la dimension

Finally, we recall a third and last result on the local Bernstein for the quantile loss from Chinot et al. (2018). The quantile loss function with parameter $\tau \in (0,1)$ is $\ell_f(x,y) = \rho_\tau(y - f(x))$ where $\rho_\tau(t) = t(\tau - I(t \le 0))$ for all $t \in \mathbb{R}$. The local Bernstein inequality holds for the quantile loss function under the following assumption.

Assumption 12 *Let $\varepsilon > 0$. For all $x \in \mathcal{X}$ let $f_{Y|X=x}$ denote the density function on Y given $X = x$*

a) There exists a constant C' such that, for all f in F, $\|f - f^*\|_{L_{2+\epsilon}} \le C'\|f - f^*\|_{L_2}$.

b) Let C' be the constant defined in a). There exist $r > 0$ and $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \le r(\sqrt{2}C')^{(2+\varepsilon)/\varepsilon}$, $f_{Y|X=x}(z) \ge \alpha$.

Under Assumption 12, the quantile loss function satisfies a local Bernstein inequality as proved in Theorem 6 in Chinot et al. (2018). We recall this result here because we will use it for the problem of robust quantile matrix completion.

Proposition 3 (Chinot et al. (2018), Theorem 6) *Grant Assumption 12 for $r > 0$. The quantile loss function with parameter $\tau \in (0,1)$ satisfies the Bersntein condition: for all $f \in F$, if $\|f - f^*\|_{L_2} \le r$ then $\|f - f^*\|_{L_2}^2 \le (4/\alpha)P\mathcal{L}_f$.*

Note that in Assumptions 10, 11 and 12, the value of C' can depend on the dimension provided that $(C')^{(2+\varepsilon)/\varepsilon} r = O(1)$. Moreover Propositions 1, 2 and 3 state that the Bernstein condition is verified over the entire set $F \cap (f^* + rB_{L_2})$. In Section 6.7 we study the problem of quantile matrix completion where C' depends on the dimension. In this precise example we use the fact that the Bernstein condition only required in $F \cap (f^* + \sqrt{C_{K,r}(2\rho^*, A)}S_{L_2} \cap 2\rho^* B)$ in Assumption 7. The localization with respect to the regularization norm allows to cover such examples when C' may depend on the dimension of the problem.
Other examples of loss functions satisfying a local Bernstein condition can be found in Chinot et al. (2018). je ne sais pas si c'est tres clair ce que jje dis ...

6.1.2 Local Rademacher complexities and Gaussian mean widths

The computation of $r_2(\gamma, \rho)$ may be involved, but can sometimes be reduced to the computation of Gaussian mean widths, thanks to Mendelson (2017). The results of Mendelson (2017) are based on the concepts of unconditional norm and isotropic random vectors.

Definition 1 *For a given vector $x = (x_i)_{i=1}^p$, let $(x_i^*)_{i=1}^p$ be the non-increasing rearrangement of $(|x_i|)_{i=1}^p$. The norm $\|\cdot\|$ in $\mathbb{R}^p$ is said κ-unconditional with respect to the canonical basis $(e_i)_{i=1}^p$ if, for every x in $\mathbb{R}^p$ and every permutation π of $\{1, \cdots, p\}$,*

$$\left\| \sum_{i=1}^p x_i e_i \right\| \le \kappa \left\| \sum_{i=1}^p x_{\pi(i)} e_i \right\| ,$$

and, for any $y \in \mathbb{R}^p$ such that, for all $1 \leq i \leq p$, $x_i^* \leq y_i^*$, then

$$\left\| \sum_{i=1}^{p} x_i e_i \right\| \leq \kappa \left\| \sum_{i=1}^{p} y_i e_i \right\| \ .$$

Definition 2 *A random vector X in $\mathbb{R}^p$ is isotropic if $\mathbb{E}[\langle t, X \rangle^2] = \|t\|_2^2$, for all $t \in \mathbb{R}^p$, where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^p$.*

Recall the main result of Mendelson (2017).

Theorem 5 *(Mendelson, 2017, Theorem 1.6) Let C_0, κ and M be real numbers. Let $V \subset \mathbb{R}^p$ such that $\sup_{v \in V} |\langle v, \cdot \rangle|$ is κ-unconditional with respect to $(e_i)_{i=1}^p$. Assume that $X \in \mathbb{R}^p$ is isotropic and satisfies, for all $1 \leq j \leq p$ and $1 \leq q \leq C_0 \log(p)$,*

$$\left\| \langle X, e_j \rangle \right\|_{L_q} \leq M\sqrt{q} \ . \tag{17}$$

Let $X_1, \ldots, X_N$ denote independent copies of X, then there exists a constant c_2 depending only on C_0 and M such that

$$\mathbb{E}\left[\sup_{v \in V} \sum_{i=1}^{N} \sigma_i \langle X_i, v \rangle \right] \leq c_2 \kappa \sqrt{N} w(V)$$

where $w(V)$ is the Gaussian mean width of V.

Recall that a real valued random variable Z is L_0-subgaussian if and only if for all $q \geq 1$, $\|Z\|_{L_q} \leq c_0 L_0 \sqrt{q}$, for some absolute constant c_0, see Theorem 1.1.5 in Chafaï et al. (2012). Hence, Theorem 5 assumes that the coordinates of the design X have at least $C_0 \log(p)$ "subgaussian" moments.

6.1.3 Sub-differential of a norm

To solve the sparsity equation – find ρ^* such that $\tilde{\Delta}(\rho^*, A) \geq 4\rho^*/5$ – we use the following classical result on the sub-differential of a norm: if $\|\cdot\|$ is a norm on $\mathbb{R}^p$, then, for all $t \in \mathbb{R}^p$, we have

$$(\partial \|\cdot\|)_t = \begin{cases} \{z^* \in S^* : \langle z^*, t \rangle = \|t\|\} & \text{if } t \neq 0 \\ B^* & \text{if } t = 0 \end{cases} . \tag{18}$$

Here, B^* is the unit ball of the dual norm associated with $\|\cdot\|$, i.e. $t \in \mathbb{R}^p \to \|t\|^* = \sup_{\|v\| \leq 1} \langle v, t \rangle$ and S^* is its unit sphere. In other words, when $t \neq 0$, the sub-differential of $\|\cdot\|$ in t is the set of all vectors z^* in the unit dual sphere S^* which are norming for t.

In the following, understanding the sub-differentials of the regularization norm is a key point for solving the sparsity equation. If one is only interested in proving "complexity" dependent bounds – which are bounds depending on $\|t^*\|$ and not the sparsity of t^* – then one can simply take $\rho^* = 20\|t^*\|$. Actually, in this case, $0 \in \Gamma_{t^*}(\rho)$, so $\tilde{\Delta}(\rho^*, A) = \rho^* \geq 4\rho^*/5$ (because $B^* = (\partial \|\cdot\|)_0 = \Gamma_{t^*}(\rho)$ according to (18)). Therefore, understanding the sub-differential of the regularization norm matters when one wants to derive statistical bounds depending on the dimension of the low-dimensional structure that contains t^*. This is something expected since a norm has sparsity inducing power if its sub-differential is a big part of the dual sphere at vectors having the sparse structure (see, for instance, the construction of atomic norms in Bhaskar et al. (2013)).

We now have all the necessary tools to derive statistical bounds for many procedures by applying Theorem 2. In each example (given by a convex and Lipschitz loss function and a regularization norm), we just have to compute the complexity function r_2, solve a sparsity equation and check the local Bernstein condition.

6.2 The minmax MOM logistic LASSO procedure

When the dimension p of the problem is large and $\|t^*\|_0 = |\{i \in \{1,\cdots,p\} : t_i^* \neq 0\}|$ is small, it is possible to derive error rate depending on the size of the support of t^* instead of the dimension p by using a ℓ_1 regularization norm. It leads to the well-known LASSO estimators, see Tibshirani (1996). For the logistic loss function, its minmax MOM formulation is the following. For a given $K \in \{1,\dots,N\}$ and $\lambda > 0$, the minmax MOM logistic LASSO procedure is defined by

$$\hat{t}_{\lambda,K} \in \underset{t\in\mathbb{R}^p}{\operatorname{argmin}} \sup_{\tilde{t}\in\mathbb{R}^p} \left(\text{MOM}_K \left[\ell_t - \ell_{\tilde{t}}\right] + \lambda(\|t\|_1 - \|\tilde{t}\|_1) \right) ,$$

with the logistic loss function defined as $\ell_t(x,y) = \log(1+\exp(-y\langle x,t\rangle))$ for all $t, x \in \mathbb{R}^p$ and $y \in \{\pm 1\}$, and with the ℓ_1 regularization norm defined for all $t \in \mathbb{R}^p$ by $\|t\|_1 = \sum_{i=1}^p |t_i|$.

We first compute the complexity function r_2. Theorem 5 can be applied to upper bound the Rademacher complexities from (10) in that case because the dual norm of ℓ_1 (i.e the ℓ_∞-norm) is 1-unconditional with respect to $(e_i)_{i=1}^p$. Then, if X is an isotropic random vector satisfying (17), Theorem 5 holds and

$$\mathbb{E} \sup_{t\in\rho B_1^p \cap rB_2^p} \left| \sum_{j\in J} \sigma_j \langle t, X_j\rangle \right| \leq c(C_0, M)\sqrt{|J|}w(\rho B_1^p \cap rB_2^p) \ ,$$

where B_1^p denote the unit ball of the ℓ_1 norm. From (Lecué and Mendelson, 2018, Lemma 5.3), we have

$$w(\rho B_1^p \cap rB_2^p) \leq c \left\{ \begin{array}{ll} r\sqrt{p} & \text{if } r \leq \rho/\sqrt{p} \\ \rho\sqrt{\log(ep\min(r^2/\rho^2,1))} & \text{if } r \geq \rho/\sqrt{p} \end{array} \right. . \tag{19}$$

Therefore, one can take

$$r_2^2(\gamma,\rho) = c(\gamma, C_0, M) \left\{ \begin{array}{ll} \frac{p}{N} & \text{if } N\rho^2 \geq c(\gamma,C_0,M)\gamma p^2 \\ \rho\sqrt{\frac{1}{N}\log\left(\frac{ep^2}{\rho^2 N}\right)} & \text{if } \log p \leq c(\gamma,C_0,M)N\rho^2 \leq c(\gamma,C_0,M)p^2 \\ \rho\sqrt{\frac{\log p}{N}} & \text{if } \log p \geq c(\gamma,C_0,M)N\rho^2. \end{array} \right. . \tag{20}$$

Let us turn to the local Bersntein assumption. We need to verify Assumption 10. Let $\varepsilon > 0$. If X is an isotropic random vector satisfying (17) and $C_0 \log(p) \geq 2+\varepsilon$, where C_0 is the constant appearing in Equation (17), then the point a) of Assumption 10 is verified with $C' = c(M, C_0)$. For any $x \in \mathbb{R}^p$, let us write $f^*(x) = \langle x, t^*\rangle$, where $t^* \in \mathbb{R}^p$. Let us assume that the oracle is such that

$$\mathbb{P}\left(|\langle X, t^*\rangle| \leq c_0\right) \geq 1 - \frac{1}{2(C')^{(4+2\varepsilon)/\varepsilon}}. \tag{21}$$

Therefore, if Equation (21) holds, the local Bernstein Assumption is verified for a constant A depending on M, C_0 and c_0 given in Proposition 1 (since the latter formula is rather complicated, we will keep the notation A all along this section).

Finally, let us turn to a solution to the sparsity equation for the ℓ_1^p norm . The result can be found in Lecué and Mendelson (2018).

Lemma 3 *(Lecué and Mendelson, 2018, Lemma 4.2) . Let us assume that X is isotropic. If the oracle t^* can be decomposed as $t^* = v + u$ with $u \in (\rho/20)B_1^p$ and $100s \leq \left(\rho/\sqrt{C_{K,r}(\rho,A)}\right)^2$ then $\Delta(\rho) \geq (4/5)\rho$, where $s = |supp(v)|$.*

Assume that t^* is a s-sparse vector, so Lemma 3 applies. We consider two cases depending on the values of K and $Nr_2^2(\gamma,\rho^*)$. When $C_{K,r}(\rho^*,A) = r_2^2(\gamma,\rho^*)$ – which holds when $K \leq c(c_0,C_0,M)Nr_2^2(\gamma,\rho^*)$ – Lemma 3 shows that $\rho^* = c(c_0,M,C_0)s\sqrt{\log{(ep/s)}/N}$ satisfies the sparsity equation. For these values, the value of r_2 given in (20) yields

$$r_2^2(\gamma,\rho^*) = c(c_0,M,C_0,\gamma)\frac{s\log(ep/s)}{N} \ .$$

Now, if $C_{K,r}(\rho,A) = c(A,L)K/N$ – which holds when $K \geq c(c_0,C_0,M)Nr_2^2(\gamma,\rho^*)$– we can take $\rho^* = c(c_0,M,C_0)\sqrt{sK/N}$. Therefore, Theorem 2 applies with

$$\rho^* = c(c_0,M,C_0)\max(s\sqrt{\log{(ep/s)}/N}, \sqrt{sK/N}) \ .$$

Finally from Remark 1, not that is necessary to have $N \geq c\log(ep/s)$, where $c>0$ is an absolute constant.

Theorem 6 *Let $\varepsilon > 0$ and (X,Y) be a random variable taking values in $\mathbb{R}^p \times \{\pm 1\}$, where X is an isotropic random vector such that for all $1 \leq j \leq p$ and $1 \leq q \leq C_0\log(p)$, $\left\|\left\langle X,e_j\right\rangle\right\|_{L_q} \leq M\sqrt{q}$ with $C_0\log(p) \geq 2+\varepsilon$. Let $f^*: x \in \mathbb{R}^p \mapsto \left\langle x,t^*\right\rangle$ be the oracle where $t^* \in \mathbb{R}^p$ is s-sparse. Assume also that the oracle satisfies Equation (21). Assume that (X,Y), $(X_i,Y_i)_{i\in\mathcal{I}}$ are i.i.d distributed and $N \geq c\log(ep/s)$. Let $K \geq 7|\mathcal{O}|/3$. With probability larger than $1-2\exp(-cK)$, the minmax MOM logistic LASSO estimator $\hat{t}_{\lambda,K}$ with*

$$\lambda = c(c_0,M,C_0)\max\left(\sqrt{\frac{\log(ep/s)}{N}}, \sqrt{\frac{K}{sN}}\right)$$

satisfies

$$\|\hat{t}_{\lambda,K} - t^*\|_1 \leq c(c_0,M,C_0)\max\left(s\sqrt{\frac{\log(ep/s)}{N}}, \sqrt{s}\sqrt{\frac{K}{N}}\right),$$
$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(c_0,M,C_0)\max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N}\right) \ ,$$
$$P\mathcal{L}_{\hat{f}_{\lambda,K}} \leq c(c_0,M,C_0)\max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N}\right) \ .$$

For $K \leq c(c_0,M,C_0)s\log(ep/s)$, the upper bound on the estimation risk and excess risk matches the minimax rates of convergence for s-sparse vectors in $\mathbb{R}^p$. It is also possible to adapt in a data-driven way to the best K and λ by using a Lepski's adaptation method such as in Devroye et al. (2016); Lecué and Lerasle (2017, 2019); Chinot et al. (2018); Chinot (2019). This step is now well understood, it is not reproduced here.

6.3 The minmax MOM logistic SLOPE

In this section, we study the minmax MOM estimator with the logistic loss function and the SLOPE regularization norm. Given $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_p > 0$, the SLOPE norm (see Bogdan et al. (2015)) is defined for for all $t \in \mathbb{R}^p$ by

$$\|t\|_{\text{SLOPE}} = \sum_{i=1}^p \beta_i t_i^* \ ,$$

where $(t_i^*)_{i=1}^p$ denotes the non-increasing re-arrangement of $(|t_i|)_{i=1}^p$. The SLOPE norm coincides with the ℓ_1 norm when $\beta_j = 1$ for all $j = 1, \cdots, p$.

Given $K \in \{1, \dots, N\}$ and $\lambda > 0$, the minmax MOM logistic SLOPE procedure is

$$\hat{t}_{\lambda,K} \in \underset{t \in \mathbb{R}^p}{\operatorname{argmin}} \sup_{\tilde{t} \in \mathbb{R}^p} \left(\text{MOM}_K \left[\ell_t - \ell_{\tilde{t}} \right] + \lambda (\|t\|_{\text{SLOPE}} - \|\tilde{t}\|_{\text{SLOPE}}) \right) , \tag{22}$$

where $\ell_t : (x, y) \in \mathbb{R}^p \times \{-1, 1\} = \log(1 + \exp(-y\langle x, t \rangle))$ for all $t \in \mathbb{R}^p$.

Let us first compute the complexity function r_2. If $V \subset \mathbb{R}^p$ is closed under permutations and reflections (sign-changes)– which is the case for B_{SLOPE}^p, the unit ball of the SLOPE norm – then $\sup_{v \in V} |\langle \cdot, v \rangle|$ is 1-unconditional. Therefore, the dual norm of $\| \cdot \|_{SLOPE}$ is 1-unconditional and Theorem 5 applies provided that X is isotropic and verifies (17). By (Lecué and Mendelson, 2018, Lemma 5.3), we have

$$\begin{aligned} \mathbb{E} \sup_{t \in \rho B_{\text{SLOPE}}^p \cap r B_2^p} \left| \sum_{i \in J} \sigma_i \langle X_i, t \rangle \right| &\leq c(C_0, M) \sqrt{|J|} w(\rho B_{\text{SLOPE}}^p \cap r B_2^p) \\ &\leq c(C_0, M) \sqrt{|J|} \begin{cases} r\sqrt{p} & \text{if } r \leq \rho/\sqrt{p} \\ \rho & \text{if } r \geq \rho/\sqrt{p} \end{cases} \end{aligned} \tag{23}$$

It follows that

$$r_2^2(\gamma, \rho) = c(C_0, \gamma, M) \begin{cases} \frac{p}{N} & \text{if } p \leq c(C_0, \gamma, M)\rho\sqrt{N} \\ \frac{\rho}{\sqrt{N}} & \text{if } p \geq c(C_0, \gamma, M)\rho\sqrt{N}. \end{cases}$$

Let us turn to the local Bernstein Assumption. Since the loss function is the same as the one used in Section 6.2, the local Bernstein assumption holds if there exists $c_0 > 0$ such that

$$\mathbb{P}(|\langle X, t^* \rangle| \leq c_0) \geq 1 - \frac{1}{2(C')^{(2+2\varepsilon)/\varepsilon}}$$

where $C' = c(M, C_0)$ is a function of M and C_0 only. The constant A in the Bernstein condition depends on c_0, C_0 and M. As for the LASSO, since the formula of A is complicated (given in Proposition 1), we write A all along this section.

A solution to the sparsity equation relative to the SLOPE norm can be found in Lecué and Mendelson (2018). We recall this result here.

Lemma 4 *(Lecué and Mendelson, 2018, Lemma 4.3) Let $1 \leq s \leq p$ and set $\mathcal{B}_s = \sum_{i \leq s} \beta_i/\sqrt{i}$. If t^* can be decomposed as $t^* = u + v$ with $u \in (\rho/20)B_{SLOPE}^p$ and v is s-sparse and if $40\mathcal{B}_s \leq \rho/\sqrt{C_{K,r}(\rho, A)}$ then $\Delta(\rho) \geq 4\rho/5$.*

Assume that t^* is exactly s-sparse, so Lemma 4 applies. We consider two cases depending on K. Consider the case where $K \leq c(c_0, C_0, M)Nr_2^2(\gamma, \rho^*)$, so $\sqrt{C_{K,r}(\rho^*, A)} = r_2(\gamma, \rho^*)$. For $\beta_j = c\sqrt{\log(ep/j)}$, one may show that $\mathcal{B}_s = c\sqrt{s \log(ep/s)}$ (see Bellec et al. (2018); Lecué and Mendelson (2018)). From (23) and Lemma 4, it follows that we can choose

$$\rho^* = c(c_0, M, C_0)s\frac{\log(ep/s)}{\sqrt{N}} \quad \text{and thus} \quad r_2^2(\gamma, \rho^*) = c(c_0, M, C_0)\frac{s\log(ep/s)}{N} . \tag{24}$$

For $C_{K,r}(\rho, A) = c(c_0, M, C_0)K/N$ holding when $K \geq c(c_0, C_0, M)Nr_2^2(\gamma, \rho^*)$, we take $\rho^* = c(c_0, C_0, M)\sqrt{sK/N}$ satisfying the sparsity equation. We can therefore apply Theorem 2 for

$$\rho^* = c(c_0, M, C_0) \max(s\sqrt{\log{(ep/s)}/N}, \sqrt{sK}/\sqrt{N}) .$$

Theorem 7 *Let $\varepsilon > 0$ and (X, Y) be random variable with values in $\mathbb{R}^p \times \{\pm 1\}$ such that X is an isotropic random vector such that for all $1 \leq j \leq p$ and $1 \leq q \leq C_0 \log(p)$, $\|\langle X, e_j \rangle\|_{L_q} \leq M\sqrt{q}$ with $C_0 \log(p) \geq 2+\varepsilon$. Let $f^* : x \in \mathbb{R}^p \mapsto \langle x, t^* \rangle$ be the oracle where $t^* \in \mathbb{R}^p$ is s-sparse. Assume also that the oracle satisfies Equation (21). Assume that $(X, Y), (X_i, Y_i)_{i \in \mathcal{I}}$ are i.i.d and $N \geq c \log(ep/s)$. Let $K \geq 7|\mathcal{O}|/3$. Let $\hat{t}_{\lambda,K}$ be the minmax MOM logistic Slope procedure introduced in (22) for the choice of weights $\beta_j = \sqrt{\log(ep/j)}$, $j = 1, \ldots, p$ and regularization parameter $\lambda = c(c_0, M, C_0) \max(1/\sqrt{N}, \sqrt{K/(sN)})$. With probability larger than $1 - 2\exp(-cK)$,*

$$\|\hat{t}_{\lambda,K} - t^*\|_{SLOPE} \leq c(c_0, M, C_0) \max\left(s\sqrt{\frac{\log(ep/s)}{N}}, \sqrt{s}\sqrt{\frac{K}{N}} \right),$$

$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(c_0, M, C_0) \max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N} \right) ,$$

$$P\mathcal{L}_{\hat{t}_{\lambda,K}} \leq c(c_0, M, C_0) \max\left(\frac{K}{N}, s\frac{\log(ep/s)}{N} \right) .$$

For $K \leq c(c_0, M, C_0) s \log(ep/s)/N$, the parameter λ is independent from the unknown sparsity s and these bounds match the minimax rates of convergence over the class of s-sparse vectors in $\mathbb{R}^p$ without any restriction on s Bellec et al. (2018). Ultimately, one can use a Lepski's adaptation method to chose in a data-driven way the number of blocks K as in Lecué and Lerasle (2019) to achieve these optimal rates without prior knowledge on the sparsity s.

6.4 The minmax MOM Huber Group-Lasso

In this section, we consider regression problems where $\mathcal{Y} = \mathbb{R}$. We consider group sparsity as notion of low-dimensionality for t^*. This setup is particularly useful when features (i.e. coordinates of X) are organized by blocks, as when one constructs dummy variables from a categorical variable.

The regularization norm used to induce this type of "structured sparsity" is called the Group LASSO (see, for example Yang and Zou (2015) and Meier et al. (2008)). It is built as follows: let $G_1, \cdots, G_M$ be a partition of $\{1, \cdots, p\}$ and define, for any $t \in \mathbb{R}^p$

$$\|t\|_{\text{GL}} = \sum_{k=1}^{M} \|t_{G_k}\|_2 \quad . \tag{25}$$

Here, for all $k = 1, \ldots, M$, t_{G_k} denotes the orthogonal projection of t onto the linear Span$(e_i, i \in G_k)$ – $(e_1, \ldots, e_p)$ being the canonical basis of $\mathbb{R}^p$.

The estimator we consider is the minmax MOM Huber Group-LASSO defined, for all $K \in \{1, \cdots, N\}$ and $\lambda > 0$, by

$$\hat{t}_{\lambda,K} \in \underset{t \in \mathbb{R}^p}{\text{argmin}} \sup_{\tilde{t} \in \mathbb{R}^p} \left(\text{MOM}_K \left[\ell_t - \ell_{\tilde{t}}\right] + \lambda(\|t\|_{GL} - \|\tilde{t}\|_{GL}) \right) ,$$

where $t \in \mathbb{R}^p \to \ell_t$ is the Huber loss function with parameter $\delta > 0$ defined as

$$\ell_t(X_i, Y_i) = \begin{cases} \frac{1}{2}(Y_i - \langle X_i, t \rangle)^2 & \text{if } |Y_i - \langle X_i, t \rangle| \leq \delta \\ \delta|Y_i - \langle X_i, t \rangle| - \frac{\delta^2}{2} & \text{if } |Y_i - \langle X_i, t \rangle| > \delta \end{cases} .$$

In particular, it is a Lipschitz loss function with $L = \delta$. Estimation bounds and oracle inequalities satisfied by $\hat{t}_{\lambda,K}$ follow from Theorem 2 as long as we can compute the complexity function r_2, we verify the local Bernstein Assumption and we find a radius ρ^* satisfying the sparsity equation. We now handle these problems starting with the computation of the complexity function r_2.

The dual norm of $\|\cdot\|_{\text{GL}}$ is $z \in \mathbb{R}^p \to \|z\|_{\text{GL}}^* = \max_{1\le k\le M} \|z_{G_k}\|_2$, it is not κ-unconditional with respect to the canonical basis $(e_i)_{i=1}^p$ of $\mathbb{R}^p$ for some absolute constant κ, so Theorem 5 does not apply directly. Therefore, in order to avoid long and technical materials on the rearrangement of empirical means under weak moment assumptions for the computation of the local Rademacher complexity from (10), we simply assume that the design vectors $(X_i)_{i\in\mathcal{I}}$ are L_0-subgaussian and isotropic: for all $i \in \mathcal{I}$, all $t \in \mathbb{R}^p$ and all $q \ge 1$

$$\big\|\langle X_i, t\rangle\big\|_{L_q} \le L_0\sqrt{q}\,\big\|\langle X_i, t\rangle\big\|_{L_2} \text{ and } \big\|\langle X_i, t\rangle\big\|_{L_2} = \|t\|_2\,. \tag{26}$$

In that case, a direct chaining argument allows to bound Rademacher processes by the Gaussian processes (see Talagrand (2014) for chaining methods):

$$\mathbb{E} \sup_{t\in\rho B_{\text{GL}}^p \cap rB_2^p} \left|\sum_{j\in J} \sigma_j \langle t, X_j\rangle\right| \le c(L_0)\sqrt{J}w(\rho B_{\text{GL}}^p \cap rB_2^p) \ .$$

Here, B_{GL}^p is the unit ball of $\|\cdot\|_{\text{GL}}$, $w(\rho B_{\text{GL}}^p \cap rB_2^p)$ is the Gaussian mean width of the interpolated body $\rho B_{\text{GL}}^p \cap rB_2^p$. It follows from the proof of Proposition 6.7 in Bellec et al. (2017) that when the M groups $G_1, \dots, G_M$ are all of same size p/M we have

$$w(\rho B_{\text{GL}}^p \cap rB_2^p) \le \begin{cases} c\rho\sqrt{\frac{p}{M} + \log\left(\frac{Mr^2}{\rho^2}\right)} & \text{if } 0 < \rho \le r\sqrt{M} \\ cr\sqrt{p} & \text{if } \rho \ge r\sqrt{M} \end{cases} .$$

This yields

$$r_2^2(\gamma,\rho) = c(\delta, L_0, \gamma) \begin{cases} \frac{\rho}{\sqrt{N}}\sqrt{\frac{p}{M} + \log\left(\frac{Mr^2}{\rho^2}\right)} & \text{if } 0 < c(\delta, L_0, \gamma)\frac{\rho}{r} \le \sqrt{M} \\ \frac{r}{\sqrt{N}}\sqrt{p} & \text{if } c(\delta, L_0, \gamma)\frac{\rho}{r} \ge \sqrt{M} \end{cases} . \tag{27}$$

Let us now turn to the local Bernstein Assumption. We need to verify Assumption 11. As we assumed that the design vectors $(X_i)_{i\in\mathcal{I}}$ are isotropic and L_0-subgaussian, it is clear that the point a) in Assumption 11 holds with $C' = L_0$. Let us take $\varepsilon = 2$ (another choice would only change the constant). For the point b), we assume that there exists $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \le 2L_0^2\sqrt{C_{K,r}(\rho, 4/\alpha)}$, $F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta) \geqslant \alpha$. Under these conditons, the local Bernstein Assumption is verified for $A = 4/\alpha$ according to Proposition 2.

Finally, we turn to the sparsity equation. The following lemma is an extension of Lemma 3 to the Group Lasso norm.

Lemma 5 *Assume that X is isotropic. Assume that $t^* = u + v$ where $\|u\|_{GL} \le \rho/20$ and v is group-sparse i.e $v_{G_k} = 0$ for all $k \notin I$ for some $I \subset \{1, \dots, M\}$. If $100|I| \le (\rho/\sqrt{C_{K,r}(\rho, 4/\alpha)})^2$, then $\Delta(\rho) \ge 4\rho/5$.*

Proof Let us define $r(\rho) := \sqrt{C_{K,r}(\rho, 4/\alpha)}$ and recall that

$$\tilde{\Delta}(\rho, 4/\alpha) = \inf_{w \in \rho S_{GL} \cap r(\rho) B_2^p} \sup_{z^* \in \Gamma_{t^*}(\rho)} \left\langle z^*, w \right\rangle \ .$$

Here, S_{GL} is the unit sphere of $\|\cdot\|_{GL}$ and $\Gamma_{t^*}(\rho)$ is the union of all sub-differentials $(\partial \|\cdot\|_{GL})_v$ for all $v \in t^* + (\rho/20) B_{GL}^p$. We want to find a condition on $\rho > 0$ insuring that $\tilde{\Delta}(\rho, 4/\alpha) \geq 4\rho/5$.

Let w be a vector in $\mathbb{R}^p$ such that $\|w\|_{GL} = \rho$ and $\|w\|_2 \leq r(\rho)$. We construct $z^* \in \mathbb{R}^p$ such that $z^*_{G_k} = w_{G_k} / \|w_{G_k}\|_2$ if $k \notin I$ (so that $\left\langle z^*_{G_k}, w_{G_k} \right\rangle = \|w_{G_k}\|_2$ for all $k \notin I$) and $z^*_{G_k} = v_{G_k} / \|v_{G_k}\|_2$ if $k \in I$ (so that $\left\langle z_{G_k}, v_{G_k} \right\rangle = \|v_{G_k}\|_2$ for all $k \in I$). We have $\left\|z^*_{G_k}\right\|_2 = 1$ for all $k \in [M]$, so $\|z^*\|^*_{GL} = 1$ (i.e. z^* is in the dual sphere of $\|\cdot\|_{GL}$) and $\left\langle z^*, v \right\rangle = \|v\|_{GL}$ (i.e. z^* is norming for v). Therefore, it follows from (18) that $z^* \in (\partial \|\cdot\|_{GL})_v$. Moreover, $\|u\|_{GL} \leq \rho/20$ hence $v \in t^* + (\rho/20) B_{GL}^p$ and so $z^* \in \Gamma_{t^*}(\rho)$. Furthermore, for this choice of sub-gradient z^*, we have

$$\begin{aligned} \left\langle z^*, w \right\rangle &= \sum_{k \in I} \left\langle z^*_{G_k}, w_{G_k} \right\rangle + \sum_{k \notin I} \left\langle z^*_{G_k}, w_{G_k} \right\rangle \geq -\sum_{k \in I} \|w_{G_k}\|_2 + \sum_{k \notin I} \|w_{G_k}\|_2 \\ &= \sum_{k=1}^M \|w_{G_k}\|_2 - 2 \sum_{k \in I} \|w_{G_k}\|_2 \geq \rho - 2\sqrt{|I|}\, r(\rho) \ . \end{aligned}$$

In the last inequality, we used that $\|w\|_{GL} = \rho$ and that

$$\sum_{k \in I} \|w_{G_k}\|_2 \leq \sqrt{|I|} \sqrt{\sum_{k \in I} \|w_{G_k}\|_2^2} \leq \sqrt{|I|}\, \|w\|_2 \leq \sqrt{|I|}\, r(\rho).$$

Then $\left\langle z^*, w \right\rangle \geq 4\rho/5$ when $\rho - 2\sqrt{|I|}\, r(\rho) \geq 4\rho/5$ which happens to be true when $100|I| \leq (\rho/r(\rho))^2$. ■

Assume that t^* is exactly s-group sparse, so Lemma 5 applies. We consider two cases depending on the value of K. When $K \leq c(L_0, \alpha, \delta) N r_2^2(\gamma, \rho^*)$, $\sqrt{C_{K,r}(\rho^*, 4/\alpha)} = r_2(\gamma, \rho^*)$. By Lemma 5 and (27), it follows that (for equal size blocks), one can choose

$$\rho^* = c(L_0, \alpha, \delta) \frac{s}{\sqrt{N}} \sqrt{\frac{p}{M} + \log M} \quad \text{and thus} \quad r^2(\gamma, \rho^*) = c(L_0, \alpha, \delta) \frac{s}{N} \left(\frac{p}{M} + \log M \right) \ . \tag{28}$$

This result has a similar flavor as the one for the Lasso. The term $s' = sp/M$ equals *block sparsity* $\times$ *size of each blocks*, i.e to the total number of non-zero coordinates in t^*: $s' = \|t^*\|_0$. Replacing the sparsity s' by sp/M in Theorem 6, we would have obtained $\rho^* = c(L_0, \alpha, \delta)(sp/M)\sqrt{\log(p)/N}$ which is larger than the bound obtained for the Group Lasso in Equation (28). It is therefore better to induce the sparsity by blocks instead of just coordinate-wise when we are aware of such block-structured sparsity. In the other case, when $K \leq c(L_0, \alpha, \delta) N r_2^2(\gamma, \rho^*)$, we have $\sqrt{C_{K,r}(\rho^*, 4/\alpha)} = c(L_0, \alpha, \delta)\sqrt{K/N}$ and so one can take $\rho^* = c(L_0, \alpha, \delta)\sqrt{sK/N}$. We can therefore apply Theorem 2 with

$$\rho^* = c(L_0, \alpha, \delta) \max\left(\frac{s}{\sqrt{N}} \sqrt{\frac{p}{M} + \log(M)}, \sqrt{s}\sqrt{\frac{K}{N}} \right) \ .$$

Theorem 8 *Let (X, Y) be a random variables with values in $\mathbb{R}^p \times \mathbb{R}$ such that $Y \in L_1$ and X is an isotropic and L_0-subgaussian random vector in $\mathbb{R}^p$. Assume that $(X, Y), (X_i, Y_i)_{i \in \mathcal{I}}$ are i.i.d. Let $f^*(\cdot) = \langle t^*, \cdot \rangle$ for some $t^* \in \mathbb{R}^p$ which is s-group sparse with respect to equal-size groups $(G_k)_{k=1}^M$. Let $K \geq 7|\mathcal{O}|/3$ and $N \geq cs(p/M + \log(M))$. Assume that there exists $\alpha > 0$ such that, for all $x \in \mathbb{R}^p$ and all $z \in \mathbb{R}$ satisfying $|z - \langle t^*, x \rangle| \leq 2L_0^2 \sqrt{C_{K,r}(2\rho^*, 4/\alpha)}$, $F_{Y|X=x}(\delta + z) - F_{Y|X=x}(z - \delta) \geq \alpha$ (where $F_{Y|X=x}$ is the cumulative ditribution function of Y given $X = x$). With probability larger than $1 - 2\exp(-cK)$, the MOM Huber group-LASSO estimator $\hat{t}_{\lambda,K}$ for*

$$\lambda = c(L_0, \alpha, \delta) \max\left(\frac{1}{\sqrt{N}} \sqrt{\frac{p}{M} + \log M}, \sqrt{\frac{K}{sN}} \right)$$

satisfies

$$\|\hat{t}_{\lambda,K} - t^*\|_{GL} \leq c(L_0, \alpha, \delta) \max\left(\frac{s}{\sqrt{N}} \sqrt{\frac{p}{M} + \log(M)}, \sqrt{s}\sqrt{\frac{K}{N}} \right),$$

$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(L_0, \alpha, \delta) \max\left(\frac{s}{N} \left(\frac{p}{M} + \log(M) \right), \frac{K}{N} \right),$$

$$P\mathcal{L}_{\hat{t}_{\lambda,K}} \leq c(L_0, \alpha, \delta) \max\left(\frac{s}{N} \left(\frac{p}{M} + \log(M) \right), \frac{K}{N} \right) \ .$$

For $K \leq c(L_0, \alpha, \delta)s(p/M + \log M)$, the regularization parameter λ is independent from the unknown group sparsity s (the choice of K can be done in data-driven way using either a Lepski method or a MOM cross validation as in Lecué and Lerasle (2019)). In the ideal i.i.d. setup (with no outliers), the same result holds for the RERM as we assumed that the class $F - f^*$ is L_0-subgaussian and for the choice of regularization parameter $\lambda = c(L_0, \alpha, \delta)(\sqrt{p/(NM)} + \sqrt{\log(M)/N})$. The minmax MOM estimator has the advantage to be robust up to $c(L_0, \alpha, \delta)s(p/M + \log M)$ outliers in the dataset.

6.5 Huber regression with total variation penalty

In this section, we investigate another type of structured sparsity with which be induced by the total variation norm. Given $t \in \mathbb{R}^p$, the Total Variation norm Osher et al. (2005) is defined as

$$\|t\|_{TV} = |t_1| + \sum_{i=1}^{p-1} |t_{i+1} - t_i| = \|Dt\|_1, \quad \text{where} \quad D = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{p \times p} \ . \quad (29)$$

The total variation norm favors vectors such that their "discrete gradient Dt is sparse" that is piecewise constant vectors t.

The estimator considered in this section is the minmax MOM Huber TV regularization defined for all $\lambda > 0$ and $K \in \{1, \cdots, N\}$ as

$$\hat{t}_{\lambda,K} \in \underset{t \in \mathbb{R}^p}{\operatorname{argmin}} \sup_{\tilde{t} \in \mathbb{R}^p} \left(\text{MOM}_K \left[\ell_t - \ell_{\tilde{t}} \right] + \lambda(\|t\|_{TV} - \|\tilde{t}\|_{TV}) \right) \ ,$$

where the loss ℓ is the Huber loss. Let $\delta > 0$,

$$\ell_t(X_i, Y_i) = \begin{cases} \frac{1}{2}(Y_i - \langle X_i, t \rangle)^2 & \text{if } |Y_i - \langle X_i, t \rangle| \leq \delta \\ \delta |Y_i - \langle X_i, t \rangle| - \frac{\delta^2}{2} & \text{if } |Y_i - \langle X_i, t \rangle| > \delta \end{cases} .$$

Statistical bounds for $\hat{t}_{\lambda,K}$ follows from Theorem 2 and the computation of r_2, ρ^* and the study of the local Bernstein assumption. We start with the computation of the complexity function r_2. Simple computations yield that the dual norm of $\| \cdot \|_{TV}$ is $z \in \mathbb{R}^p \mapsto \|z\|_{TV}^* = \|(D^{-1})^T z\|_\infty = \max_{1 \leq k \leq p} |\sum_{i=1}^k z_i|$ which is not κ-unconditional with respect to the canonical basis $(e_i)_{i=1}^p$ of $\mathbb{R}^p$ for some absolute constant κ. Therefore, Theorem 5 does not apply directly. To upper bound the Rademacher complexity from (10), we assume that the design vectors $(X_i)_{i \in \mathcal{I}}$ are L_0-subgaussian and isotropic (see Equation (26)) as in Section 6.4. A direct chaining argument allows to bound the Rademacher complexity by the Gaussian mean width (see Talagrand (2014) for chaining methods):

$$\mathbb{E} \sup_{t \in \rho B_{\text{TV}}^p \cap r B_2^p} \left| \sum_{j \in J} \sigma_j \langle t, X_j \rangle \right| \leq c(L_0) \sqrt{J} w(\rho B_{\text{TV}}^p \cap r B_2^p)$$

A recent result from Chatterjee and Goswami (2019) allows to compute the Gaussian Mean-Width of B_{TV}^p.

Lemma 1 (Lemma 4.5 Chatterjee and Goswami (2019)) *For any $p \geq 2$, there exists a universal constant $c > 0$ such that*

$$w(B_{TV}^p) \leq c\big(\log(p)\log(1+2p) + 1\big) \leq c \log^2(p) \quad . \tag{30}$$

It follows from Equation (30) that $w(\rho B_{\text{TV}}^p \cap r B_2^p) \leq \min(\rho w(B_{\text{TV}}^p), r w(B_2^p)) \leq c \min(\rho \log^2(p), r\sqrt{p})$ and so one can take

$$r_2^2(\gamma, \rho) = c(\delta, L_0, \gamma) \min\left(\frac{\rho \log^2(p)}{\sqrt{N}}, \frac{p}{N} \right) .$$

Let us now turn to the local Bernstein Assumption. The loss function and the model being the same as the ones in Section 6.4 the Bernstein Assumption is verified with a constant $A = 4/\alpha$, if there exists a constant $\alpha > 0$ such that for all $x \in \mathcal{X}$ and all $z \in \mathbb{R}$ satisfying $|z - f^*(x)| \leq 2L_0^2 \sqrt{C_{K,r}(\rho, 4/\alpha)}$, $F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta) \geqslant \alpha$.

Let us turn to the sparsity equation. The following Lemma solves the sparsity equation for the TV regularization.

Lemma 6 *Let us assume that X is isotropic . If the oracle t^* can be decomposed as $t^* = v + u$ for $u \in (\rho/20) B_{TV}^p$ and $400s \leq (\rho/\sqrt{C_{K,r}(\rho, 4/\alpha)})^2$, then $\Delta(\rho) \geq 4\rho/5$, where $s = |supp(Dv)|$.*

Compared with Lemma 3, sparsity in Lemma 6 is granted on the linear transformation Dt^* (also called discrete gradient of t^*) rather than on the oracle t^*.

Proof Let us denote $\sqrt{C_{K,r}(\rho, 4/\alpha)} := r(\rho)$. Let us recall that

$$\tilde{\Delta}(\rho, 4/\alpha) = \inf_{w \in \rho S_{TV} \cap r(\rho) B_2^p} \sup_{z^* \in \Gamma_{t^*}(\rho)} \langle z^*, w \rangle$$

where S_{TV} is the unit sphere of $\|\cdot\|_{TV}$ and $\Gamma_{t^*}(\rho)$ is the union of all sub-differentials $(\partial \|\cdot\|_{TV})_v$ for all $v \in t^* + (\rho/20) B_{TV}^p$. We want to find a condition on $\rho > 0$ insuring that $\tilde{\Delta}(\rho, 4/\alpha) \geq 4\rho/5$.

Recall that the oracle t^* can be decomposed as $t^* = u + v$, where $u \in (\rho/20)B_{TV}$ and thus $\|t^* - v\|_{TV} \le \rho/20$. Let I denote the support of Dv and s its cardinality. Let I^C be the complementary of I. Let $w \in \rho S_{TV}^p \cup r(\rho)B_2^p$.

We construct $z^* = D^T u^*$, such for all i in I, $u_i^* = sign\big((Dv)_i\big)$ and for all i in I^C, $u_i^* = sign\big((Dw)_i\big)$. Such a choice of z^* implies that $\big\langle z^*, v\big\rangle = \big\langle u^*, Dv\big\rangle = \sum_{i\in I} sign\big((Dv)_i\big)(Dv)_i = \|v\|_{TV}$ i.e z^* is norming for v. Moreover, we have $\|z^*\|_{TV}^* = \|(D^{-1})^T z^*\|_\infty = \|u^*\|_\infty = 1$ hence $z^* \in S_{TV}^*$. Then it follows from (18) that $z^* \in (\partial\| \cdot \|_{TV})_v$ and since $u \in (\rho/20)B_{TV}$ we have $z^* \in \Gamma_{t^*}(\rho)$.

Now let us denote by $P_I w$ the orthogonal projection of w onto Span$(e_i, i \in I)$. From the choice of z^* we get

$$\begin{aligned}\big\langle z^*, w\big\rangle = \big\langle D^T u^*, w\big\rangle = \big\langle u^*, Dw\big\rangle &= \big\langle u^*, P_I Dw\big\rangle + \big\langle u^*, P_{I^C} Dw\big\rangle \\ &\ge -\|P_I Dw\|_1 + \|P_{I^C} Dw\|_1 = \|Dw\|_1 - 2\|P_I Dw\|_1\end{aligned}$$

Moreover we have $\|P_I Dw\|_1 \le \sqrt{s}\|P_I Dw\|_2 \le \sqrt{s}\|Dw\|_2$ and

$$\|Dw\|_2 = \|(I_P + D^-)w\|_2 \le \|w\|_2 + \|D^- w\|_2 \le 2\|w\|_2 \le 2r(\rho) \ ,$$

where

$$D^- = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ -1 & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & -1 & 0 \end{bmatrix} \ .$$

Since $\|Dw\|_1 = \|w\|_{TV} = \rho$, we get

$$\Delta(\rho) \ge \rho - 4\sqrt{s}r(\rho) \ge \frac{4\rho}{5}$$

when $\rho \ge 20\sqrt{s}r(\rho)$. ■

Let us now identify a radius ρ^* satisfying the sparsity equation using Lemma 6. We place ourselves under the assumption from Lemma 6 that is when t^* is such that Dt^* is approximately s-sparse. There are two cases to study according to the value of K. For the case where $\sqrt{C_{K,r}(\rho^*, 4/\alpha)} = r_2(\gamma, \rho^*)$– which holds when $K \le c(L_0, \alpha, \delta)Nr_2^2(\gamma, \rho^*)$– , we can take

$$\rho^* = c(L_0, \alpha, \delta)\frac{s\log^2(p)}{\sqrt{N}} \quad \text{and} \quad r_2^2(\gamma, \rho^*) = c(L_0, \alpha, \delta)\frac{s\log^4(p)}{N}.$$

For $C_{K,r}(\rho^*, 4/\alpha) = c(L_0, \alpha, \delta)K/N$– which holds when $K \ge c(L_0, \alpha, \delta)Nr_2^2(\gamma, \rho^*)$– we can take $\rho^* = c(L_0, \alpha, \delta)\sqrt{sK/N}$. We can therefore apply Theorem 2 with

$$\rho^* = c(L_0, \alpha, \delta)\max\left(\frac{s\log^2(p)}{\sqrt{N}}, \sqrt{sK/N}\right) \ .$$

To simplify the presentation, we assume that Dt^* is exactly s-sparse. We may only assume it is approximatively s-sparse using the more involved formalism of Lemma 6.

Theorem 9 *Let (X, Y) be a random variables with values in $\mathbb{R}^p \times \mathbb{R}$ such that $Y \in L_1$ and X is an isotropic and L_0-subgaussian random vector in $\mathbb{R}^p$. Assume that $(X,Y), (X_i, Y_i)_{i\in\mathcal{I}}$ are i.i.d. Let $f^*(\cdot) = \langle t^*, \cdot \rangle$, where t^* is such that Dt^* is s-sparse. Let $K \geq 7|\mathcal{O}|/3$ and $N \geq cs\log^4(p)$. Assume that there exists $\alpha > 0$ such that, for all $x \in \mathbb{R}^p$ and all $z \in \mathbb{R}$ satisfying $|z - \langle t^*, x\rangle| \leq 2L_0^2\sqrt{C_{K,r}(2\rho^*, 4/\alpha)}$, $F_{Y|X=x}(\delta + z) - F_{Y|X=x}(z-\delta) \geq \alpha$ (where $F_{Y|X=x}$ is the cumulative distribution function of Y given $X = x$). With probability larger than $1 - 2\exp(-cK)$, the MOM Huber TV estimator $\hat{t}_{\lambda,K}$ for*

$$\lambda = c(L_0, \alpha, \delta) \max\left(\frac{\log^2(p)}{\sqrt{N}}, \sqrt{\frac{K}{sN}}\right)$$

satisfies

$$\|\hat{t}_{\lambda,K} - t^*\|_{TV} \leq c(L_0,\alpha,\delta)\max\left(s\frac{\log^2(p)}{\sqrt{N}}, \sqrt{s}\sqrt{\frac{K}{N}}\right)$$
$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(L_0,\alpha,\delta)\max\left(s\frac{\log^4(p)}{N}, \frac{K}{N}\right) \ ,$$
$$P\mathcal{L}_{\hat{t}_{\lambda,K}} \leq c(L_0,\alpha,\delta)\max\left(s\frac{\log^4(p)}{N}, \frac{K}{N}\right) \ .$$

For $K \leq c(L_0,\alpha,\delta)s\log^4(p)$, the regularization parameter λ is independent from the unknown sparsity s and $\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(L_0,\alpha,\delta)s\log^4(p)/N$ which is the right order for s-sparse regression problems (up to the $\log^2 p$ term). Without any outliers the same conclusion (for $K = c(L_0,\alpha,\delta)s\log^4(p)$) holds also for the RERM for $\lambda = c(L_0,\alpha,\delta)\log^2(p)/\sqrt{N}$ since the Assumptions on the design X imply that the class $F - f^*$ is L_0-subgaussian. The minmax MOM estimator has the advantage to be robust up to $s\log^4(p)$ outliers.

6.6 Huber Fused Lasso

In this section, we investigate the fusion of two sparsity structures which will be induced by a mixture of the Total Variation and ℓ_1 norms. The resulting regularization norm is known as the fused Lasso (see Tibshirani et al. (2005)) defined for some mixture parameters $\beta, \eta > 0$ for all $t \in \mathbb{R}^p$ by

$$\|t\|_{FL} = \eta\|t\|_1 + \beta\|t\|_{TV} \ .$$

This type of norm is expected to promote signals having both a small number of non-zero coefficients (thanks to the ℓ_1^p-norm) and a sparse discrete gradient (thanks to the TV norm) i.e. sparse and constant by blocks signals.

The estimator we consider in this section is defined for all $\lambda > 0$ and $K \in \{1, \cdots, N\}$ as

$$\hat{t}_{\lambda,K} \in \underset{t\in\mathbb{R}^p}{\operatorname{argmin}} \sup_{\tilde{t}\in\mathbb{R}^p}\left(\text{MOM}_K\left[\ell_t - \ell_{\tilde{t}}\right] + \lambda(\|t\|_{FL} - \|\tilde{t}\|_{FL})\right) \ ,$$

where $t \in \mathbb{R}^p \to \ell_t$ denotes the Huber loss function with parameter $\delta > 0$.

Theorem 2 may be used to derive robust statistical bounds for $\hat{t}_{\lambda,K}$. To do so we first have to compute a complexity function r_2. The dual norm of $\|\cdot\|_{FL}$ denoted by $z \in \mathbb{R}^p \mapsto \|z\|_{FL}^*$ is not K-unconditional with respect to the canonical basis $(e_i)_{i=1}^p$ of $\mathbb{R}^p$; so Theorem 5 does not apply directly. To compute the Rademacher complexity from (10), we assume that the design vectors $(X_i)_{i\in\mathcal{I}}$ are subgaussian and isotropic as in Section 6.4 and 6.5. A direct chaining argument allows to bound the Rademacher complexity by the Gaussian mean width (see Talagrand (2014) for chaining methods):

$$\mathbb{E}\sup_{t\in\rho B_{\text{FL}}^p\cap rB_2^p}\left|\sum_{j\in J}\sigma_j\langle t, X_j\rangle\right| \leq c(L_0)\sqrt{|J|}w(\rho B_{\text{FL}}^p \cap rB_2^p)$$

where B_{FL}^p denote the unit ball for $\|\cdot\|_{FL}$ in $\mathbb{R}^p$. Since $B_{FL}^p\subset\eta^{-1}B_1^p$ and $B_{FL}^p\subset\beta^{-1}B_{TV}^p$, it follows from Sections 6.2 and 6.5 that

$$w(\rho B_{FL}^p\cap rB_{L_2^p}) \leq c(\delta, L_0, \gamma)\min\left(\rho\log^2(p)\min\left(\frac{1}{\eta},\frac{1}{\beta}\right), r\sqrt{p}\right) \tag{31}$$

(to simplify the presentation of the results, we used the same upper bound on the Gaussian mean widths of both the ℓ_1^p and TV unit balls that is $w(B_1^p), w(B_{TV}^p)\leq c\log^2(p)$ instead of using the $\sqrt{\log p}$ sharper bound for $w(B_1^p)$). We can therefore take

$$r_2^2(\gamma,\rho) = c(\delta, L_0,\gamma)\min\left(\frac{\rho\log^2(p)}{\sqrt{N}}\min\left(\frac{1}{\eta},\frac{1}{\beta}\right), \frac{p}{N}\right). \tag{32}$$

Let us turn to the local Bernstein Assumption. The loss function and the model being the same as the ones in Section 6.4 the local Bernstein Assumption is verified with a constant $A = 4/\alpha$, if there exists a constant $\alpha > 0$ such that for all $x\in\mathcal{X}$ and all $z\in\mathbb{R}$ satisfying $|z - f^*(x)|\leq 2L_0^2\sqrt{C_{K,r}(\rho, 4/\alpha)}$, $F_{Y|X=x}(z+\delta) - F_{Y|X=x}(z-\delta)\geqslant\alpha$.

Finally, let us turn to the sparsity equation. To that end, we will use the following result.

Lemma 7 *Assume that X is isotropic and that t^* can be decomposed as $t^* = v + u$, with $u\in(\rho/20)B_{FL}^p$ for some $\rho > 0$. Let $s_1 = |\text{supp}(v)|$ and $s_{TV} = |\text{supp}(Dv)|$. If $\rho\geq\big(10\eta\sqrt{s_1} + 20\beta\sqrt{s_{TV}}\big)\sqrt{C_{K,r}(\rho, 4/\alpha)}$ then $\Delta(\rho)\geq 4\rho/5$ (we recall that η and β are the mixture parameters of the Fused Lasso norm).*

Proof Let us denote $\sqrt{C_{K,r}(\rho,4/\alpha)} := r(\rho)$. Let us recall that

$$\tilde{\Delta}(\rho, 4/\alpha) = \inf_{w\in\rho S_{FL}\cap r(\rho)B_2^p}\ \sup_{z^*\in\Gamma_{t^*}(\rho)}\langle z^*, w\rangle$$

where S_{FL} is the unit sphere of $\|\cdot\|_{FL}$ and $\Gamma_{t^*}(\rho)$ is the union of all sub-differentials $(\partial\|\cdot\|_{FL})_v$ for all $v\in t^* + (\rho/20)B_{FL}^p$. We want to find a condition on $\rho > 0$ insuring that $\tilde{\Delta}(\rho, 4/\alpha)\geq 4\rho/5$.

For all $t\in\mathbb{R}^p$, the sub-differential of the fused Lasso norm is given by (see Theorem 23.8 in Rockafellar (1997)) $(\partial\|\cdot\|_{FL})_t = \eta(\partial\|\cdot\|_1)_t + \beta(\partial\|\cdot\|_{TV})_t$ and we also recall that $(\partial\|\cdot\|_{TV})_t = D^\top\partial\|\cdot\|_1(Dt)$.

Let $w\in\rho S_{FL}\cap r(\rho)B_2^p$ and let us denote $I = \text{supp}(v)$ and $J = \text{supp}(Dv)$. We construct $z^* = \eta u^* + \beta D^T q^*$ such that

$$u_i^* = \begin{cases}\text{sign}(v_i) & \text{for } i\in I\\ \text{sign}(w_i) & \text{for } i\in I^C\end{cases} \text{ and } q_i^* = \begin{cases}\text{sign}((Dv)_i) & \text{for } i\in J\\ \text{sign}((Dw)_i) & \text{for } i\in J^C.\end{cases}$$

We have $\langle u^*, v\rangle = \|v\|_1$ and $u^* \in S_1^*$. Likewise $\langle D^T q^*, v\rangle = \|Dv\|_1$ and $D^T q^* \in S_{TV}^*$. Therefore, $u^* \in (\partial \|\cdot\|_1)_v$ and $D^T q^* \in (\partial \|\cdot\|_{TV})_v$ and it follows that $z^* \in (\partial \|\cdot\|_{FL})_v$. Since $u \in (\rho/20)B_{FL}^p$ we have $z^* \in \Gamma_{t^*}(\rho)$.

Let us denote by $P_I w$ (resp. $P_J w$) the orthogonal projection of w onto Span$(e_i, i \in I)$ (resp. Span$(e_i, i \in J)$). We have

$$\begin{aligned}
\langle z^*, w\rangle &= \eta\langle u^*, w\rangle + \beta\langle D^T q^*, w\rangle \\
&= \eta\langle u^*, P_I w\rangle + \eta\langle u^*, P_{I^C} w\rangle + \beta\langle q^*, P_J Dw\rangle + \beta\langle q^*, P_{J^C} Dw\rangle \\
&\geq -\eta\|P_I w\|_1 + \eta\|P_{I^C} w\|_1 - \beta\|P_J Dw\|_1 + \beta\|P_{J^C} Dw\|_1 \\
&= \eta\|w\|_1 - 2\eta\|P_I w\|_1 + \beta\|Dw\|_1 - 2\beta\|P_J Dw\|_1 \\
&= \rho - 2\eta\|P_I w\|_1 - 2\beta\|P_J Dw\|_1.
\end{aligned}$$

Moreover we have $\|P_J Dw\|_1 \leq \sqrt{s_{TV}}\|P_I Dw\|_2 \leq 2\sqrt{s_{TV}} r(\rho)$. Similarly, $\|P_I w\|_1 \leq \sqrt{s_1} r(\rho)$. Finally

$$\Delta(\rho) \geq \rho - 2\eta\sqrt{s_1} r(\rho) - 4\beta\sqrt{s_{TV}} r(\rho) \geq \frac{4\rho}{5}$$

when $\rho \geq \left(10\eta\sqrt{s_1} + 20\beta\sqrt{s_{TV}}\right) r(\rho)$. ■

We now use Lemma 7 to identify a radius ρ^* satisfying the sparsity equation. There are two cases. For the case where $\sqrt{C_{K,r}(\rho^*, 4/\alpha)} = r_2(\gamma, \rho^*)$– which holds when $K \leq c(L_0, \alpha, \delta) N r_2^2(\gamma, \rho^*)$– , we can take

$$\begin{aligned}
\rho^* &= c(L_0, \alpha, \delta) \frac{\log^2(p)\left(\eta\sqrt{s_1} + \beta\sqrt{s_{TV}}\right)^2 \min\left(\frac{1}{\eta}, \frac{1}{\beta}\right)}{\sqrt{N}} \\
r_2(\gamma, \rho^*) &= c(L_0, \alpha, \delta) \frac{\log^2(p) \min\left(\sqrt{s_1} + \frac{\beta}{\eta}\sqrt{s_{TV}}, \frac{\eta}{\beta}\sqrt{s_1} + \sqrt{s_{TV}}\right)}{\sqrt{N}} \quad .
\end{aligned}$$

By taking $\beta = \min(1, \sqrt{s_1/s_{TV}})$ and $\eta = \min(1, \sqrt{s_{TV}/s_1})$ we get

$$\begin{aligned}
\rho^* &= c(L_0, \alpha, \delta) \frac{\min(s_1, s_{TV})}{\sqrt{N}} \log^2(p) \\
r_2(\gamma, \rho^*) &= c(L_0, \alpha, \delta) \frac{\min(\sqrt{s_1}, \sqrt{s_{TV}})}{\sqrt{N}} \log^2(p) \quad .
\end{aligned}$$

For $C_{K,r}(\rho^*, 4/\alpha) = c(L_0, \alpha, \delta) K/N$ – which holds when $K \geq c(L_0, \alpha, \delta) N r_2^2(\gamma, \rho^*)$ – ,we can take $\rho^* = c(L_0, \alpha, \delta)\sqrt{\min(s_1, s_{TV})K/N}$. We can therefore apply Theorem 2 with

$$\rho^* = c(L_0, \alpha, \delta) \max\left(\frac{\min(s_1, s_{TV})}{\sqrt{N}} \log^2(p), \sqrt{\min(s_1, s_{TV})K/N}\right).$$

To simplify the presentation, we assume that t^* is exactly s_1-sparse and Dt^* is exaclty s_{TV}-sparse. We may only assume that t^* and Dt^* are respectively approximatively s_1 and s_{TV}-sparse using the more involved formalism of Lemma 7.

Theorem 10 *Let (X, Y) be a random variables with values in $\mathbb{R}^p \times \mathbb{R}$ such that $Y \in L_1$ and X is an isotropic and L_0-subgaussian random vector in $\mathbb{R}^p$. Assume that $(X, Y), (X_i, Y_i)_{i\in\mathcal{I}}$ are i.i.d. Let $f^*(\cdot) = \langle t^*, \cdot \rangle$ for t^* in $\mathbb{R}^p$ such that t^* is s_1-sparse and Dt^* is s_{TV}-sparse. Let $K \geq 7|\mathcal{O}|/3$ and $N \geq c \min(s_1, s_{TV}) \log^4(p)$ Assume that there exists $\alpha > 0$ such that, for all $x \in \mathbb{R}^p$ and all $z \in \mathbb{R}$ satisfying $|z - \langle t^*, x\rangle| \leq 2L_0^2\sqrt{C_{K,r}(2\rho^*, 4/\alpha)}$, $F_{Y|X=x}(\delta + z) - F_{Y|X=x}(z - \delta) \geq \alpha$ (where $F_{Y|X=x}$ is the cumulative distribution function of Y given $X = x$). Let us take $\beta = \min(1, \sqrt{s_1/s_{TV}})$ and $\eta = \min(1, \sqrt{s_{TV}/s_1})$. With probability larger than $1 - 2\exp(-cK)$, the MOM Huber FL estimator $\hat{t}_{\lambda,K}$ for*

$$\lambda = c(L_0, \alpha, \delta) \max\left(\frac{\log^2(p)}{\sqrt{N}}, \sqrt{\frac{K}{\min(s_1, s_{TV})N}}\right)$$

satisfies

$$\|\hat{t}_{\lambda,K} - t^*\|_{TV} \leq \rho^* = c(L_0, \alpha, \delta) \max\left(\frac{\min(s_1, s_{TV})}{\sqrt{N}}\log^2(p), \sqrt{K/(\min(s_1, s_{TV})N)}\right)$$

$$\|\hat{t}_{\lambda,K} - t^*\|_2^2 \leq c(L_0, \alpha, \delta) \max\left(\frac{\min(s_1, s_{TV})}{N}\log^4(p), \frac{K}{N}\right) ,$$

$$P\mathcal{L}_{\hat{t}_{\lambda,K}} \leq c(L_0, \alpha, \delta) \max\left(\frac{\min(s_1, s_{TV})}{N}\log^4(p), \frac{K}{N}\right) .$$

For $K \leq c(L_0, \alpha, \delta) \min(s_1, s_{TV}) \log^4(p)$, we obtain a bound depending on $\min(s_1, s_{TV})$. As a consequence, we can use Fused Lasso even when only the sparsity to ℓ_1 (or TV) holds. However, note that we obtained this result by choosing η and β depending on the unknown sparsity s_1 and s_{TV} since the latter bounds automatically adapt to the sparsity structure. If both type of sparsity holds (short support and constant by blocks) only the best one is kept. This remarkable feature of the Fused Lasso regularization can be achieved only when the mixture parameters η and β are chosen depending on the ratio of the two sparsity numbers. Given that s_{TV} and s_1 are usually unknown, one should again use an adaptation step on the choice of these mixture parameters. Since s_{TV} and s_1 are in the finite set $\{1, \ldots, p\}$, this can be easily done using the Lepski's method. The resulting estimator will have the same properties as $\hat{t}_{\lambda,K}$ but with a data-driven way to chose η and β. We do not perform this extra step here since it is now well understood (see Devroye et al. (2016); Lecué and Lerasle (2019)).

6.7 Robust quantile matrix completion via the minmax MOM quantile trace norm procedure

The aim of this section is to provide an example where the $L_{2+\varepsilon}/L_2$ assumption from point *a)* in Assumptions 10, 11 and 12 is satisfied but with a "constant" C' depending on the dimension of the problem. Our aim is to show that even in such a situation, Theorem 2 can be efficiently applied without any loss in the rate of convergence; the only price we pay is on the number of observations we need to insure such a result.

The matrix completion problem with continuous entries has been tackled using a square loss in Candès and Plan (2010); Koltchinskii et al. (2011); Klopp (2014); Lecué and Mendelson (2018); Mai and Alquier (2015) and robust loss functions in Elsener and van de Geer (2018); Alquier et al. (2017) for bounded or subgaussian classes of functions in the ideal i.i.d. setup. Our aim here is to make no assumption on the class F, that is to take $F = \{\langle \cdot, M\rangle : M \in \mathbb{R}^{m\times T}\}$ the class of all linear functions, and construct robust procedures, that is procedures which can handle any type of outliers in the database.

Let us start with the particular form of design used in matrix completion which is a generalization the uniform distribution over the canonical basis of $\mathbb{R}^{m\times T}$.

Assumption 13 (Matrix completion design) *The variable X takes value in the canonical basis $(E_{1,1}, \cdots, E_{m,T})$ of $\mathbb{R}^{m\times T}$. There are positive constants $\underline{c}, \bar{c}$ such that for any $(p,q) \in \{1,\dots,m\}\times\{1,\dots,T\}$, $\underline{c}/(mT) \le \mathbb{P}(X = E_{p,q}) \le \bar{c}/(mT)$.*

As we will see below, a design distribution such as in Assumption 13 does not satisfy a $L_{2+\varepsilon}/L_2$ norm equivalence assumption over F for an absolute constant C'.

In the ideal noiseless and i.i.d. setup, we observe $Y_i = \langle X_i, M^*\rangle, i = 1,\dots,N$ where $X_1,\dots,X_N$ are distributed like X satisfying Assumption 13. The aim is to reconstruct exactly M^* when M^* is known to have low rank Candès and Tao (2010).

In the noisy setup, quantile matrix completion can be particularly useful for large and small values of τ (for example the 0.05 and 0.95 quantiles). Actually, these allow to build confidence intervals for $Y|X = E_{p,q}$ for all entries (p,q) of the oracle matrix M^* that we want to recover. For this type of problem, global Bernstein conditions have been proved under a boundedness assumption on the class F in van de Geer (2016). In particular, this assumption implies that M^* has entries uniformly bounded by some known absolute constant b. There are situations where this assumption is naturally satisfied, for instance, when M^* is a users/items matrix whose entries are grades such as $\{0,1,2,3,4,5\}$. In other applications, M^* is a large table of similarities measures using an unbounded metric or kernel. In that case, the boundedness assumption is not satisfied and therefore previous results do not apply.

The main difficulty is that the matrix design under Assumption 13 is degenerated in the sense that it is not a typical diagonal random vector (such as a standard Gaussian vector in $\mathbb{R}^{m\times T}$). In this situation, a $L_{2+\varepsilon}/L_2$ norm equivalence assumption holds over the entire space F but with a dimension dependent constant C' of the order of $(mT)^{1/4}$: for all $M \in \mathbb{R}^{m\times T}$, $\|\langle X, M\rangle\|_{L_4} \le (\bar{c}/\underline{c})(mT)^{1/4}\|\langle X, M\rangle\|_{L_2}$. The constant $C' \sim (mT)^{1/4}$ cannot be improved since it is achieved (up to constants) by $M = E_{p,q}$ whatever $(p,q) \in [m]\times[T]$. Nevertheless, we show in the following that Theorem 2 still applies without loss of performance even though C' depends on the dimension. We first recall a result from Alquier et al. (2017) which we will use as a benchmark. It proves estimation bounds and an oracle inequality for a regularized ERM procedure using the S_1-norm as a low-rank inducing norm. We recall that the S_1-norm of $M \in \mathbb{R}^{m\times T}$ is the sum of its singular values.

Theorem 11 (Theorem 11 in Alquier et al. (2017)) *Assume that Assumption 13 holds and that the data are i.i.d.. Let $b > 0$ and assume that $M^* \in bB_\infty$. Assume that for any (p,q), $Y|X = E_{p,q}$ has a density g with respect to the Lebesgue measure such that $g(u) > 1/c_0$ for some constant $c_0 > 0$ for any u such that $|u - M^*_{p,q}| \le 2b$. Let $s \in \{1,\dots,\min(m,T)\}$ and assume that M^* has rank at most s. Then, with probability at least $1 - c(c_0,\bar{c},\underline{c},b)\exp\left(-c(c_0,\bar{c},\underline{c},b)s\max(m,T)\log(m+T)\right)$, the estimator*

$$\widehat{M} \in \underset{M\in bB_\infty}{\operatorname{argmin}}\left(\frac{1}{N}\sum_{i=1}^N \rho_\tau(Y_i - \langle X_i, M\rangle) + \lambda\|M\|_{S_1}\right) \tag{33}$$

with $\lambda = c(c_0, \bar{c}, \underline{c}, b)\sqrt{\log(m+T)/(N\min(m,T))}$ *satisfies*

$$\frac{1}{mT}\left\|\widehat{M} - M^*\right\|_{S_1} \leq c(c_0, \bar{c}, \underline{c}, b) \min\left\{ s\sqrt{\frac{\log(m+T)}{N\min(m,T)}}, \frac{\|M^*\|_{S_1}}{mT}\right\},$$

$$\frac{1}{\sqrt{mT}}\left\|\widehat{M} - M^*\right\|_{S_2} \leq c(c_0, \bar{c}, \underline{c}, b) \min\left\{ \sqrt{\frac{s(m+T)\log(m+T)}{N}}, \|M^*\|_{S_1}^{\frac{1}{2}} \left(\frac{\log(m+T)}{N\min(m,T)}\right)^{\frac{1}{4}}\right\}$$

$$\mathcal{E}_{quantile}(\widehat{M}) \leq c(c_0, \bar{c}, \underline{c}, b) \min\left\{ \frac{s(m+T)\log(m+T)}{N}, \|M^*\|_{S_1}\sqrt{\frac{\log(m+T)}{N\min(m,T)}}\right\}$$

where $\mathcal{E}_{quantile}(\widehat{M})$ *is the excess quantile risk of* $\widehat{M}$.

The S_2 estimation rates obtained in Theorem 11 are the same as the one obtained in Rohde and Tsybakov (2011); Koltchinskii et al. (2011) for the penalized least squares estimator that is $\sqrt{s(m+T)\log(m+T)/N}$.

We now show how the results from Theorem 11 can be extended in several ways using a trace-norm regularized minimax MOM estimator using the quantile loss function: for all $\lambda > 0$

$$\widehat{M}_{\lambda,K} \in \underset{M\in\mathbb{R}^{m\times T}}{\mathrm{argmin}} \sup_{\tilde{M}\in\mathbb{R}^{m\times T}} \left(\mathrm{MOM}_K\left[\ell_M - \ell_{\tilde{M}}\right] + \lambda(\|M\|_{S_1} - \|\tilde{M}\|_{S_1})\right).$$

We show below that $\widehat{M}_{\lambda,K}$ satisfies the same results as in Theorem 11 but under weaker assumptions: we don't need M^* to have bounded entries and we allow the dataset to be corrupted. To prove this result we apply Theorem 2 and so we need to check its assumptions.

We start with the local Bernstein condition and to that end Proposition 3 will be helpful: we only need to check Assumption 12 that we do now. We have already seen that point *a)* of Assumption 12 is satisfied for $C' = (\bar{c}/\underline{c})(mT)^{1/4}$. Given that we make no assumption on the class F of linear functions, we see that f^* is the Bayes rules, that is for every $x \in \mathcal{X} := \{E_{p,q} : (p,q) \in [m]\times[T]\}$, we have $f^*(x) = q_\tau^{Y|X=x} = \langle M^*, x\rangle$ where $M^* = (q_\tau^{Y|X=E_{p,q}})_{(p,q)\in[m]\times[T]}$. Hence, point *b)* of Assumption 12 is satisfied. To check point *c)* we make the same assumption as in Theorem 11: there are constants $C_0 > 0$ and $\alpha > 0$ such that for all $z \in \mathbb{R}$ satisfying $|z - M^*_{p,q}| \leq C_0$, $f_{Y|X=E_{p,q}}(z) \geq \alpha$ (where $f_{Y|X=E_{p,q}}$ is a density function w.r.t. the Lebesgue measure of Y given $X = E_{p,q}$). Under this assumption, point *c)* will be satisfied if $2(C')^2 C_{K,r}(2\rho^*, 4/\alpha) \leq C_0$ where $C_{K,r}$ and ρ^* are two quantities that we are now computing using their definitions from Definition 5.

We first start with the computation of the fixed point $r_2^2(\gamma,\rho)$ for all $\rho > 0$ and some $\gamma > 0$. When the design X is distributed according to the matrix completion design assumption from Assumption 13 it is (up to absolute constants) isotropic: for all $M \in \mathbb{R}^{m\times T}, \underline{c}(mT)^{-1}\|M\|_{S_2}^2 \leq \mathbb{E}\langle X, M\rangle^2 \leq \bar{c}(mT)^{-1}\|M\|_{S_2}^2$. A rB_{L_2} localization is therefore equivalent (up to constants) to a $\sqrt{mT}rB_{S_2}$ localization (where B_{S_2} is the unit ball of the S_2 norm which is the Frobenius norm: the ℓ_2 norm of the spectrum). In our setup, we therefore have for all $J \subset \mathcal{I}$ such that $|J| \geq N/2$

$$\mathbb{E}\left[\sup_{f\in(F-f^*)\cap\rho B\cap rB_{L_2}}\left|\sum_{i\in J}\sigma_i f(X_i)\right|\right] \leq \mathbb{E}\sup_{M\in\rho B_{S_1}\cap(r\sqrt{mT}/\underline{c})B_{S_2}}\left\langle M, \sum_{i\in J}\sigma_i X_i\right\rangle = (\star)$$

where we recall that for all $p \geq 1$, S_p is the Schatten-p norm, i.e. the ℓ_p norm of the spectrum. A sharp estimate of $(\star)$ requires estimates of the Euclidean norm of the first singular values of $\sum_{i \in J} \sigma_i X_i$ when $(X_i)_{i \in J}$ are i.i.d. distributed according to the matrix design assumption. Such a result is not yet available. We will therefore use a loose estimate here which causes extra log factors. The key argument to compute the statistical complexity of the problem is Corollary 9.1 in Tropp (2012) (which is also used in Proposition 1 from Koltchinskii et al. (2011)); it yields

$$(\star) \leq \rho \left\| \sum_{i \in J} \sigma_i f(X_i) \right\|_{S_\infty} \leq c(\bar{c}, \underline{c}) \rho \sqrt{\frac{|J| \log(m+T)}{\min(m,T)}} \tag{34}$$

and therefore, one can take

$$r_2^2(\gamma, \rho) = \frac{c(\bar{c}, \underline{c})\rho}{\gamma} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \quad .$$

We now compute a radius ρ^* solution of the sparsity equation. It follows from Equation (15) in Alquier et al. (2017) that, if M^* has rank s and $K = c(\bar{c}, \underline{c}, \alpha) N r_2^2(\gamma, \rho^*)$, then one can pick

$$\rho^* = c(\bar{c}, \underline{c}, \alpha) \min\left(smT \sqrt{\frac{\log(m+T)}{N \min(m,T)}}, \|M^*\|_{S_1} \right) \quad .$$

We see that the condition $2(C')^2 C_{K,r}(2\rho^*, 4/\alpha) \leq C_0$ is satisfied when $K = c(\bar{c}, \underline{c}, \alpha) N r_2^2(\gamma, \rho^*)$ and $r_2(\gamma, 2\rho^*) \leq c(\bar{c}, \underline{c}, \alpha)(mT)^{-1/2}$, i.e. $N \geq c(\bar{c}, \underline{c}, \alpha, C_0) s (mT)^{3/2} \log(m+T)/\min(m,T)$. We now have all the ingredients to apply Theorem 2 to get the following result.

Theorem 12 *Assume that $(X_i, Y_i)_{i \in \mathcal{I}}$ are i.i.d. such that $Y_i \in L_1$ and the design matrices X_i is distributed according to the matrix completion design from Assumption 13. Assume that there are constants $C_0 > 0$ and $\alpha > 0$ such that for any $(p,q) \in [m] \times [T]$, $Y|X = E_{p,q}$ has a density g with respect to the Lebesgue measure such that $g(u) \geq \alpha$ for any u such that $|u - M^*_{p,q}| \leq C_0$. Let $s \in \{1, \ldots, \min(m,T)\}$ and assume that M^* has rank at most s. Let $K = c(\bar{c}, \underline{c}, \alpha) \min(s(m+T)\log(m+T), \|M^*\|_{S_1} (N \log(m+T)/\min(m,T))^{1/2}$. Assume that $N \geq c(\bar{c}, \underline{c}, \alpha) s (mT)^{3/2} \log(m+T)/\min(m,T)$ and that $K \geq 7|\mathcal{O}|/3$. Then, with probability at least $1 - c(\bar{c}, \underline{c}, \alpha) \exp\left(-c(\bar{c}, \underline{c}, \alpha) s \max(m,T) \log(m+T)\right)$, the trace-norm regularized minmax MOM estimator $\widehat{M}_{\lambda,K}$ with $\lambda = c(\bar{c}, \underline{c}, \alpha)\sqrt{\log(m+T)/(N \min(m,T))}$ satisfies*

$$\frac{1}{mT} \left\| \widehat{M}_{\lambda,K} - M^* \right\|_{S_1} \leq c(\bar{c}, \underline{c}, \alpha) \min\left\{ s\sqrt{\frac{\log(m+T)}{N \min(m,T)}}, \frac{\|M^*\|_{S_1}}{mT} \right\},$$

$$\frac{1}{\sqrt{mT}} \left\| \widehat{M}_{\lambda,K} - M^* \right\|_{S_2} \leq c(\bar{c}, \underline{c}, \alpha) \min\left\{ \sqrt{\frac{s(m+T)\log(m+T)}{N}}, \|M^*\|_{S_1}^{\frac{1}{2}} \left(\frac{\log(m+T)}{N \min(m,T)} \right)^{\frac{1}{4}} \right\}$$

$$\mathcal{E}_{quantile}(\widehat{M}_{\lambda,K}) \leq c(\bar{c}, \underline{c}, \alpha) \min\left\{ \frac{s(m+T)\log(m+T)}{N}, \|M^*\|_{S_1} \sqrt{\frac{\log(m+T)}{N \min(m,T)}} \right\}$$

where $\mathcal{E}_{quantile}(\widehat{M}_{\lambda,K})$ is the excess quantile risk of $\widehat{M}_{\lambda,K}$.

For simplicity, we did not write the result for all K in Theorem 12 like in the other examples. We chose $K = c(\bar{c}, \underline{c}, \alpha)Nr_2^2(\gamma, \rho^*)$ to recover the same probability estimate and same results as in Theorem 11. However, this choice requires to know the rank (or at least an upper bound on the rank) of M^* and $\|M^*\|_{S_1}$ which are usually not available in practice. To solve this issue we can either write the result for all K as in the previous examples and construct $\widehat{M}_{\lambda,K}$ for this choice of K or use a Lepskii's adaptation method as in Lecué and Lerasle (2019).

Theorem 12 extends Theorem 11 in two ways: we drop the boundedness assumption onto the entries of M^* and we allow for corrupted databases (in both inputs and outputs). In particular, the rates of convergence and residual terms in the estimation bounds and the oracle inequality are not downgraded even though the L_4/L_2 assumption holds with a dimension dependent constant C' of the order of $(mT)^{1/4}$. However, the price we pay for this loose L_4/L_2 norm equivalence is the extra assumption on the number of data $N \geq c(\bar{c}, \underline{c}, \alpha)s(mT)^{3/2}\log(m+T)/\min(m,T)$. This extra assumption is due to the analysis of the local Bernstein condition via Proposition 12 which uses the L_4/L_2 assumption. We can relax this assumption into a $L_{2+\epsilon}/L_2$ assumption (for any $\epsilon > 0$), if the sample size is large enough. One could also try to prove directly the local Bernstein condition as it has been done under the boundedness assumption (see Elsener and van de Geer (2018)).

A final word about the proof technique: Theorem 11 follows from Talagrand's concentration inequality which applies only for bounded class of functions. Beyond the boundedness assumption, we need to use other type of arguments (see, for instance, the difference of analysis between the subgaussian and the bounded cases in Alquier et al. (2017)). Here the main result Theorem 2 covers almost any situation with a single analysis and, in particular, the L_∞-bounded case. Indeed, in that case the local Bernstein is satisfied (see Elsener and van de Geer (2018)) and therefore Theorem 2 applies without any restriction on the number of data.

7. Simulations

This section provides a simulation study to illustrate our theoretical findings. Minmax MOM estimators are approximated using an alternating proximal block gradient descent/ascent with a wisely chosen block of data as in Lecué and Lerasle (2019). At each iteration, the block on which the descent/ascent is performed is chosen according to its "centrality" (see algorithm 1 below). Two examples from high-dimensional statistics are considered 1) Logistic classification with a ℓ_1 penalization and 2) Huber regression with a Group-Lasso penalization.

7.1 Presentation of the algorithm

Let $\mathcal{X} = \mathbb{R}^p$ and let $F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$. The oracle $f^* = \mathrm{argmin}_{f \in F} P\ell_f(X,Y)$ is such that $f^*(\cdot) = \langle t^*, \cdot \rangle$ for some $t^* \in \mathbb{R}^p$. The minmax MOM estimator is defined as

$$\hat{t}_{\lambda,K} \in \underset{t \in \mathbb{R}^p}{\mathrm{argmin}} \sup_{\tilde{t} \in \mathbb{R}^p} MOM_K(\ell_t - \ell_{\tilde{t}}) + \lambda(\|t\| - \|\tilde{t}\|) \tag{35}$$

where ℓ is a convex and Lipschitz loss function and $\| \cdot \|$ is a norm in $\mathbb{R}^p$.

Following the idea of Lecué and Lerasle (2019), the minmax problem (35) is approximated by a proximal block gradient ascent-descent algorithm, see Algorithm 1. At each step, one considers the block of data realizing the median and perform an ascent/descent step onto this block. The regularization step is obtained via the proximal operator

$$\text{prox}_{\lambda\|\cdot\|} : x \in \mathbb{R}^p \to \underset{y\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2}\|x-y\|_2^2 + \lambda\|y\| \right\} .$$

Algorithm 1: Proximal Descent-Ascent gradient method with median blocks

Input: A number of blocks K, initial points t_0 and $\tilde{t}_0$ in $\mathbb{R}^p$, two sequences of step sizes $(\eta_t)_t$ and $(\tilde{\eta}_t)_t$ and T a number of epochs

Output: An approximating solution of the minimax problem (35)

for $i = 1, \cdots, T$ **do**

Construct a random equipartition $B_1 \sqcup \cdots \sqcup B_K$ of $\{1, \cdots, N\}$

Find $k \in [K]$ such that $\text{MOM}_K(\ell_{t_i} - \ell_{\tilde{t}_i}) = P_{B_k}(\ell_{t_i} - \ell_{\tilde{t}_i})$

Update:

$$t_{i+1} = \text{prox}_{\lambda\|\cdot\|}\big(t_i - \eta_i \nabla_t (t \to P_{B_k}\ell_t)_{|t=t_i}\big)$$

$$\tilde{t}_{i+1} = \text{prox}_{\lambda\|\cdot\|}\big(\tilde{t}_i - \tilde{\eta}_i \nabla_{\tilde{t}} (\tilde{t} \to P_{B_k}\ell_{\tilde{t}})_{|\tilde{t}=\tilde{t}_i}\big)$$

end

To make the presentation simple in Algorithm 1, we have not perform any line search or any sophisticated stopping rule (see, Lecué and Lerasle (2019) for more involved line search and stopping rules). To compare the statistical and robustness performances of the minmax MOM and RERM, we perform a proximal gradient descent to approximate the RERM, see Algorithm 2 below.

Algorithm 2: Proximal gradient descent algorithm

Input: Initial points t_0 in $\mathbb{R}^p$ and a sequence of stepsizes $(\eta_t)_t$

Output: RERM estimator.

for $i = 1, \cdots, T$ **do**

$$t_{i+1} = \text{prox}_{\lambda\|\cdot\|}\big(t_i - \eta_i \nabla_t (t \to P_N \ell_t)_{|t=t_i}\big)$$

end

The number of blocks K is chosen by MOM cross-validation (see Lecué and Lerasle (2019) for more precision on that procedure). The sequences of stepsizes are constant along the algorithm $(\eta_t)_t := \eta$ and $(\tilde{\eta}_t)_t = \tilde{\eta}$ and are also chosen by MOM cross-validation.

7.2 Organisation of the results

In all simulations, the links between inputs and outputs are given in the regression and classification problems in $\mathbb{R}^p$ respectively by the following model:

$$\text{in regression: } Y = \big\langle X, t^* \big\rangle + \zeta; \qquad \text{in classification: } Y = \text{sign}\big(\big\langle X, t^* \big\rangle + \zeta\big) \tag{36}$$

where the distribution of X and ζ depend on the considered framework:

- **First framework:** X is a standard Gaussian random vector in $\mathbb{R}^p$ and ζ is a real-valued standard Gaussian variable independent of X with variance σ^2.

- **Second framework:** X is a standard Gaussian random vector in $\mathbb{R}^p$ and $\zeta \sim \mathcal{T}(2)$ (student distribution with 2 degrees of freedom). This framework is used to verify the robustness w.r.t the noise.

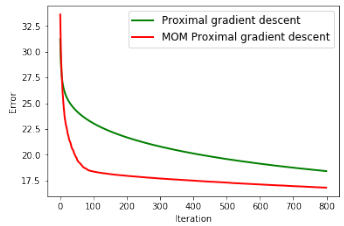
- **Third framework:** $X = (x_1, \cdots, x_p)$ with $x_1, \ldots, x_p \overset{i.i.d.}{\sim} \mathcal{T}(2)$ and ζ is a real-valued standard Gaussian variable independent of X with variance σ^2. Here we want to test the robustness w.r.t heavy-tailed design $(X_i)_i$.

- **Fourth framework:** $X = (x_1, \cdots, x_p)$ with $x_1, \ldots, x_p \overset{i.i.d.}{\sim} \mathcal{T}(2)$ and $\zeta \sim \mathcal{T}(2)$. We also corrupt the database with $|\mathcal{O}|$ outliers which are such that for all $i \in \mathcal{O}$, $X_i = (10^5)_{i=1}^p$ and $Y = 1$. Here we verify the robustness w.r.t possible outliers in the dataset.

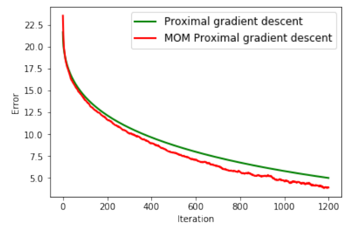
In a both first and second frameworks, the RERM and minmax MOM estimators are expected to perform well according to Theorem 1 and Theorem 2 even though the noise ζ can be heavy-tailed. In the third framework, the design vector X is no longer subgaussian, as a consequence Theorem 1 does not apply and we have no guarantee for the RERM. On the contrary, Theorem 2 provides statistical guarantees for the minmax MOM estimators. Nevertheless, it should also be noticed that the study of RERM under moment assumptions on the design can also be performed, see for instance Lecué and Mendelson (2017). In that case, the rates of convergence are still the same but the deviation is only polynomial whereas it is exponential for the minmax MOM estimators. Therefore, in the third example, we may expect similar performance for both estimators but with a larger variance in the results for the RERM. In the fourth framework, the database has been corrupted by outliers (in both outputs Y_i and inputs X_i); in that case, only minmax MOM estimators are expected to perform well as long as $|\mathcal{O}|$ is not too large compare with K, the number of blocks.

7.3 Sparse Logistic regression

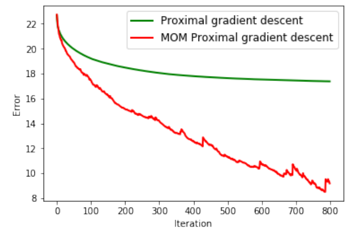
Let ℓ denote the Logistic loss (i.e. $t \in \mathbb{R}^p \to \ell_t(x, y) = \log(1 + \exp(-y\langle x, t\rangle)), \forall x \in \mathbb{R}^p, y \in \mathcal{Y} = \{\pm 1\}$), and let the ℓ_1 norm in $\mathbb{R}^p$ be the regularization norm. Figure 1 presents the results of our simulations for $N = 1000$, $p = 400$ and $s = 30$. In subfigures (a), (b) and (c) the error is the L_2 error, which is here $\left\|\hat{t}_{K,\lambda}^T - t^*\right\|_2$, between the output $\hat{t}_{K,\lambda}^T$ of the algorithm and the true $t^* \in \mathbb{R}^p$. In subfigure (d), an increasing number of outliers is added. The error rate is the proportion of misclassification on a test dataset. The stepsizes, the number of block and the parameteter of regularization are all chosen by MOM cross-validation (see Lecué and Lerasle (2019) for more details on the MOM cross-validation procedure) Subfigure (a) shows convergence of the error for both algorithms in the first framework. Similar performances are observed for both algorithms but Algorithm 1 converges faster than Algorithm 2. It may be because the computation of the gradient on a smaller batch of data in step **5** and **6** of Algorithm 1 is faster than the one on the entire database in step **2** of Algorithm 2 and that the choice of the median blocks at each descent/ascent step is particularly good in Algorithm 1. Subfigure (b) shows the results in the second framework. The convergence for the alternating gradient-ascent descent algorithm is a bit faster as the one from Algorithm 2, but the performances are the same. Subfigure (c) shows results in the third setup where ζ is Gaussian and the feature vector $X = (x_1, \cdots, x_p)$ is heavy-tailed, i.e. $x_1, \ldots, x_p$ are i.i.d. with $x_1 \sim \mathcal{T}(2)$ – a Student with degree 2. Minmax MOM estimators perform better than RERM. It highlights the fact that minmax MOM estimators have optimal subgaussian performance even without the sub-gaussian assumption on the design while RERM are expected to have downgraded statistical properties in heavy-tailed scenariis. Subfigure (d) shows result in the fourth setup where an increasing number of outliers is added in the dataset. Outliers are $X = (10^5)_1^p$ and $Y_i = 1$ a.s.. While RERM has deteriorated performance just after one outliers was added to the dataset, minmax MOM estimators maintains good performances up to 10% of outliers.



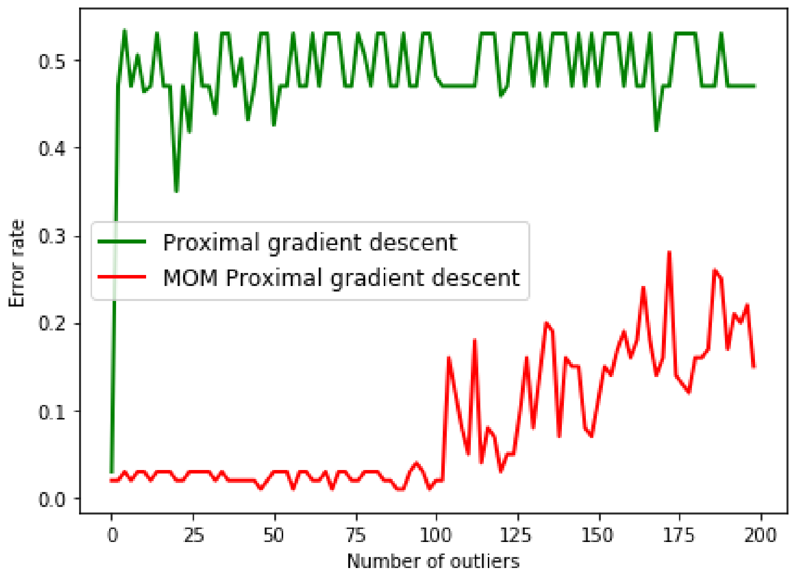
(a) Gaussian design and Gaussian noise.



(b) Heavy-tailed noise ζ (Student distribution of order 2) and standard Gaussian design.



(c) Gaussian noise and heavy-tailed design (Student distribution of order 2).



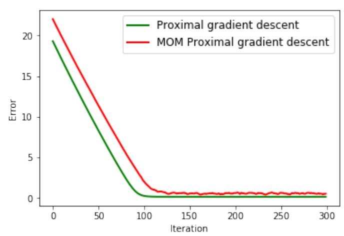
(d) Student of order 2 design and noise corrupted by outliers.

Figure 1: ℓ_2 estimation error rates of RERM and minmax MOM proximal descent algorithms (for the logistic loss and the ℓ_1 regularization norm) versus time in (a), (b) and (c) and versus number of outliers in (d) in the classification model (36) for $N = 1000$, $p = 400$ and $s = 30$.

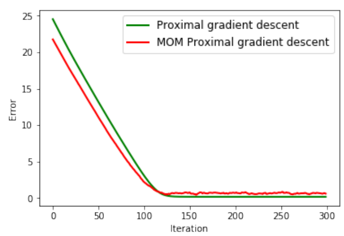
7.4 Huber regression with a Group Lasso penalty

Let ℓ denote the Huber loss function $t \in \mathbb{R}^d \to \ell_t(x, y) = (y - \langle x, t \rangle)^2/2$ if $|y - \langle x, t \rangle| \leq \delta$ and $\ell_t(x, y) = \delta|y - \langle x, t \rangle| - \delta^2/2$ other wise for all $x \in \mathbb{R}^p$ and $y \in \mathcal{Y} = \mathbb{R}$. Let $G_1, \cdots, G_M$ be a partition of $\{1, \cdots, p\}$, $\|t\| = \|t\|_{\text{GL}} = \sum_{k=1}^M \|t_{G_k}\|_2$. Figure 1 presents the results of our simulation for $N = 1000$, $p = 400$ for 30 blocks with a block-sparsity parameter $s = 5$. In subfigures (a), (b) and (c), the error is the L_2-error between the output of the algorithm and the oracle t^* – which corresponds here to a ℓ_2^p estimation error, given that the design in all cases is isotropic. In subfigure (d) the prediction error on a (non-corrupted) test set of both the RERM and the minmax MOM estimators are depicted.

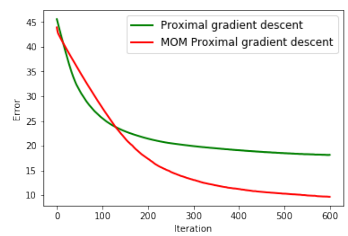
The conclusion are the same as for the Lasso Logistic regression: Algorithm 1 (regularized minmax MOM) has better performances than algorithm 2 (RERM) in case of heavy-tailed inliers and when outliers pollute the dataset while both are robust w.r.t heavy-tailed noise.



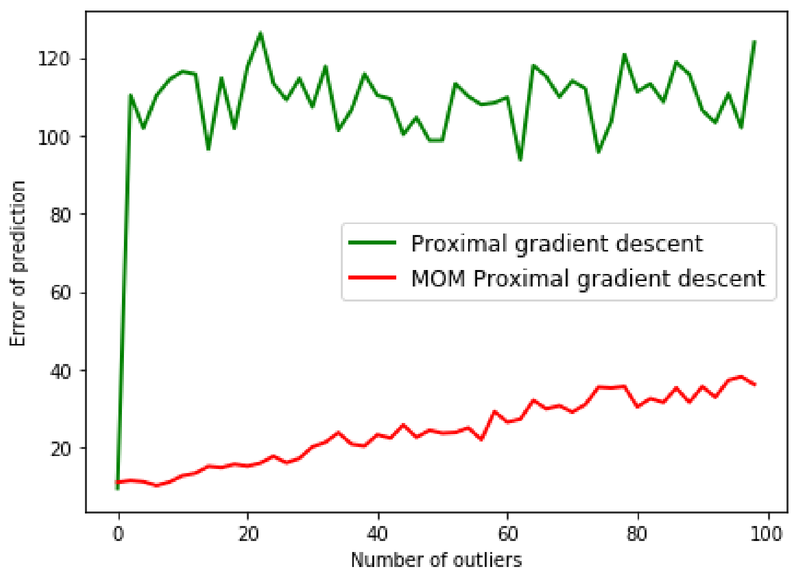
(a) Simulations from model (36) with standard Gaussian design and Gaussian noise



(b) Simulation with heavy-tailed noise ζ and standard Gaussian design



(c) Simulations with Gaussian noise heavy tailed design (Student distribution)



(d) Error of prediction in function of the number of outliers in the dataset

Figure 2: Results for the Huber regression with Group-Lasso penalization

8. Conclusion

We obtain estimation and prediction results for RERM and regularized minmax MOM estimators for any Lipschitz and convex loss functions and for any regularization norm. When the norm has some sparsity inducing properties the statistical bounds depend on the dimension of the low-dimensional structure where the oracle belongs. We develop a systematic way to analyze both estimators by identifying three key idea 1) the local complexity function r_2 2) the sparsity equation 3) the local Bernstein condition. All these quantities and condition depend only of the structure and complexity of a local set around the oracle. This local set is ultimately proved to be the smallest set containing our estimators. We show the versatility of our main meta-theorems in an extensive applications section covering three different loss functions and five sparsity inducing regularization norm. Some of them inducing highly structured sparsity concept such as the Fused Lasso norm.

On top of these results, we show that the minmax MOM approach is robust to outliers and to heavy-tailed data and that the computation of the key objects such as the complexity functions r_2 and a radius ρ^* satisfying the sparsity equation can be done in this corrupted heavy-tailed scenario. Moreover, we show in a simulation section that they can be computed by a simple modification of existing proximal gradient descent algorithms by simply adding a selection step of the central block of data in these algorithms. The resulting algorithms is indeed robust to heavy-tailed data and to few outliers (in both input and output variables).

Acknowledgements

Guillaume Lecué is supported by a grant of the French National Research Agency (ANR), Investissements d'Avenir (LabEx Ecodec/ANR-11-LABX-0047).

9. Proof Theorem 1

All along this section we will write $r(\rho)$ for $r(A,\rho)$. Let $\theta = 1/(3A)$. The proof is divided into two parts. First, we identify an event where the RERM $\hat{f} := \hat{f}_\lambda^{RERM}$ is controlled. Then, we prove that this event holds with large probability. Let ρ^* satisfying the A-sparsity Equation from Definition 4 and let $\mathcal{B} = \rho^* B \cap r(\rho^*) B_{L_2}$ and consider

$$\Omega := \left\{ \forall f \in F \cap (f^* + \mathcal{B}), \quad \left| (P - P_N)\mathcal{L}_f \right| \le \theta r^2(\rho^*) \right\} \ .$$

Proposition 4 *Let λ be as in (6) and let ρ^* satisfy the A- sparsity from Definition 4. On Ω, one has*

$$\|\hat{f} - f^*\| \le \rho^*, \quad \|\hat{f} - f^*\|_{L_2} \le r(\rho^*) \text{ and } P\mathcal{L}_{\hat{f}} \le A^{-1} r^2(\rho^*) \ .$$

Proof Prove first that $\hat{f} \in f^* + \mathcal{B}$. Recall that

$$\forall f \in F, \qquad \mathcal{L}_f^\lambda = \mathcal{L}_f + \lambda(\|f\| - \|f^*\|) \ .$$

Since $\hat{f}$ satisfies $P_N \mathcal{L}_{\hat{f}}^\lambda \leqslant 0$, it is sufficient to prove that $P_N \mathcal{L}_f^\lambda > 0$ for all $f \in F \backslash (f^* + \mathcal{B})$ to get the result. The proof relies on the following homogeneity argument. If $P_N \mathcal{L}_{f_0} > 0$ on the border of $f^* + \mathcal{B}$, then $P_N \mathcal{L}_f > 0$ for all $f \in F \setminus \{f^* + \mathcal{B}\}$.

Let $f \in F \setminus \{f^* + \mathcal{B}\}$. By convexity of $\{f^* + \mathcal{B}\} \cap F$, there exists $f_0 \in F$ and $\alpha > 1$ such that $f - f^* = \alpha(f_0 - f^*)$ and $f_0 \in \partial(f^* + \mathcal{B})$ where $\partial(f^* + \mathcal{B})$ denotes the border of $f^* + \mathcal{B}$ (see, Figure 3).

For all $i \in \{1, \cdots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be the random function defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \ell(u + f^*(X_i), Y_i) - \ell(f^*(X_i), Y_i) \ . \tag{37}$$

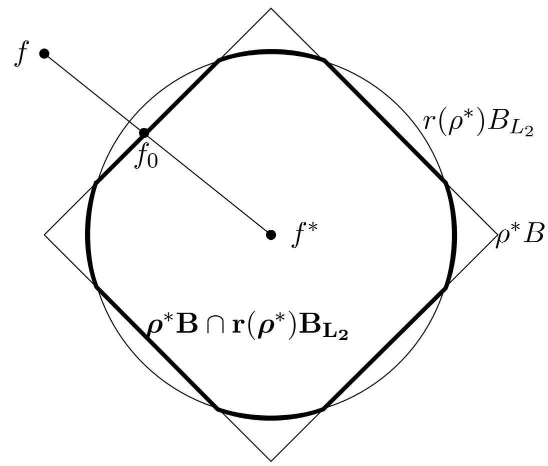


Figure 3: Construction of f_0.

By construction, for any i, $\psi_i(0) = 0$ and ψ_i is convex because ℓ is. Hence, $\alpha\psi_i(u) \leq \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha \geq 1$. In addition, $\psi_i(f(X_i) - f^*(X_i)) = \ell(f(X_i), Y_i) - \ell(f^*(X_i), Y_i)$. Therefore,

$$P_N \mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N} \psi_i\big(f(X_i) - f^*(X_i)\big) = \frac{1}{N}\sum_{i=1}^{N} \psi_i\big(\alpha(f_0(X_i) - f^*(X_i))\big)$$
$$\geq \frac{\alpha}{N}\sum_{i=1}^{N} \psi_i(f_0(X_i) - f^*(X_i)) = \alpha P_N \mathcal{L}_{f_0} \ . \tag{38}$$

For the regularization term, by the triangular inequality,

$$\|f\| - \|f^*\| = \|f^* + \alpha(f_0 - f^*)\| - \|f^*\| \geq \alpha(\|f_0\| - \|f^*\|) \ .$$

From the latter inequality, together with (38), it follows that

$$P_N \mathcal{L}_f^\lambda \geq \alpha P_N \mathcal{L}_{f_0}^\lambda \ . \tag{39}$$

As a consequence, if $P_N \mathcal{L}_{f_0}^\lambda > 0$ for all $f_0 \in F \cap \partial(f^* + \mathcal{B})$ then $P_N \mathcal{L}_f^\lambda > 0$ for all $f \in F \backslash (f^* + \mathcal{B})$.

In the remaining of the proof, assume that Ω holds and let $f_0 \in F \cap \partial(f^* + \mathcal{B})$. As $f_0 \in F \cap (f^* + \mathcal{B})$, on Ω,

$$|(P - P_N)\mathcal{L}_{f_0}| \leq \theta r^2(\rho^*) \ . \tag{40}$$

By definition of $\mathcal{B}$, as $f_0 \in \partial(f^* + \mathcal{B})$, either: 1) $\|f_0 - f^*\| = \rho^*$ and $\|f_0 - f^*\|_{L_2} \leq r(\rho^*)$ so $\alpha = \|f - f^*\| / \rho^*$ or 2) $\|f_0 - f^*\|_{L_2} = r(\rho^*)$ and $\|f_0 - f^*\| \leq \rho^*$ so $\alpha = \|f - f^*\|_{L_2} / r(\rho^*)$. We treat these cases independently.

Assume first that $\|f_0 - f^*\| = \rho^*$ and $\|f_0 - f^*\|_{L_2} \leq r(\rho^*)$. Let $v \in E$ be such that $\|f^* - v\| \leq \rho^*/20$ and $g \in \partial \|\cdot\| (v)$. We have

$$\|f_0\| - \|f^*\| \geq \|f_0\| - \|v\| - \|f^* - v\| \geq \big\langle g, f_0 - v \big\rangle - \|f^* - v\|$$
$$\geqslant \big\langle g, f_0 - f^* \big\rangle - 2\|f^* - v\| \geqslant \big\langle g, f_0 - f^* \big\rangle - \rho^*/10 \ .$$

As the latter result holds for all $v \in f^* + (\rho^*/20)B$ and $g \in \partial \|\cdot\| (v)$, since $f_0 - f^* \in \rho^* S \cap r(\rho^*)B_{L_2}$, it yields

$$\|f_0\| - \|f^*\| \geq \Delta(\rho^*) - \rho^*/10 \geq 7\rho^*/10 \ . \tag{41}$$

Here, the last inequality holds because ρ^* satisfies the sparsity equation. Hence,

$$P_N \mathcal{L}_f^\lambda = P_N \mathcal{L}_f + \lambda \left(\|f\| - \|f^*\| \right) \geq \alpha (P_N \mathcal{L}_{f_0} + 7\lambda\rho^*/10) \ . \tag{42}$$

Thus, on Ω, since $\lambda > 10\theta r^2(\rho^*)^2/(7\rho^*)$,

$$P_N \mathcal{L}_{f_0} + 7\lambda\rho^*/10 = P\mathcal{L}_{f_0} + (P_N - P)\mathcal{L}_{f_0} + 7\lambda\rho^*/10 \geq -\theta r^2(\rho^*) + 7\lambda\rho^*/10 > 0 \ .$$

Assume now that $\|f_0 - f^*\|_{L_2} = r(\rho^*)$ and $\|f_0 - f^*\| \leq \rho^*$. By Assumption 5, on Ω,

$$\begin{aligned} P_N \mathcal{L}_f^\lambda &\geqslant P_N \mathcal{L}_{f_0} - \lambda \|f_0 - f^*\| \geqslant P\mathcal{L}_{f_0} + (P_N - P)\mathcal{L}_{f_0} - \lambda\rho^* \\ &\geqslant A^{-1} \|f_0 - f^*\|_{L_2}^2 - \theta r^2(\rho^*) - \lambda\rho^* \geqslant (A^{-1} - \theta) r^2(\rho^*) - \lambda\rho^* \ . \end{aligned}$$

From (6), $\lambda < (A^{-1} - \theta) r^2(\rho^*)^2/\rho^*$, thus $P_N \mathcal{L}_f^\lambda > 0$. Together with (42), this proves that $\hat{f} \in f^* + \mathcal{B}$. Now, on Ω, this implies that $|(P - P_N)\mathcal{L}_{\hat{f}}| \leq \theta r^2(\rho^*)$, so by definition of $\hat{f}$,

$$P\mathcal{L}_{\hat{f}} = P_N \mathcal{L}_{\hat{f}}^\lambda + (P - P_N)\mathcal{L}_{\hat{f}} + \lambda(\|f^*\| - \|\hat{f}\|) \leqslant \theta r^2(\rho^*) + \lambda\rho^* \leqslant A^{-1} r^2(\rho^*) \ .$$

∎

To prove that Ω holds with large probability, the following result from Alquier et al. (2017) is useful.

Lemma 2 *(Alquier et al., 2017, Lemma 9.1) Grant Assumptions 2 and 4. Let $F' \subset F$ denote a subset with finite L_2-diameter $d_{L_2}(F')$. For every $u > 0$, with probability at least $1 - 2\exp(-u^2)$*

$$\sup_{f,g \in F'} |(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)| \leq \frac{16 L L_0}{\sqrt{N}} \left(w(F') + u d_{L_2}(F') \right) \ .$$

It follows from Lemma 2 that for any $u > 0$, with probability larger that $1 - 2\exp(-u^2)$,

$$\begin{aligned} \sup_{f \in F \cap (f^* + \mathcal{B})} \left| (P - P_N)\mathcal{L}_f \right| &\leqslant \sup_{f,g \in F \cap (f^* + \mathcal{B})} \left| (P - P_N)(\mathcal{L}_f - \mathcal{L}_g) \right| \\ &\leqslant \frac{16 L L_0}{\sqrt{N}} \left(w(F \cap (f^* + \mathcal{B})) + u d_{L_2}(F \cap (f^* + \mathcal{B})) \right) \ . \end{aligned}$$

It is clear that $d_{L_2}(F \cap (f^* + \mathcal{B})) \leqslant r(\rho^*)$. By definition of the complexity function (3), for $u = \theta\sqrt{N} r(\rho^*)/(32 L L_0)$, we have with probability at least $1 - 2\exp\left(-\theta^2 N r^2(\rho^*)/(32 L L_0)^2 \right)$,

$$\forall f \in F \cap (f^* + \mathcal{B}), \qquad \left| (P - P_N)\mathcal{L}_f \right| \leq \theta r^2(\rho^*) \ .$$

10. Proof Theorem 2

All along the proof, the following notations will be used repeatedly.

$$\theta = \frac{1}{34A}, \quad \gamma = \theta/(192L) \quad \hat{f} = \hat{f}_{K,\lambda} \ .$$

The proof is divided into two parts. First, we identify an event where the minmax MOM estimator $\hat{f}$ is controlled. Then, we prove that this event holds with large probability. Let $K \geqslant 7|\mathcal{O}|/3$, and $\kappa \in \{1,2\}$ let

$$C_{K,r,\kappa} = \max\left(\frac{96L^2K}{\theta^2 N}, r_2^2(\gamma, \kappa\rho^*)\right) \quad \text{and} \quad \lambda = 10\theta\frac{C_{K,r,2}}{\rho^*} \ .$$

Let $\mathcal{B}_\kappa = \sqrt{C_{K,r,\kappa}} B_{L_2} \cap \kappa\rho^* B$. Consider the following event

$$\Omega_K = \left\{\forall \kappa \in \{1,2\},\ \forall f \in F \cap f^* + \mathcal{B}_\kappa,\ \sum_{k=1}^K I\left(\left|(P_{B_k} - P)(\ell_f - \ell_{f^*})\right| \leq \theta C_{K,r,\kappa}\right) \geqslant \frac{K}{2}\right\} \quad (43)$$

10.1 Deterministic argument

Lemma 3 *$\hat{f} - f^* \in \mathcal{B}_\kappa$ if there exists $\eta > 0$ such that*

$$\sup_{f \in f^* + F \backslash \mathcal{B}_\kappa} MOM_K\left[\ell_{f^*} - \ell_f\right] + \lambda\big(\|f^*\| - \|f\|\big) < -\eta \ , \quad (44)$$

$$\sup_{f \in F} MOM_K\left[\ell_{f^*} - \ell_f\right] + \lambda\big(\|f^*\| - \|f\|\big) \leq \eta \ . \quad (45)$$

Proof For any $f \in F$, denote by $S(f) = \sup_{g \in F} \text{MOM}_K[\ell_f - \ell_g] + \lambda\big(\|f\| - \|g\|\big)$. If (44) holds, by homogeneity of MOM_K, any $f \in f^* + F \backslash \mathcal{B}_\kappa$ satisfies

$$S(f) \geqslant \inf_{f \in f^* + F \backslash \mathcal{B}_\kappa} \text{MOM}_K[\ell_f - \ell_{f^*}] + \lambda\big(\|f\| - \|f^*\|\big) > \eta \ . \quad (46)$$

On the other hand, if (45) holds,

$$S(f^*) = \sup_{f \in F} \text{MOM}_K[\ell_{f^*} - \ell_f] + \lambda\big(\|f^*\| - \|f\|\big) \leqslant \eta \ .$$

Thus, by definition of $\hat{f}$ and (45),

$$S(\hat{f}) \leqslant S(f^*) \leqslant \eta \ .$$

Therefore, if (44) and (45) hold, $\hat{f} \in f^* + \mathcal{B}_\kappa$. ■

It remains to show that, on Ω_K, Equations (44) and (45) hold for $\kappa = 2$.

Let $\kappa \in \{1,2\}$ and $f \in F \cap \mathcal{B}_\kappa$. On Ω_K, there exist more than $K/2$ blocks B_k such that

$$\left|(P_{B_k} - P)(\ell_f - \ell_{f^*})\right| \leq \theta C_{K,r,\kappa} \ . \quad (47)$$

It follows that

$$\sup_{f\in f^*+F\cap\mathcal{B}_\kappa} \text{MOM}_K\left[\ell_{f^*}-\ell_f\right]\le\theta C_{K,r,\kappa}$$

In addition, $\|f\|-\|f^*\|\le\kappa\rho^*$. Therefore, from the choice of λ, on Ω_K, one has

$$\sup_{f\in f^*+F\cap\mathcal{B}_\kappa} \text{MOM}_K\left[\ell_{f^*}-\ell_f\right]+\lambda\big(\|f^*\|-\|f\|\big)\le(1+10\kappa)\theta C_{K,r,\kappa}\ . \tag{48}$$

Assume that f belongs to $F\backslash\mathcal{B}_\kappa$. By convexity of F, there exists $f_0\in f^*+F\cap\mathcal{B}_\kappa$ and $\alpha>1$ such that

$$f=f^*+\alpha(f_0-f^*)\ . \tag{49}$$

For all $i\in\{1,\cdots,N\}$, let $\psi_i:\mathbb{R}\to\mathbb{R}$ be the random function defined for all $u\in\mathbb{R}$ by

$$\psi_i(u)=\ell(u+f^*(X_i),Y_i)-\ell(f^*(X_i),Y_i)\ . \tag{50}$$

The functions ψ_i are convex and satisfy $\psi_i(0)=0$. Thus $\alpha\psi_i(u)\le\psi_i(\alpha u)$ for all $u\in\mathbb{R}$ and $\alpha>1$ and $\psi_i(f(X_i)-f^*(X_i))=\ell(f(X_i),Y_i)-\ell(f^*(X_i),Y_i)$. Hence, for any block B_k,

$$\begin{aligned}P_{B_k}\mathcal{L}_f&=\frac{1}{|B_k|}\sum_{i\in B_k}\psi_i\big(f(X_i)-f^*(X_i)\big)=\frac{1}{|B_k|}\sum_{i\in B_k}\psi_i\big(\alpha(f_0(X_i)-f^*(X_i))\big)\\&\ge\frac{\alpha}{|B_k|}\sum_{i\in B_k}\psi_i\big(f_0(X_i)-f^*(X_i)\big)=\alpha P_{B_k}\mathcal{L}_{f_0}\ .\end{aligned} \tag{51}$$

By the triangular inequality,

$$\|f\|-\|f^*\|=\|f^*+\alpha(f_0-f^*)\|-\|f^*\|\ge\alpha(\|f_0\|-\|f^*\|).$$

Together with (51), this yields, for all block B_k

$$P_{B_k}\mathcal{L}_f^\lambda\ge\alpha P_{B_k}\mathcal{L}_{f_0}^\lambda\ . \tag{52}$$

As $f_0\in F\cap\mathcal{B}_\kappa$, on Ω_K,

$$|(P-P_{B_k})\mathcal{L}_{f_0}|\le\theta C_{K,r,\kappa}. \tag{53}$$

As f_0 can be chosen in $\partial(f^*+\mathcal{B}_\kappa)$, either: 1) $\|f_0-f^*\|=\kappa\rho^*$ and $\|f_0-f^*\|_{L_2}\le\sqrt{C_{K,r,\kappa}}$ or 2) $\|f_0-f^*\|_{L_2}=\sqrt{C_{K,r,\kappa}}$ and $\|f_0-f^*\|\le\kappa\rho^*$.

Assume first that $\|f_0-f^*\|=\kappa\rho^*$ and $\|f_0-f^*\|_{L_2}\le\sqrt{C_{K,r,\kappa}}$. Since the sparsity equation is satisfied for $\rho=\rho^*$, it is also satisfied for $\kappa\rho^*$. By (41),

$$\lambda\big(\,\|f_0\|-\|f^*\|\,\big)\ge7\lambda\kappa\rho^*/10=7\kappa C_{K,r,2}\ . \tag{54}$$

Therefore, on Ω_K, there are more than $K/2$ blocks B_k where

$$P_{B_k}\mathcal{L}_f^\lambda\ge\alpha P_{B_k}\mathcal{L}_{f_0}^\lambda\ge\alpha\bigg(-\theta C_{K,r,\kappa}+\frac{7\kappa\lambda\rho^*}{10}\bigg)\geqslant\alpha(7\kappa-1)\theta C_{K,r,2}\ . \tag{55}$$

It follows that

$$\text{MOM}_K\left[\ell_f-\ell_{f^*}\right]+\lambda\big(\|f\|-\|f^*\|\big)\geqslant\alpha\theta\big(7\kappa C_{K,r,2}-C_{K,r,\kappa}\big)C_{K,r,2}\ . \tag{56}$$

Assume that $\|f_0 - f^*\|_{L_2} = \sqrt{C_{K,r,\kappa}}$ and $\|f_0 - f^*\| \le \kappa\rho^*$. By Assumption 7, on Ω_K, there exist more than $K/2$ blocks B_k where

$$P_{B_k}\mathcal{L}_f^\lambda \geqslant P_{B_k}\mathcal{L}_{f_0} - \lambda\|f_0 - f^*\| \ge P\mathcal{L}_{f_0} + (P_{B_k} - P)\mathcal{L}_{f_0} - \lambda\kappa\rho^*$$
$$\ge A^{-1}\|f_0 - f^*\|_{L_2}^2 - \theta C_{K,r,\kappa} - \kappa\lambda\rho^* = \theta(33C_{K,r,\kappa} - 10\kappa C_{K,r,2}) \ .$$

It follows that

$$\text{MOM}_K\left[\ell_f - \ell_{f^*}\right] + \lambda\big(\|f\| - \|f^*\|\big) \geqslant \alpha\theta(33C_{K,r,\kappa} - 10\kappa C_{K,r,2}) \ . \tag{57}$$

From Equations (48), (56) and (57) with $\kappa = 1$, it follows that

$$\sup_{f\in F}\text{MOM}_K\left[\ell_{f^*} - \ell_f\right] + \lambda\big(\|f^*\| - \|f\|\big) \le 11\theta C_{K,r,2} \ . \tag{58}$$

Therefore, (45) holds with $\eta = 11\theta C_{K,r,2}$. Now, Equations (56) and (57) with $\kappa = 2$ yield

$$\sup_{f\in f^*+F\setminus\mathcal{B}_2}\text{MOM}_K\left[\ell_{f^*} - \ell_f\right] + \lambda\big(\|f^*\| - \|f\|\big) \leqslant -13\alpha\theta C_{K,r,2} < -11\theta C_{K,r,2} \ .$$

Therefore, Equation (44) holds with $\eta = 11\theta C_{K,r,2}$. Overall, Lemma 3 shows that $\hat{f} \in \mathcal{B}_2$. On Ω_K, this implies that there exist more than $K/2$ blocks B_k where $P\mathcal{L}_{\hat{f}} \le P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r,2}$. In addition, by definition of $\hat{f}$ and (58),

$$\text{MOM}_K\left[\ell_{\hat{f}} - \ell_{f^*}\right] + \lambda(\|\hat{f}\| - \|f^*\|) \leqslant \sup_{f\in F}\text{MOM}_K\left[\ell_{f^*} - \ell_f\right] + \lambda(\|f^*\| - \|f\|) \le 11\theta C_{K,r,2} \ .$$

This means that there exist at least $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_{\hat{f}} + \lambda(\|\hat{f}\| - \|f^*\|) \leqslant 11\theta C_{K,r,2}$. As $\|\hat{f}\| - \|f^*\| \geqslant -\|\hat{f} - f^*\| \geqslant -2\rho^*$, on these blocks, $P_{B_k}\mathcal{L}_{\hat{f}} \le 31\theta C_{K,r,2}$. Therefore, there exists at least one block B_k for which simultaneously $P\mathcal{L}_{\hat{f}} \le P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r,2}$ and $P_{B_k}\mathcal{L}_{\hat{f}} \le 31\theta C_{K,r,2}$. This shows that $P\mathcal{L}_{\hat{f}} \le 32\theta C_{K,r,2} \le A^{-1}C_{K,r,2}$.

10.2 Control of the stochastic event

Proposition 5 *Grant Assumptions 2, 3, 6 and 7. Let $K \ge 7|\mathcal{O}|/3$. Then Ω_K holds with probability larger than $1 - 2\exp(-K/504)$.*

Proof Let $\mathcal{F} = F \cap (f^* + \mathcal{B}_\kappa)$ and let $\phi(t) = \mathbb{1}\{t \ge 2\} + (t-1)\mathbb{1}\{1 \le t \le 2\}$. This function satisfies $\forall t \in \mathbb{R}^+ \quad \mathbb{1}\{t \ge 2\} \le \phi(t) \le \mathbb{1}\{t \ge 1\}$. Let $W_k = ((X_i, Y_i))_{i\in B_k}$ and, for any $f \in \mathcal{F}$, let $G_f(W_k) = (P_{B_k} - P)(\ell_f - \ell_{f^*})$. Let also $C_{K,r,\kappa} = \max\left(96L^2K/(\theta^2 N), r_2^2(\gamma, \kappa\rho^*)\right)$. For any $f \in \mathcal{F}$, let

$$z(f) = \sum_{k=1}^{K}\mathbb{1}\{|G_f(W_k)| \le \theta C_{K,r,\kappa}\} \ .$$

Proposition 5 will be proved if $z(f) \geqslant K/2$ with probability larger than $1 - e^{-K/504}$. Let $\mathcal{K}$ denote the set of indices of blocks which have not been corrupted by outliers, $\mathcal{K} = \{k \in \{1, \cdots, K\}$:

$B_k \subset \mathcal{I}\}$, where we recall that $\mathcal{I}$ is the set of informative data. Basic algebraic manipulations show that

$$z(f) \geqslant |\mathcal{K}| - \sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) - \sum_{k\in\mathcal{K}} \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \quad . \tag{59}$$

The last term in (59) can be bounded from below as follows. Let $f \in \mathcal{F}$ and $k \in \mathcal{K}$,

$$\mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \leqslant \mathbb{P}\left(|G_f(W_k)| \geq \frac{\theta C_{K,r,\kappa}}{2}\right) \leqslant \frac{4\mathbb{E}G_f(W_k)^2}{(\theta C_{K,r,\kappa})^2}$$
$$\leqslant \frac{4K^2}{\theta^2 C_{K,r,\kappa}^2 N^2} \sum_{i\in B_k} \mathbb{E}[(\ell_f - \ell_{f^*})^2(X_i,Y_i)] \leq \frac{4L^2K}{\theta^2 C_{K,r,\kappa}^2 N} \|f-f^*\|_{L_2}^2 \quad .$$

The last inequality follows from Assumption 6. Since $\|f - f^*\|_{L_2} \leq \sqrt{C_{K,r,\kappa}}$,

$$\mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \leqslant \frac{4L^2K}{\theta^2 C_{K,r,\kappa} N} \quad .$$

As $C_{K,r,\kappa} \geqslant 96L^2K/(\theta^2 N)$,

$$\mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \leq \frac{1}{24} \quad .$$

Plugging this inequality in (59) yields

$$z(f) \geq |\mathcal{K}|(1 - \frac{1}{24}) - \sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \quad . \tag{60}$$

Using the Mc Diarmid's inequality, with probability larger than $1 - \exp(-|\mathcal{K}|/288)$,we get

$$\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right)$$
$$\leq \frac{|\mathcal{K}|}{24} + \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right) \quad .$$

By the symmetrization lemma, it follows that, with probability larger than $1 - \exp(-|\mathcal{K}|/288)$,

$$\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \right)$$
$$\leqslant \frac{|\mathcal{K}|}{24} + 2\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{K}} \sigma_k \phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) \quad .$$

As ϕ is 1-Lipschitz with $\phi(0) = 0$, the contraction lemma from Ledoux and Talagrand (2013)and yields

$$\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\left(\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2(\theta C_{K,r,\kappa})^{-1}|G_f(W_k)|)\right)$$
$$\leqslant \frac{|\mathcal{K}|}{24} + \frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\sigma_k\frac{G_f(W_k)}{C_{K,r,\kappa}}$$
$$= \frac{|\mathcal{K}|}{24} + \frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\sigma_k\frac{(P_{B_k}-P)(\ell_f-\ell_{f^*})}{C_{K,r,\kappa}}$$

For any $k \in \mathcal{K}$, let $(\sigma_i)_{i\in B_k}$ independent from $(\sigma_k)_{k\in\mathcal{K}}$, $(X_i)_{i\in\mathcal{I}}$ and $(Y_i)_{i\in\mathcal{I}}$. The vectors $(\sigma_i\sigma_k(\ell_f - \ell_{f^*})(X_i,Y_i))_{i,f}$ and $(\sigma_i(\ell_f-\ell_{f^*})(X_i,Y_i))_{i,f}$ have the same distribution. Thus, by the symmetrization and contraction lemmas, with probability larger than $1-\exp(-|\mathcal{K}|/288)$,

$$\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\left(\phi(2C_{K,r,\kappa}^{-1}|G_f(W_k)|) - \mathbb{E}\phi(2C_{K,r,\kappa}^{-1}|G_f(W_k)|)\right)$$
$$\le \frac{|\mathcal{K}|}{24} + \frac{8}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\frac{1}{|B_k|}\sum_{i\in B_k}\sigma_i\frac{(\ell_f-\ell_{f^*})(X_i,Y_i)}{C_{K,r,\kappa}}$$
$$= \frac{|\mathcal{K}|}{24} + \frac{8K}{\theta N}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(\ell_f-\ell_{f^*})(X_i,Y_i)}{C_{K,r,\kappa}}$$
$$\le \frac{|\mathcal{K}|}{24} + \frac{8LK}{\theta N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right| . \qquad (61)$$

Now either 1) $K \le \theta^2 r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$ or 2) $K > \theta^2 r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$. Assume first that $K \le \theta^2 r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$, so $C_{K,r,\kappa} = r_2^2(\gamma,\kappa\rho^*)$ and by definition of the complexity parameter

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right| = \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{r_2^2(\gamma,\kappa\rho^*)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| \le \frac{\gamma|\mathcal{K}|N}{K} .$$

If $K > \theta^2 r_2^2(\gamma,\kappa\rho^*)N/(96L^2)$, $C_{K,r,\kappa} = 96L^2K/(\theta^2 N)$. Write $\mathcal{F} = \mathcal{F}_1\cup\mathcal{F}_2$, where

$$\mathcal{F}_1 := \{f\in\mathcal{F} : \|f-f^*\|_{L_2} \leqslant r_2(\gamma,\kappa\rho^*)\}, \qquad \mathcal{F}_2 = \mathcal{F}\setminus\mathcal{F}_1 .$$

Then,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r,\kappa}}\right|$$
$$= \frac{1}{C_{K,r,\kappa}}\mathbb{E}\left[\sup_{f\in\mathcal{F}_1}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| \vee \sup_{f\in\mathcal{F}_2}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\right] .$$

For any $f \in \mathcal{F}_2$, $g = f^* + (f - f^*) r_2(\gamma, \kappa\rho^*)/\sqrt{C_{K,r,\kappa}} \in \mathcal{F}_1$ and

$$\left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i (f - f^*)(X_i) \right| = \frac{\sqrt{C_{K,r,\kappa}}}{r_2(\gamma, \kappa\rho^*)} \left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i (g - f^*)(X_i) \right| .$$

It follows that

$$\sup_{f \in \mathcal{F}_2} \left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i (f - f^*)(X_i) \right| \leqslant \frac{\sqrt{C_{K,r,\kappa}}}{r_2(\gamma, \kappa\rho^*)} \sup_{f \in \mathcal{F}_1} \left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i (f - f^*)(X_i) \right| .$$

Hence,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i \frac{(f - f^*)(X_i)}{C_{K,r,\kappa}} \right| \leqslant \frac{1}{r_2(\gamma, \kappa\rho^*) \sqrt{C_{K,r,\kappa}}} \mathbb{E} \sup_{f \in \mathcal{F}_1} \left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i (f - f^*)(X_i) \right| .$$

By definition of r_2, this implies

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i \frac{(f - f^*)(X_i)}{C_{K,r,\kappa}} \right| \leqslant \frac{r_2(\gamma, \kappa\rho^*)}{\sqrt{C_{K,r,\kappa}}} \frac{\gamma |\mathcal{K}| N}{K} \leqslant \frac{\gamma |\mathcal{K}| N}{K} .$$

Plugging this bound in (61) yields, with probability larger than $1 - e^{-|\mathcal{K}|/288}$

$$\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \Big(\phi(2C_{K,r,\kappa}^{-1} |G_f(W_k)|) - \mathbb{E}\phi(2C_{K,r,\kappa}^{-1} |G_f(W_k)|) \Big) \leqslant |\mathcal{K}| \left(\frac{1}{24} + \frac{8L\gamma}{\theta} \right) = \frac{|\mathcal{K}|}{12} .$$

Plugging this inequality into (60) shows that, with probability at least $1 - e^{-|\mathcal{K}|/288}$,

$$z(f) \geqslant \frac{7|\mathcal{K}|}{8} .$$

As $K \geqslant 7|\mathcal{O}|/3$, $|\mathcal{K}| \geqslant K - |\mathcal{O}| \geqslant 4K/7$, hence, $z(f) \geqslant K/2$ holds with probability at least $1 - e^{-K/504}$. Since it has to hold for any κ in $\{1, 2\}$, the final probablity is $1 - 2e^{-K/504}$. ■

11. Proof Theorem 3

The proof is very similar to the one of Theorem 2. We only present the different arguments we use coming from the localization with the excess risk. The proof is split into two parts. First we identify an event $\bar{\Omega}_K$ in the same way is Ω_K in (43) where the L_2-localization is replaced by the excess risk localization. For $\kappa \in \{1, 2\}$ let $\mathcal{B}_\kappa = \{f \in E : \ P\mathcal{L}_f \leq \bar{r}^2(\gamma, \kappa\rho^*), \ \|f - f^*\| \leq \ \kappa\rho^*\}$ and

$$\bar{\Omega}_K = \left\{ \forall \kappa \in \{1, 2\}, \forall f \in F \cap \mathcal{B}_\kappa, \sum_{k=1}^{K} I\big\{ |(P_{B_k} - P)\mathcal{L}_f| \leq \frac{1}{20} \bar{r}^2(\gamma, 2\rho^*) \big\} \geq K/2 \right\}$$

Let us us the following notations,

$$\lambda = \frac{11\bar{r}^2(\gamma, 2\rho^*)}{40\rho^*}, \quad \hat{f} = \hat{f}_K^\lambda \quad \text{and} \quad \gamma = 1/3840L$$

Finally recal that the complexity parameter is defined as

$$\bar{r}(\gamma, \rho) = \inf\left\{r > 0 : \max\left(\frac{E(r,\rho)}{\gamma}, \sqrt{384000}V_K(r,\rho)\right) \le r^2\right\}$$

where

$$E(r,\rho) = \sup_{J\subset\mathcal{I}:|J|\ge N/2} \mathbb{E} \sup_{f\in F:P\mathcal{L}_f\le r^2,\ \|f-f^*\|\le\rho} \left|\frac{1}{|J|}\sum_{i\in J}\sigma_i(f-f^*)(X_i)\right|$$

$$V_K(r,\rho) = \max_{i\in\mathcal{I}} \sup_{f\in F:P\mathcal{L}_f\le r^2,\ \|f-f^*\|\le\rho} \left(\sqrt{\mathbb{V}ar_{P_i}(\mathcal{L}_f)}\right)\sqrt{\frac{K}{N}}$$

First, we show that on the event $\bar{\Omega}_K$, $P\mathcal{L}_{\hat{f}} \le \bar{r}^2(\gamma, 2\rho^*)$ and $\|f - f^*\| \le 2\rho^*$. Then we will control the probability of $\bar{\Omega}_K$.

Lemma 4 *Grant Assumptions 2 and 3. Let ρ^* satisfy the sparsity equation from Definition 6. On the event $\bar{\Omega}_K$, $P\mathcal{L}_{\hat{f}} \le \bar{r}^2(\gamma, 2\rho^*)$ and $\|f - f^*\| \le 2\rho^*$.*

Proof Let $f \in F\backslash\mathcal{B}_\kappa$. From Lemma 6 in Chinot et al. (2018) there exist $f_0 \in F$ and $\alpha > 0$ such that $f - f^* = \alpha(f_0 - f^*)$ and $f_0 \in \partial\mathcal{B}_\kappa$. By definition of $\mathcal{B}_\kappa$, either 1)$P\mathcal{L}_{f_0} = \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| \le \kappa\rho^*$ or 2) $P\mathcal{L}_{f_0} \le \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| = \kappa\rho^*$.

Assume that $P\mathcal{L}_{f_0} = \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| \le \kappa\rho^*$. On $\bar{\Omega}_K$, there exist at least $K/2$ blocks B_k such that $P_{B_k}\mathcal{L}_{f_0} \ge P\mathcal{L}_{f_0} - (1/20)\bar{r}^2(\gamma, \kappa\rho^*) = (19/20)\bar{r}^2(\gamma, \kappa\rho^*)$. It follows that, on at least $K/2$ blocks B_k

$$P_{B_k}\mathcal{L}_f^\lambda \ge \alpha P_{B_k}\mathcal{L}_{f_0}^\lambda = \alpha\big(P_{B_k}\mathcal{L}_{f_0} + \lambda(\|f_0\| - \|f^*\|)\big) \ge (19/20)\bar{r}^2(\gamma, \kappa\rho^*) - 11\kappa\bar{r}^2(\gamma, 2\rho^*)/40 \tag{62}$$

Assume that $P\mathcal{L}_{f_0} \le \bar{r}^2(\gamma, \kappa\rho^*)$ and $\|f_0 - f^*\| = \kappa\rho^*$. From the sparsity equation defined in Definition 6 we get $\|f_0\| - \|f^*\| \ge 7\kappa\rho^*/10$. And on more than $K/2$ blocks B_k

$$P_{B_k}\mathcal{L}_f^\lambda \ge -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) + 7\lambda\kappa\rho^*/10 = -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) + 77\kappa\bar{r}^2(\gamma, 2\rho^*)/400 \tag{63}$$

Now let us consider $f \in F \cap \mathcal{B}_\kappa$. On $\bar{\Omega}_K$, there exist at least $K/2$ blocks B_k such that

$$P_{B_k}\mathcal{L}_f^\lambda \ge -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) - \lambda\kappa\rho^* = -(1/20)\bar{r}^2(\gamma, \kappa\rho^*) - 11\kappa\bar{r}^2(\gamma, 2\rho^*)/40 \tag{64}$$

As Equations (62), (63) and (64) hold for more than $K/2$ blocks it follows for $\kappa = 1$ that

$$\sup_{f\in F} \text{MOM}_K\left[\ell_{f^*} - \ell_f\right] + \lambda(\|f^*\| - \|f\|) \le (13/40)\bar{r}^2(\gamma, 2\rho^*) \ . \tag{65}$$

From Equations (62), (63) and (64) with $\kappa = 2$ we get

$$\sup_{f\in F\setminus\mathcal{B}_2} \text{MOM}_K\left[\ell_{f^*}-\ell_f\right]+\lambda(\|f^*\|-\|f\|) < (13/40)\bar{r}^2(\gamma,2\rho^*) \ . \tag{66}$$

From Equations (65) and (66) and a slight modification of Lemma 3 it easy to see that on $\bar{\Omega}_K$, $P\mathcal{L}_{\hat{f}} \le \bar{r}^2(\gamma,2\rho^*)$ and $\|f-f^*\|\le\rho^*$. ■

Proposition 6 *Grant Assumptions 2, 3 and 8. Then $\bar{\Omega}_K$ holds with probability larger than $1-2\exp(-cK)$*

Sketch of proof. The proof of Proposition 6 follows the same line as the one of Proposition 5. Let us precise the main differences. For all $f\in F\cap\mathcal{B}_\kappa$ we set, $z'(f)=\sum_{k=1}^K I\{|G_f(W_k)|\le(1/20)\bar{r}^2(\gamma,\kappa\rho^*)\}$ where $G_f(W_k)$ is the same quantity as in the proof of Proposition 5. Let us consider the contraction ϕ introduced in Proposition 5. By definition of $V_K(r)$ and $\bar{r}^2(\gamma,\kappa\rho^*)$ we have

$$\begin{aligned}
&\mathbb{E}\phi(40|G_f(W_k)|/\bar{r}^2(\gamma,\kappa\rho^*)) \le \mathbb{P}\left(|G_f(W_k)|\ge\frac{\bar{r}^2(\gamma,\kappa\rho^*)}{40}\right)\le\frac{(40)^2}{\bar{r}^4(\gamma,\kappa\rho^*)}\mathbb{E}G_f(W_k)^2\\
&=\frac{(40)^2}{\bar{r}^4(\gamma,\kappa\rho^*)}\mathbb{V}ar(P_{B_k}\mathcal{L}_f)\le\frac{(40)^2K^2}{\bar{r}^4(\gamma,\kappa\rho^*)N^2}\sum_{i\in B_k}\mathbb{V}ar_{P_i}(\mathcal{L}_f)\\
&\le\frac{(40)^2K}{\bar{r}^4(\gamma,\kappa\rho^*)N}\sup\{\mathbb{V}ar_{P_i}(\mathcal{L}_f): f\in F\cap\mathcal{B}_\kappa, i\in\mathcal{I}\}\le 1/24 \ .
\end{aligned}$$

Using Mc Diarmid's inequality, the Giné-Zinn symmetrization argument and the contraction lemma twice and the Lipschitz property of the loss function, such as in the proof of Proposition 5, we obtain for all $x>0$, with probability larger than $1-\exp(-|\mathcal{K}|/288)$, for all $f\in\mathcal{F}'$,

$$z'(f)\ge 11|\mathcal{K}|/12-\frac{160LK}{\theta N}\mathbb{E}\sup_{f\in F\cap\mathcal{B}_\kappa}\frac{1}{\bar{r}^2(\gamma,\kappa\rho^*)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|. \tag{67}$$

From the definition of $\bar{r}^2(\gamma,\kappa\rho^*)$ it follows that $\mathbb{E}\sup_{f\in F\cap\mathcal{B}_\kappa}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\le\gamma\bar{r}^2(\gamma,\kappa\rho^*)$ and $z'(f)\ge|\mathcal{K}|(11/12-160L^2\gamma)=7|\mathcal{K}|/8$. The rest of the proof is totally similar.

11.1 Proof of Theorem 4

From Assumption 2, it holds $V_K(r)\le LV_K'(r)$, where for all $r>0$,

$$V_K'(r)=\sqrt{K/N}\max_{i\in\mathcal{I}}\sup_{f\in F: P\mathcal{L}_f\le r^2,\ \|f-f^*\|\le\rho}\|f-f^*\|_{L_2} \ .$$

By Assumption 9,

$$\sqrt{c}V_K\left(\sqrt{384000}L\sqrt{\frac{\bar{A}K}{N}},2\rho^*\right)\le 384000L^2\frac{\bar{A}K}{N} \ .$$

From the definition of $r_2^2(\gamma, 2\rho^*)$ and Assumption 9, it follows

$$\frac{1}{\gamma} E\left(\frac{r_2(\gamma/\bar{A}, 2\rho^*)}{\sqrt{\bar{A}}}\right) \leq \frac{r_2^2(\gamma/\bar{A}, 2\rho^*)}{\bar{A}} \quad .$$

Hence, $\bar{r}^2(\gamma, 2\rho^*) \leq \max\left(r_2^2(\gamma/\bar{A}, 2\rho^*)/\sqrt{\bar{A}}, 384000L^2\bar{A}K/N\right)$ and the proof is complete.

References

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. ISSN 0022-0000. doi: 10.1006/jcss.1997.1545. URL http://dx.doi.org/10.1006/jcss.1997.1545. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

P. Alquier, V. Cottet, and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *to appear in Ann. Statist., arXiv preprint arXiv:1702.01402*, 2017.

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294, 2014. ISSN 2049-8764. doi: 10.1093/imaiai/iau005. URL https://doi.org/10.1093/imaiai/iau005.

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5): 2766–2794, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS918. URL http://dx.doi.org/10.1214/11-AOS918.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statist. Sci.*, 27(4):450–468, 2012. ISSN 0883-4237. doi: 10.1214/12-STS394. URL https://doi.org/10.1214/12-STS394.

Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017. ISSN 0020-9910. doi: 10.1007/s00222-016-0673-5. URL https://doi.org/10.1007/s00222-016-0673-5.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006. ISSN 0178-8051. URL https://doi.org/10.1007/s00440-005-0462-3.

Pierre C Bellec. Localized gaussian width of m-convex hulls with applications to lasso and convex aggregation. *arXiv preprint arXiv:1705.10696*, 2017.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Towards the study of least squares estimators with convex penalty. *In Séminaire et Congrès, number 31. Société mathématique de France*, 2017.

Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. *Ann. Statist.*, 46(6B):3603–3642, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1670. URL https://doi.org/10.1214/17-AOS1670.

Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Process.*, 61(23):5987–5999, 2013. ISSN 1053-587X. doi: 10.1109/TSP.2013.2273443. URL https://doi.org/10.1109/TSP.2013.2273443.

Lucien Birgé. Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.*, 20(3):201–223, 1984. ISSN 0246-0203. URL http://www.numdam.org/item?id=AIHPB_1984__20_3_201_0.

Mał gorzata 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. ISSN 1932-6157. doi: 10.1214/15-AOAS842. URL https://doi.org/10.1214/15-AOAS842.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2. doi: 10.1007/978-3-642-20192-9. URL https://doi.org/10.1007/978-3-642-20192-9. Methods, theory and applications.

T Tony Cai, Zhao Ren, Harrison H Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

E. J. Candès and Y. Plan. Matrix Completion With Noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2044061. URL https://doi.org/10.1109/TIT.2010.2044061.

Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*. Citeseer, 2012.

Sabyasachi Chatterjee and Subhajit Goswami. New risk bounds for 2d total variation denoising. *arXiv preprint arXiv:1902.01215*, 2019.

G. Chinot, G. Lecué, and M. Lerasle. Statistical learning with lipschitz and convex loss functions. *arXiv:1810.01090*, 2018.

Geoffrey Chinot. Robust learning and complexity dependent bounds for regularized problems. *arXiv preprint arXiv:1902.02238*, 2019.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *Ann. Statist.*, 46(6B):3481–3509, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1666. URL https://doi.org/10.1214/17-AOS1666.

Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.

Yehoram Gordon, Alexander E Litvak, Shahar Mendelson, and Alain Pajor. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *Journal of Approximation Theory*, 149(1):59–73, 2007.

P. J. Huber and E. Ronchetti. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

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. ISSN 0304-3975. doi: 10.1016/0304-3975(86)90174-X. URL http://dx.doi.org/10.1016/0304-3975(86)90174-X.

O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1): 282–303, 2014.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. ISBN 978-3-642-22146-0. URL https://doi.org/10.1007/978-3-642-22147-7. 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].

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS894. URL http://dx.doi.org/10.1214/11-AOS894.

G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. *To appear in Stochastic Processes and their applications*, 2017.

G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *to appear in Ann. Statist. arXiv preprint arXiv:1711.10306*, 2019.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. *J. Mach. Learn. Res.*, 18:Paper No. 146, 48, 2017. ISSN 1532-4435.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1562. URL https://doi.org/10.1214/17-AOS1562.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

T. T. Mai and P. Alquier. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9:823–841, 2015.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6): 1808–1829, 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939240. URL https://doi.org/10.1214/aos/1017939240.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

Shahar Mendelson. On multiplier processes under weak moment assumptions. In *Geometric Aspects of Functional Analysis*, pages 301–318. Springer, 2017.

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. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2): 460–489, 2005.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. ISBN 0-691-01586-4. Reprint of the 1970 original, Princeton Paperbacks.

Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892, 2011.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. ISBN 978-3-642-54074-5; 978-3-642-54075-2. doi: 10.1007/978-3-642-54075-2. URL https://doi.org/10.1007/978-3-642-54075-2. Modern methods and classical problems.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12 (4):389–434, 2012. ISSN 1615-3375. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32 (1):135–166, 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120131. URL https://doi.org/10.1214/aos/1079120131.

Sara van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham], 2016. ISBN 978-3-319-32773-0; 978-3-319-32774-7. doi: 10.1007/978-3-319-32774-7. URL https://doi.org/10.1007/978-3-319-32774-7. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971. ISSN 0040-361x.

Vladimir Naumovich Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Yi Yang and Hui Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.

Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust M-estimation: finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.*, 46(5):1904–1931, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1606. URL https://doi.org/10.1214/17-AOS1606.