Statistical learning with Lipschitz and convex loss functions

Geoffrey Chinot, Guillaume Lecué and Matthieu Lerasle

October 3, 2018

Abstract

We obtain risk bounds for Empirical Risk Minimizers (ERM) and minmax Median-Of-Means (MOM) estimators based on loss functions that are both Lipschitz and convex.

Results for the ERM are derived without assumptions on the outputs and under subgaussian assumptions on the design as in [2] but relaxing the global Bernstein condition of this paper into a local assumption as in [40]. Similar results are shown for minmax MOM estimators in a close setting where the design is only supposed to satisfy moment assumptions, relaxing the subgaussian hypothesis necessary for ERM. Unlike alternatives based on MOM's principle [24, 29], the analysis of minmax MOM estimators is not based on the small ball assumption (SBA) of [22]. In particular, the basic example of non parametric statistics where the learning class is the linear span of localized bases, that does not satisfy SBA [37] can now be handled.

Finally, minmax MOM estimators are analyzed in a setting where the local Bernstein condition is also dropped out. It is shown to achieve an oracle inequality with exponentially large probability under minimal assumptions insuring the existence of all objects.

1 Introduction

The paper studies learning problems where the loss function is simultaneously Lipschitz and convex. This situation happens in classical examples such as quantile, Huber and L_1 regression or logistic and hinge classification [40]. As the Lipschitz property allows to remove all assumptions on the outputs, these losses have been quite popular in robust statistics [17]. Empirical risk minimizers (ERM) based on Lipschitz losses such as the Huber loss have received recently an important attention [44, 15, 2].

Based on a dataset $\{(X_i, Y_i) : i = 1, \dots, N\}$ of points in $\mathcal{X} \times \mathcal{Y}$, a class F of functions and a risk function $R(\cdot)$ defined on F, the statistician should estimate an oracle $f^* \in \operatorname{argmin}_{f \in F} R(f)$. The risk function $R(\cdot)$ is often defined as the expectation of a known loss function $\ell : (f, x, y) \in F \times \mathcal{X} \times \mathcal{Y} \to \ell_f(x, y) \in \mathbb{R}$ with respect to the unknown distribution P of a random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$: $R(f) = \mathbb{E}\ell_f(X, Y)$. Hereafter, the risk is assumed to have this form for a loss function ℓ such that, for any (f, x, y), $\ell_f(x, y) = \bar{\ell}(f(x), y)$, for some function $\bar{\ell} : \bar{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, where the set $\bar{\mathcal{Y}}$ is a convex set containing all possible values of $f(x)$. ℓ is said Lipschitz and convex when the following assumption holds.

Assumption 1. *There exists $L > 0$ such that, for any $y \in \mathcal{Y}$, $\bar{\ell}(\cdot, y)$ is L-Lipschitz and convex.*

Many classical loss functions satisfy Assumption 1 as shown by the following examples.

- The **logistic loss** defined, for any $u \in \bar{\mathcal{Y}} = \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\bar{\ell}(u, y) = \log(1 + \exp(-yu))$ satisfies Assumption 1 with $L = 1$.
- The **hinge loss** defined, for any $u \in \bar{\mathcal{Y}} = \mathbb{R}$ and $y \in \mathcal{Y} = \{-1, 1\}$, by $\bar{\ell}(u, y) = \max(1 - uy, 0)$ satisfies Assumption 1 with $L = 1$.
- The **Huber loss** defined, for any $\delta > 0$, $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by

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

satisfies Assumption 1 with $L = \delta$.

- The **quantile loss** is defined, for any $\tau \in (0,1)$, $u, y \in \mathcal{Y} = \bar{\mathcal{Y}} = \mathbb{R}$, by $\bar{\ell}(u,y) = \rho_\tau(u-y)$ where, for any $z \in \mathbb{R}$, $\rho_\tau(z) = z(\tau - I\{z \le 0\})$. It satisfies Assumption 1 with $L = 1$. For $\tau = 1/2$, the quantile loss is the L_1 loss.

All along the paper, the following assumption is also granted.

Assumption 2. *The class F is convex.*

When (X, Y) and the data $((X_i, Y_i))_{i=1}^N$ are independent and identically distributed (i.i.d.), for any $f \in F$, the empirical risk $R_N(f) = (1/N)\sum_{i=1}^N \ell_f(X_i, Y_i)$ is a natural estimator of $R(f)$. The empirical risk minimizers (ERM) [41] obtained by minimizing $f \in F \to R_N(f)$ are expected to be close to the oracle f^*. This procedure and its regularized versions have been extensively studied in learning theory [20]. When the loss is both convex and Lipschitz, results have been obtained in practice [4, 12] and theory [40]. Risk bounds with exponential deviation inequalities for the ERM can be obtained without assumptions on the outputs Y, but stronger assumptions on the design X, which must satisfy subgaussian or boundedness assumptions. Moreover, fast rates of convergence [39] can only be obtained under margin type assumptions such as the Bernstein condition [8, 40].

The Lipschitz assumption and global Bernstein conditions (that hold over the entire F as in [2]) imply boundedness in L_2-norm of the class F, see the discussion preceding Assumption 4 for details. This boundedness is not satisfied in linear regression with unbounded design so the results of [2] don't apply to this basic example. To bypass this restriction, the global condition has to be relaxed into a "local" one as in [15, 40], see also Assumption 4 below.

The main constraint in our results on ERM is the subgaussian assumption on the design. This constraint can be relaxed by considering alternative estimators based on the "median-of-means" (MOM) principle of [35, 9, 18, 1] and the minmax procedure of [3, 5]. The resulting minmax MOM estimators have been introduced in [24] for least-squares regression as an alternative to other MOM based procedures [29, 30, 31, 23]. In the case of convex and Lipschitz loss functions, these estimators satisfy the following properties 1) as the ERM, they are efficient without assumptions on the noise 2) they achieve optimal rates of convergence under weak stochastic assumptions on the design, relaxing the subgaussian or boundedness hypotheses used for ERM and 3) the rates are not downgraded by the presence of some outliers in the dataset.

These improvements of MOM estimators upon ERM are not surprising. For univariate mean estimation, rate optimal subgaussian deviation bounds can be shown under minimal L_2 moment assumptions for MOM estimators [14] while the empirical mean needs each data to have subgaussian tails to achieve such bounds [13]. In least-squares regression, MOM-based estimators [29, 30, 31, 23, 24] inherit these properties, whereas the ERM has downgraded statistical properties under moment assumptions [26]. Furthermore, MOM procedures are resistant to outliers: results hold in the "$\mathcal{O} \cup \mathcal{I}$" framework of [23, 24], where inliers or informative data (indexed by $\mathcal{I}$) only satisfy weak moments assumptions and the dataset may contain outliers (indexed by $\mathcal{O}$) on which no assumption is made, see Section 3. This robustness, that almost comes for free from a technical point of view, is another important advantage of MOM estimators compared to ERM in practice. Figure 1 [1] illustrates this fact, showing that statistical performance of the standard logistic regression are strongly affected by a single corrupted observation, while the minmax MOM estimator maintains good statistical performance even with 5% of corrupted data.

Compared to [29, 24], considering convex-Lipschitz losses instead of the square loss allows to simplify simultaneously some assumptions and the presentation of the results for MOM estimators: L_2-assumptions on the noise in [29, 24] can be removed and complexity parameters driving risk of ERM and MOM estimators only involve a single stochastic linear process, see Eq. (3) and (6) below. Also, contrary to the analysis in least-squares regression, the small ball assumption [22, 33] is not required here. Recall that this assumption states that there are absolute constants κ and β such that, for all $f \in F$, $\mathbb{P}[|f(X) - f^*(X)| \ge$

[1] All figures can be reproduced from the code available at https://github.com/lecueguillaume/MOMpower

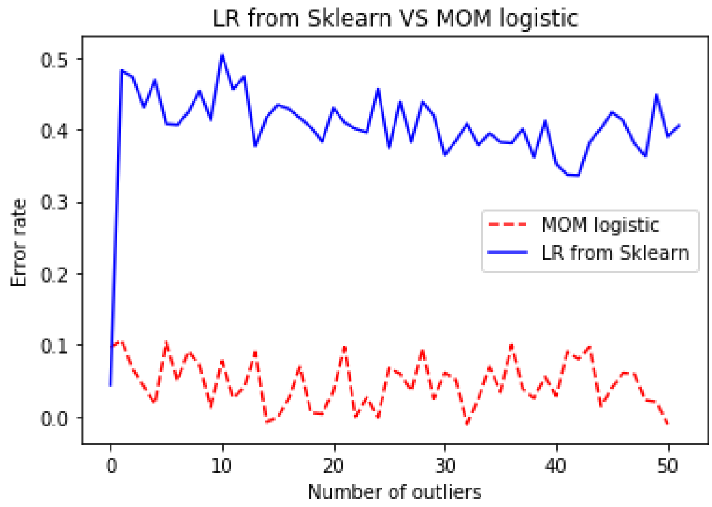

Figure 1: MOM Logistic Regression VS Logistic regression from Sklearn ($p = 50$ and $N = 1000$)

$\kappa \|f - f^*\|_{L_2}] \geq \beta$. It is interesting as it involves only moments of order 1 and 2 of the functions in F. However, it does not hold in classical frameworks such as histograms, see [37, 16] and Section 5.

Finally, minmax MOM estimators are studied in a framework where the Bernstein condition is dropped out. In this setting, they are shown to achieve an oracle inequality with exponentially large probability (see Section 4). The results are slightly weaker in this relaxed setting: the excess risk is bounded but not the L_2 risk and the rates of convergence are "slow" in $1/\sqrt{N}$ in general. Fast rates of convergence in $1/N$ can still be recovered from this general result if a local Bernstein type condition is satisfied though, see Section 4 for details. This last result shows that minmax MOM estimators can be safely used with Lipschitz and convex losses, assuming only that inliers data are independent with moments giving sense to all objects necessary to state the results.

To approximate minmax MOM estimators, an algorithm inspired from [24, 27] is also proposed. Convergence of this algorithm has been proved in [27] under strong assumptions, but, to the best of our knowledge, convergence rates have not been established. Nevertheless, the simulation study presented in Section 6 shows that the algorithm presents interesting empirical performance in the setting of this paper.

The paper is organized as follows. Optimal results for the ERM are presented in Section 2. Minmax MOM estimators are introduced and analysed in Section 3 under a local Bernstein condition and in Section 4 without the Bernstein condition. A discussion of the main assumptions is provided in Section 5 while Section 6 provides a typical example where all assumptions are satisfied and a simulation study where a natural algorithm associated to the minmax MOM estimator for logistic loss is presented. The proofs of the main theorems are gathered in Sections A, B and C.

Notations Let $\mathcal{X}, \mathcal{Y}$ be measurable spaces. Let F be a class of measurable functions $f : \mathcal{X} \to \mathcal{Y}$ and let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random variable with distribution P. Let μ denote the marginal distribution of X. For any probability measure Q on $\mathcal{X} \times \mathcal{Y}$, and any function $g \in L_1(Q)$, let $Qg = \int g(x, y)Q(\mathrm{d}(x, y))$. Let $\ell : F \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $(f, x, y) \mapsto \ell_f(x, y)$ denote a loss function measuring the error made when predicting y by $f(x)$. It is always assumed that there exist $\bar{\mathcal{Y}}$ a convex set containing all possible values of $f(x)$ for $f \in F$, $x \in \mathcal{X}$ and a function $\bar{\ell} : \bar{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ such that, for any $(f, x, y) \in F \times \mathcal{X} \times \mathcal{Y}$, $\bar{\ell}(f(x), y) = \ell_f(x, y)$. Let $R(f) = P\ell_f = \mathbb{E}\ell_f(X, Y)$ for f in F denote the risk and let $\mathcal{L}_f = \ell_f - \ell_{f^*}$ denote the excess loss. If $F \subset L_1(P) := L_1$ and Assumption 1 holds, an equivalent risk can be defined even if $Y \notin L_1$. Actually, for any $f_0 \in F$, $\ell_f - \ell_{f_0} \in L_1$ so one can define $R(f) = P(\ell_f - \ell_{f_0})$. W.l.o.g. the set of risk minimizers is assumed to be reduced to a singleton $\mathrm{argmin}_{f\in F} R(f) = \{f^*\}$. f^* is called the oracle as $f^*(X)$ provides the prediction of Y with minimal risk among functions in F. For any f and $p > 0$, let $\|f\|_{L_p} = (P|f|^p)^{1/p}$, for any $r \geqslant 0$, let $rB_{L_2} = \{f \in F : \|f\|_{L_2} \leq r\}$ and $rS_{L_2} = \{f \in F : \|f\|_{L_2} = r\}$. For any set H for which it makes sense, $H + f^* = \{h + f^* \text{ s.t } h \in H\}$, $H - f^* = \{h - f^* \text{ s.t } h \in H\}$.

2 ERM in the subgaussian framework

This section studies the ERM, improving some results from [2]. In particular, the global Bernstein condition in [2] is relaxed into a local hypothesis following [40]. All along this section, data $(X_i, Y_i)_{i=1}^N$ are independent and identically distributed with common distribution P. The ERM is defined for $f \in F \to P_N \ell_f = (1/N)\sum_{i=1}^N \ell_f(X_i, Y_i)$ by

$$\hat{f}^{ERM} = \arg\min_{f \in F} P_N \ell_f \quad . \tag{1}$$

The results for the ERM are shown under a subgaussian assumption on the design. This result is the benchmark for the following minmax MOM estimators.

Definition 1. *Let $B \geq 1$. F is called B-subgaussian (with respect to X) when for all $f \in F$ and all $\lambda > 1$*

$$\mathbb{E}\exp(\lambda |f(X)|/\|f\|_{L_2}) \leq \exp(\lambda^2 B^2/2) \enspace .$$

Assumption 3. *The class $F - F$ is B-subgaussian, where $F - F = \{f_1 - f_2 : f_1, f_2 \in F\}$.*

Under this subgaussian assumption, statistical complexity can be measured via Gaussian mean-widths.

Definition 2. *Let $H \subset L_2$. Let $(G_h)_{h \in H}$ be the canonical centered Gaussian process indexed by H (in particular, the covariance structure of $(G_h)_{h\in H}$ is 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 $w(H) = \mathbb{E}\sup_{h \in H} G_h$.*

The complexity parameter driving the performance of $\hat{f}^{ERM}$ is presented in the following definition.

Definition 3. *The* ***complexity parameter*** *is defined as*

$$r_2(\theta) = \inf\{r > 0 : 32Lw((F - f^*) \cap rB_{L_2}) \leq \theta r^2 \sqrt{N}\}$$

where $L > 0$ is the Lipschitz constant from Assumption 1.

Let $A > 0$. In [8], the class F is called $(1, A)$-Bernstein if, for all $f \in F$, $P\mathcal{L}_f^2 \leqslant AP\mathcal{L}_f$. Under Assumption 1, F is $(1, AL^2)$ Bernstein if the following stronger assumption is satisfied

$$\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f \quad . \tag{2}$$

This stronger version was used, for example in [2] to study ERM. However, under Assumption 1, Eq (2) implies that

$$\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f \leqslant AL\|f - f^*\|_{L_1} \leqslant AP\mathcal{L}_f \leqslant AL\|f - f^*\|_{L_2} \quad .$$

Therefore, $\|f - f^*\|_{L_2} \leqslant AL$ for any $f \in F$. The class F is bounded in L^2-norm, which is restrictive as this assumption is not verified by the class of linear functions for example. To bypass this issue, the following condition is introduced.

Assumption 4. *There exists a constant $A > 0$ such that for all $f \in F$ if $\|f - f^*\|_{L_2} \leq r_2(\theta)$ then $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$.*

In Assumption 4, Bernstein condition is granted in a L_2-neighborhood of f^* only. Outside of this neighborhood, there is no restriction on the excess loss. This relaxed assumption is satisfied on linear models and Lipschitz-convex loss functions under moment assumptions as will be checked in Section 5. The following theorem is the main result of this section.

Theorem 1. *Grant Assumptions 1, 2, 3 and 4 and let $\theta = 1/(2A)$, $\hat{f}^{ERM}$ defined in (1) satisfies, with probability larger than*

$$1 - 2\exp\left(-\frac{\theta^2 N r_2^2(\theta)}{16^3 L^2}\right) , \tag{3}$$

$$\|\hat{f}^{ERM} - f^*\|_{L_2}^2 \leq r_2^2(\theta) \text{ and } P\mathcal{L}_{\hat{f}^{ERM}} \leq \theta r_2^2(\theta) \ . \tag{4}$$

Theorem 1 is proved in Section A.1. It holds without restrictions on the outputs Y. Theorem 1 shows deviation bounds both in L_2 norm and for the excess risk, which are both minimax optimal as proved in [2]. As in [2], a similar result can be derived if the subgaussian Assumption 3 is replaced by a boundedness in L_∞ assumption. An extension of Theorem 1 can be shown, where Assumption 4 is replaced by the following hypothesis: there exists κ such that for all $f \in F$ in a L_2-neighborhood of f^*, $\|f - f^*\|_{L_2}^{2\kappa} \leqslant AP\mathcal{L}_f$. The case $\kappa = 1$ is the most classical and its analysis contains all the ingredients for the study of the general case with any parameter $\kappa \geq 1$. More general Bernstein conditions can also be considered as in [40, Chapter 7]. These extensions are left to the interested reader.

3 Minmax MOM estimators

This section presents and studies minmax MOM estimators, comparing them to ERM.

3.1 The estimators

The framework of this section is a relaxed version of the i.i.d. setup considered in Section 2. Following [23, 24], there exists a partition $\mathcal{O} \cup \mathcal{I}$ of $\{1, \cdots, N\}$ in two subsets which is unknown to the statistician. No assumption is granted on the set of "outliers" $(X_i, Y_i)_{i\in\mathcal{O}}$. "Inliers", indexed by $\mathcal{I}$, are only assumed to satisfy the following assumption.

Assumption 5. *$(X_i, Y_i)_{i\in\mathcal{I}}$ are independent and for all $i \in \mathcal{I}$, (X_i, Y_i) has distribution P_i, X_i has distribution μ_i and for any $p > 0$ and any function g for which it makes sense $\|g\|_{L_p(\mu_i)} = (P_i|g|^p)^{1/p}$. We assume that, for any $i \in \mathcal{I}$, $\|f - f^*\|_{L_2} = \|f - f^*\|_{L_2(\mu_i)}$ and $P_i\mathcal{L}_f = P\mathcal{L}_f$.*

Assumption 5 holds in the i.i.d case but it covers other situations where informative data $(X_i, Y_i)_{i\in\mathcal{I}}$ may have different distributions. Recall the definition of MOM estimators of univariate means. Let $(B_k)_{k=1,\dots,K}$ denote a partition of $\{1, \dots, N\}$ into blocks B_k of equal size N/K (if N is not a multiple of K, just remove some data). For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $k \in \{1, \dots, K\}$, let $P_{B_k} f = (K/N)\sum_{i\in B_k} f(X_i, Y_i)$. MOM estimator is the median of these empirical means:

$$\text{MOM}_K\big(f\big) = \text{Med}(P_{B_1} f, \cdots, P_{B_K} f) \ .$$

The estimator $\text{MOM}_K\big(f\big)$ achieves rate optimal subgaussian deviation bounds, assuming only that $Pf^2 < \infty$, see for example [14]. The number K is a tuning parameter. The larger K, the more outliers is allowed. When $K = 1$, $\text{MOM}_K\big(f\big)$ is the empirical mean, when $K = N$, the empirical median.

Following [24], remark that the oracle is also solution of the following minmax problem:

$$f^* \in \arg\min_{f\in F} P\ell_f = \arg\min_{f\in F}\sup_{g\in F} P(\ell_f - \ell_g) \ .$$

Minmax MOM estimators are obtained by plugging MOM estimators of the unknown expectations $P(\ell_f - \ell_g)$ in this formula:

$$\hat{f} \in \arg\min_{f\in F}\sup_{g\in F} \text{MOM}_K\big(\ell_f - \ell_g\big) \ . \tag{5}$$

The minmax MOM construction can be applied systematically as an alternative to ERM. For instance, it yields a robust version of logistic classifiers. The minmax MOM estimator with $K = 1$ is the ERM.

The linearity of the empirical process P_N is important to use localisation technics and derive “fast rates” of convergence for ERM [21], improving “slow rates” derived with the approach of [42], see [39] for details on “fast and slow rates”. The idea of the minmax reformulation comes from [3], where this strategy allows to overcome the lack of linearity of some alternative robust mean estimators. [23] introduced minmax MOM estimators to least-squares regression.

3.2 Theoretical results

3.2.1 Setting

The assumptions required for the study of estimator (5) are essentially those of Section 2 except for the subgaussian Assumption 3 which is replaced by the moment Assumption 5. Instead of Gaussian mean width, the complexity parameter is expressed as a fixed point of local Rademacher complexities [7, 10, 6]. Let $(\sigma_i)_{i=1,\dots,N}$ denote i.i.d. Rademacher random variables (uniformly distributed on $\{-1,1\}$), independent from $(X_i,Y_i)_{i\in\mathcal{I}}$. Let

$$\tilde{r}_2(\gamma)=\inf\left\{r>0,\forall J\subset\mathcal{I}:|J|\geq\frac{N}{2},\mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\leq r}\left|\sum_{i\in J}\sigma_i(f-f^*)(X_i)\right|\leq r^2|J|\gamma\right\}\ . \tag{6}$$

The outputs do not appear in the complexity parameter. This is an interesting feature of Lipschitz losses. Since Assumption 4 depends on the complexity parameter in the subgaussian framework, it is necessary to adapt the Bernstein assumption to this framework.

Assumption 6. *There exists a constant $A>0$ such that for all $f\in F$ if $\|f-f^*\|_{L_2}^2\leqslant C_{K,r}$ then $\|f-f^*\|_{L_2}^2\leqslant AP\mathcal{L}_f$ where*

$$C_{K,r}=\max\left(\tilde{r}_2^2(\gamma),864A^2L^2\frac{K}{N}\right)\ . \tag{7}$$

Assumptions 6 and 4 have a similar flavor as both require the Bernstein condition in a L_2 neighborhood of f^* with radius given by the rate of convergence of the associated estimator (see Theorems 1 and 2). For $K\leq(\tilde{r}_2^2(\gamma)N)/(846A^2L^2)$ the neighborhood $\{f\in F:||f-f^*||_{L_2}\leq\sqrt{C_{K,r}}\}$ is a L_2-ball around f^* of radius $\tilde{r}_2(\gamma)$ which can be of order $1/\sqrt{N}$ (see Section 3.2.3). As a consequence, Assumption 6 holds in examples where the small ball assumption does not (see discussion after Assumption 9).

3.2.2 Main results

We are now in position to state the main result regarding the statistical properties of estimator (5) under a local Bernstein condition.

Theorem 2. *Grant Assumptions 1, 2, 5 and 6 and assume that $|\mathcal{O}|\leq3N/7$. Let $\gamma=1/(575AL)$ and $K\in\left[7|\mathcal{O}|/3,N\right]$. The minmax MOM estimator $\hat{f}$ defined in (5) satisfies, with probability at least*

$$1-\exp(-K/2016), \tag{8}$$

$$\|\hat{f}-f^*\|_{L_2}^2\leq C_{K,r}\ \text{and}\ P\mathcal{L}_{\hat{f}}\leq\frac{2}{3A}C_{K,r}\ . \tag{9}$$

Suppose that $K=\tilde{r}_2^2(\gamma)N$, which is possible as long as $|\mathcal{O}|\lesssim N\tilde{r}_2^2(\gamma)$. The deviation bound is then of order $\tilde{r}_2^2(\gamma)$ and the probability estimate $1-\exp(-N\tilde{r}_2^2(\gamma)/2016)$. Therefore, minmax MOM estimators achieve the same statistical bound with the same deviation as the ERM as long as $\tilde{r}_2^2(\gamma)$ and $r_2(\theta)$ are of the same order. Using generic chaining [38], this comparison is true under Assumption 3. It can also be shown under weaker moment assumption, see [34] or the example of Section 3.2.3.

When $\tilde{r}_2^2(\gamma)\asymp r_2(\theta)$, the bounds are rate optimal as shown in [2]. This is why these bounds are called rate optimal subgaussian deviation bounds. While these hold for ERM in the i.i.d. setup with subgaussian design in the absence of outliers (see Theorem 1), they hold for minmax MOM estimators in a setup where inliers may not be i.i.d., nor have subgaussian design and up to $N\tilde{r}_2^2(\gamma)$ outliers may have contaminated the dataset.

This section is concluded by presenting an estimator achieving (5) simultaneously for all K. For all $K\in\{1,\dots,N\}$ and $f\in F$, define $T_K(f)=\sup_{g\in F}\text{MOM}_K\big(\ell_f-\ell_g\big)$ and let

$$\hat{R}_K=\{g\in F: T_K(g)\leq (1/3A)C_{K,r}\} \quad . \tag{10}$$

Now, building on the Lepskii's method, define a data-driven number of blocks

$$\hat{K}=\inf\left(K\in\{1,\dots,N\}:\bigcap_{J=K}^{N}\hat{R}_J\neq\emptyset\right) \tag{11}$$

and let $\tilde{f}$ such that

$$\tilde{f}\in\bigcap_{J=\hat{K}}^{N}\hat{R}_J \quad . \tag{12}$$

Theorem 3. *Grant Assumptions 1, 2, 5 and 6 and assume that $|\mathcal{O}|\leq 3N/7$. Let $\gamma=1/(575AL)$. The estimator $\tilde{f}$ defined in (12) is such that for all $K\in\big[7|\mathcal{O}|/3,N\big]$, with probability at least*

$$1-4\exp(-K/2016),$$

$$\|\tilde{f}-f^*\|_{L_2}^2\leq C_{K,r} \text{ and } P\mathcal{L}_{\tilde{f}}\leq\frac{2}{3A}C_{K,r} \quad .$$

Theorem 3 states that $\tilde{f}$ achieves the results of Theorem 2 simultaneously for all K. This extension is useful as the number $|\mathcal{O}|$ of outliers is typically unknown in practice. However, contrary to $\hat{f}$, the estimator $\tilde{f}$ requires the knowledge of A and $\tilde{r}(\gamma)$. This kind of limitation is not surprising, it holds in least-squares regression [24] and comes from the univariate mean estimation case [14, Theorem 3.2].

3.2.3 A basic example

The following example illustrates the optimality of the rates provided in Theorem 2.

Lemma 1 ([19]). *In the $\mathcal{O}\cup\mathcal{I}$ framework with $F=\{\big\langle t,\cdot\big\rangle: t\in\mathbb{R}^p\}$, we have $\tilde{r}_2^2(\gamma)\leq Rank(\Sigma)/(2\gamma^2N)$, where $\Sigma=\mathbb{E}[XX^T]$ is the $p\times p$ covariance matrix of X.*

The proof of Lemma 1 is recalled in Section B for the sake of completeness.

Section 5 shows that Assumptions 4 and 6 are satisfied when $F=\{\big\langle t,\cdot\big\rangle: t\in\mathbb{R}^p\}$ and X is a vector with i.i.d. entries having only a few finite moments. Theorem 2 applies therefore in this setting and the Minmax MOM estimator (5) achieves the optimal fast rate of convergence $\text{Rank}(\Sigma)/N$.

4 Relaxing the Bernstein condition

This section shows that minmax MOM estimators satisfy sharp oracle inequalities with exponentially large deviation under minimal stochastic assumptions insuring the existence of all objects. These results are slightly weaker than those of the previous section: the L_2 risk is not controlled and only slow rates of convergence hold in this relaxed setting. However, the bounds are sufficiently precise to imply fast rates of convergence for the excess risk as in Theorems 2 if a slightly stronger Bernstein condition holds.

Given that data may not have the same distribution as (X,Y), the following relaxed version of Assumption 5 is introduced.

Assumption 7. *$(X_i,Y_i)_{i\in\mathcal{I}}$ are independent and for all $i\in\mathcal{I}$, (X_i,Y_i) has distribution P_i, X_i has distribution μ_i. For any $i\in\mathcal{I}$, $F\subset L_2(\mu_i)$ and $P_i\mathcal{L}_f=P\mathcal{L}_f$ for all $f\in F$.*

When Assumption 6 does not necessary hold, the localization argument has to be modified. Instead of the L_2-norm, the excess risk $f \in F \to P\mathcal{L}_f$ is used to define neighborhoods around f^*. The associated complexity is then defined for all $\gamma > 0$ and $K \in \{1, \cdots, N\}$ by

$$\bar{r}_2(\gamma) = \inf\left\{ r > 0 : \max\left(\frac{E(r)}{\gamma}, \sqrt{1536} V_K(r) \right) \leq r^2 \right\} \tag{13}$$

where

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

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

The main difference with $r_2(\theta)$ in Definition 2 or $\tilde{r}_2(\gamma)$ in (6) is the extra variance term $V_K(r)$. Under the Bernstein condition, this term is negligible in front of the "expectation term" $E(r)$ see [8]. In the general setting considered here, the variance term is handled in the complexity parameter.

Theorem 4. *Grant Assumptions 1, 2, 7 and assume that* $|\mathcal{O}| \leq 3N/7$. *Let* $\gamma = 1/(768L)$ *and* $K \in \left[7|\mathcal{O}|/3, N\right]$. *The minmax MOM estimator* $\hat{f}$ *defined in* (5) *satisfies, with probability at least* $1 - \exp(-K/2016)$, $P\mathcal{L}_{\hat{f}} \leq \bar{r}_2^2(\gamma)$.

Recall that Assumptions 1 and 2 are only meaning that the loss function is convex and Lipschitz and that the class F is convex. The stochastic assumption 7 says that inliers are independent and define the same excess risk as (X, Y) over F. In particular, Theorem 4 holds without assumptions on the outliers $(X_i, Y_i)_{i \in \mathcal{O}}$ or the outputs $(Y_i)_{i \in \mathcal{I}}$ of the inliers. Moreover, the risk bound holds with exponentially large probability without assuming sub Gaussian design, a small ball hypothesis or a Bernstein condition. This generality can be achieved by combining MOM estimators with convex-Lipschitz loss functions.

The following result discuss relationships between Theorems 2 and 4. Introduce the following modification of the Bernstein condition.

Assumption 8. *Let* $\gamma = 1/(768L)$. *There exists a constant* $A > 0$ *such that for all* $f \in F$ *if* $P\mathcal{L}_f \leqslant C'_{K,r}$ *then* $\|f - f^*\|_{L_2}^2 \leqslant AP\mathcal{L}_f$ *where, for* $\tilde{r}_2(\gamma)$ *defined in* (6),

$$C'_{K,r} = \max\left(\frac{\tilde{r}_2^2(\gamma/A)}{A}, \frac{1536 A L^2 K}{N} \right) . \tag{14}$$

Assumption 8 is slightly stronger than Assumption 4 since the neighborhood around f^* where the condition holds is (slightly) larger. If Assumption 8 holds then Theorem 5 implies the same statistical bounds for (5) as Theorem 2 up to constants, as shown by the following result.

Theorem 5. *Grant Assumptions 1, 2, 7 and assume that* $|\mathcal{O}| \leq 3N/7$. *Assume that the local Bernstein condition Assumption 8 holds. Let* $\gamma = 1/(768L)$ *and* $K \in \left[7|\mathcal{O}|/3, N\right]$. *The minmax MOM estimator* $\hat{f}$ *defined in* (5) *satisfies, with probability at least* $1 - \exp(-K/2016)$,

$$\left\| \hat{f} - f^* \right\|_{L_2}^2 \leq \max\left(\tilde{r}_2^2(\gamma/A), \frac{1536 L^2 A^2 K}{N} \right) \text{ and } P\mathcal{L}_{\hat{f}} \leq \max\left(\frac{\tilde{r}_2^2(\gamma/A)}{A}, \frac{1536 L^2 A K}{N} \right).$$

Proof. First, $V_K(r) \leq L V'_K(r)$ for all $r > 0$ where $V'_K(r) = \sqrt{K/N} \max_{i \in \mathcal{I}} \sup_{f \in F : P\mathcal{L}_f \leq r^2} \|f - f^*\|_{L_2(\mu_i)}$. Moreover, $r \to E(r)/r^2$ and $r \to V'_K(r)/r^2$ are non-increasing, therefore by Assumption 8 and the definition of $\tilde{r}_2(\gamma)$, $V'_K(r)$,

$$\frac{1}{\gamma} E\left(\frac{\tilde{r}_2(\gamma/A)}{\sqrt{A}} \right) \leq \frac{\tilde{r}_2^2(\gamma/A)}{A} \quad \text{and} \quad \sqrt{1536} V'_K \left(\sqrt{1536} L \sqrt{\frac{AK}{N}} \right) \leq \frac{1536 A^2 L K}{N} .$$

Hence, $\bar{r}_2^2(\gamma) \leq \max(\tilde{r}_2^2(\gamma/A)/A, 1536 L^2 A (K/N))$. ∎

5 Bernstein's assumption

This section shows that the local Bernstein condition holds for various loss functions and design X. All along the section, the oracle f^* and the Bayes rules which minimizes the risk $f \to R(f)$ over all measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ are assumed to be equal. In that case, the Bernstein's condition [8] coincides with the margin assumption [39, 32]. The class F and design X are also assumed to satisfy a "local L_4/L_2-assumption". Let $r^2 \in \{r_2^2(\theta), C_{K,r}\}$, to verify Assumption 4, choose $r^2 = r_2^2(\theta)$ while for Assumption 6, choose $r^2 = C_{K,r}$.

Assumption 9. *There exists $C' > 0$ such that for all $f \in F$ satisfying $\|f - f^*\|_{L_2} \leq r$, $\|f - f^*\|_{L_4} \leq C'\|f - f^*\|_{L_2}$*

Assumption 9 is a local version of a "L_4/L_2" norm equivalence assumption over $F - \{f^*\}$ which has been used for the study of MOM estimators (see [29]) since it implies the small ball condition. Examples of distributions satisfying the global (i.e. over all F) version of Assumption 9 can be found in [33].

Assumption 9 is local, it holds only in a L_2 neighborhood of f^* and not on the entire set $F - \{f^*\}$. There are situations where the constant C' depends on the dimension p of the model. In that case, the results in [29, 24] provide sub-optimal statistical upper bounds. For instance, if X is uniformly distributed on $[0,1]$ and $F = \{\sum_{j=1}^p \alpha_j I_{A_j} : (\alpha_j)_{j=1}^p \in \mathbb{R}^p\}$ where I_{A_j} is the indicator of $A_j = [(j-1)/p, j/p]$ then for all $f \in F$, $\|f - f^*\|_{L_4} \leq \sqrt{p}\|f - f^*\|_{L_2}$ so $C' = \sqrt{p}$. As shown in the following subsections, the rates given in Theorem 2 or Theorem 3 are not deteriorated in this example.

5.1 Quantile loss

The proof is based on [15, Lemma 2.2]. Recall that $\ell_f(x,y) = (y - f(x))(\tau - I\{y - f(x) \leq 0\})$ and that the Bayes estimator is the quantile $q_\tau^{Y|X=x}$ of order τ of the conditional distribution of Y given $X = x$.

Assumption 10. *Let C' be the constant defined in Assumption 9. There exists $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq 2(C')^2 r$ we have $f_{Y|X=x}(z) \geq \alpha$, where $f_{Y|X=x}$ is the conditional density function of Y given $X = x$.*

Theorem 6. *Grant Assumptions 9 and 10. Assume that the oracle f^* is the Bayes estimator i.e. $f^*(x) = q_\tau^{Y|X=x}$ for all $x \in \mathcal{X}$. Then, the local Bernstein condition holds: for all $f \in F$ such that $\|f - f^*\|_{L_2} \leqslant r$, $\|f - f^*\|_{L_2}^2 \leqslant \frac{4}{\alpha} P\mathcal{L}_f$.*

The proof is postponed to Section C.1.

Consider the example from Section 3.2.3, assume that $K \lesssim \mathrm{Rank}(\Sigma)$ and recall that $r^2 = C_{K,r} = \tilde{r}_2^2(\gamma) \leq \mathrm{Rank}(\Sigma)/(2\gamma^2 N)$. If $C' = \sqrt{p}$, Assumption 10 holds as long as $N \gtrsim p^3$ and there exists $\alpha > 0$ such that, for all $x \in \mathcal{X}$ and for all z in $\mathbb{R}$ such that $|z - f^*(x)| \lesssim 1$, $f_{Y|X=x}(z) \geq \alpha$. In this situation, the rates given in Theorems 2 and 3 are still $\mathrm{Rank}(\Sigma)/N$. In particular, they are not deteriorated if $C' = \sqrt{p}$ while those given in [29, 24] are of order $p\mathrm{Rank}(\Sigma)/N$. This gives a partial answer, in our setting, to the issue raised in [37] regarding results based on the small ball method.

5.2 Huber Loss

Consider the Huber loss function defined for all $f \in F$ and $x \in \mathcal{X}, y \in \mathbb{R}$ by $\ell_f(x,y) = \rho_H(y - f(x))$ where $\rho_H(t) = t^2/2$ if $|t| \leq \delta$ and $\rho_H(t) = \delta|t| - \frac{\delta^2}{2}$ otherwise. Introduce the following assumption.

Assumption 11. *Let C' be the constant defined in Assumption 9. There exists $\alpha > 0$ such that for all $x \in \mathcal{X}$ and all z in $\mathbb{R}$ such that $|z - f^*(x)| \leq 2(C')^2 r$, $F_{Y|X=x}(z + \delta) - F_{Y|X=x}(z - \delta) \geqslant \alpha$, where $F_{Y|X=x}$ is the conditional cumulative function of Y given $X = x$.*

Under this assumption, the local Bernstein condition can be checked as shown by the following result.

Theorem 7. *Grant Assumptions 9 and 11. Assume that the oracle f^* is the Bayes rules, meaning that for all $x \in \mathcal{X}, f^*(x) \in \operatorname{argmin}_{u\in\mathbb{R}} \mathbb{E}[\rho_H(Y-u)|X=x]$. Then, for all $f \in F$, if $\|f-f^*\|_{L_2} \leqslant r$ then $\|f-f^*\|_{L_2}^2 \leqslant (4/\alpha)P\mathcal{L}_f$.*

The proof is postponed to Section C.2. The interested reader can check that the remark following Theorem 6 applies to this example too.

5.3 Hinge loss

Let η denote the regression function $\eta : x \in \mathcal{X} \to \mathbb{E}[Y|X=x]$ for all $x \in \mathcal{X}$ and recall that the Bayes rule is defined by $f^{**} : x \in \mathcal{X} \to \operatorname{sgn}(\eta(x))$. The function f^{**} minimizes $f \mapsto R(f)$ over all measurable functions from $\mathcal{X}$ to $\mathcal{Y} = \{-1, 1\}$ [43]. Assume that f^* is the Bayes rules, i.e. that $f^{**} \in F$. Then, Assumption 4 holds if the following margin assumption is satisfied, see [25] and [39, Proposition 1]:

$$\text{there exists } \tau > 0 \text{ such that } \eta(X) \geq \tau \ \text{ a.s } . \tag{15}$$

The following result was proved in [25].

Theorem 8. *Let F be a class of functions from $\mathcal{X}$ to $[-1,1]$. If $f^*(x) = sgn(\eta(x))$ for all $x \in \mathcal{X}$ and (15) is satisfied, the Bernstein condition holds for the hinge loss with $A = 1/(2\tau)$.*

5.4 Logistic classification

Consider the logistic loss and recall that the Bayes rule, called the log-odds ratio, is defined by $f^{**} : x \to \log[\eta(x)/(1-\eta(x))]$ for all $x \in \mathcal{X}$, where $\eta(x) = \mathbb{P}(Y=1|X=x)$ for all $x \in \mathcal{X}$. Consider the following assumption on η.

Assumption 12. *There exists $c_0 > 0$ such that*

$$\mathbb{P}\left(\frac{1}{1+\exp(c_0)} \leq \eta(X) \leq \frac{1}{1+\exp(-c_0)}\right) \geq 1 - e^{-c_0} \ .$$

Assumption 12 excludes trivial cases where deterministic predictors equal to 1 or -1 are optimal.

Theorem 9. *Grant Assumptions 9 and 12 and assume that the Bayes rules is in F. Then, there exists $A > 0$ such that*

$$\forall f \in F : \|f-f^*\|_{L_2} \leqslant r, \qquad P\mathcal{L}_f \geqslant A\|f-f^*\|_{L_2}^2 \ .$$

Theorem 9 is proved in Section C.3. The explicit form of the constant A is presented in the proof. A close inspection of the proof reveals that Assumption 12 is only used to bound $f^*(X)$ with large probability. If the oracle is bounded, the same result holds without assuming that f^* is the Bayes rules.

6 Simulation study

This section provides a short simulation study that illustrates our theoretical findings. The section starts with an example where all assumptions of Theorem 2 are simultaneously satisfied.

6.1 Application: robust classification

Let $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \{-1, 1\}$, let ℓ denote the logistic loss recalled in Section 1. Recall that Assumption 1 holds with $L = 1$. Let $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^p\}$ be a class of linear functions, so Assumption 2 holds. Assume that $X = (\xi_1, \cdots, \xi_p)$ in $\mathbb{R}^p$, where $(\xi_j)_{j=1}^p$ are centered, independent and identically distributed and that (X, Y) satisfies a logistic model

$$\log\left(\frac{\mathbf{P}(Y=1|X)}{\mathbf{P}(Y=-1|X)}\right) = \langle X, t^* \rangle + \epsilon \ . \tag{16}$$

By Theorem 9, Assumptions 4 and 6 hold if

$$\forall f \in F \text{ s.t } \|f - f^*\|_{L_2} \le r, \; \|f - f^*\|_{L_4} \le C' \|f - f^*\|_{L_2} \tag{17}$$

for $r^2 \in \{r_2^2(\theta), C_{K,r}\}$. Now, since $\xi_1, \cdots, \xi_p$ are centered i.i.d. in $\mathbb{R}^p$, basic algebraic computations show that (17) holds if

$$\mathbb{E}(|\xi_1|^4)^{1/4} \le A\mathbb{E}(|\xi_1|^2)^{1/2}.$$

This assumption is satisfied, for example, by student distributions $\mathcal{T}(5)$ that do not have a fifth moment. The results allow non-symmetric noise such as $\epsilon \sim \mathcal{LN}(0,1)$ where $\mathcal{LN}(\mu,\sigma)$ denotes the log-normal distribution with mean $\exp(\mu + \sigma/2)$ and variance $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$.

Consider the model (16) with X and ϵ as above. Let $(X_i, Y_i)_{i\in\mathcal{I}}$ independent and distributed as (X, Y). Let also $(X_i, Y_i)_{i\in\mathcal{O}}$ where $\mathcal{I} \cup \mathcal{O} = \{1, \cdots, N\}$. For any $K \ge 7|\mathcal{O}|/3$, the minmax MOM estimator is solution of the following problem:

$$\hat{t}_K^{\text{MOM}} \in \arg\min_{t\in\mathbb{R}^p} \sup_{\tilde{t}\in\mathbb{R}^p} \text{MOM}_K(\ell_t - \ell_{\tilde{t}}) \;. \tag{18}$$

As all assumptions in Theorem 2 are satisfied, if the number of outliers is such that $|\mathcal{O}| \le p$ and the number of blocks is $K = p$ then there exist absolute positive constants a_1, a_2, a_3 and a_4 depending only on A such that, with probability larger than $1 - a_1 \exp(-a_2 p)$,

$$\|\hat{t}_p^{\text{MOM}} - t^*\|_2^2 \le a_3 \frac{p}{N} \text{ and } P\mathcal{L}_{\hat{t}} \le a_4 \frac{p}{N} \;. \tag{19}$$

6.2 Simulations

This section presents algorithms derived from the minmax MOM construction (18). These algorithms are inspired from [24]. Consider the setup of Section 6.1 where Theorem 2 applies: $X = (\xi_1, \cdots, \xi_p)$, where $(\xi_j)_{j=1}^p$ are independent and identically distributed, with $\xi_1 \sim \mathcal{T}(5)$, $\epsilon \sim \mathcal{LN}(0,1)$ and

$$\log\left(\frac{\mathbf{P}(Y = 1|X)}{\mathbf{P}(Y = -1|X)}\right) = \langle X, t^* \rangle + \epsilon \;.$$

Following [24], a gradient ascent-descent step is performed on the empirical incremental risk $(t, \tilde{t}) \to P_{B_k}(\ell_t - \ell_{\tilde{t}})$ constructed on the block B_k of data realizing the median of the empirical incremental risk. Initial points $t_0 \in \mathbb{R}^p$ and $\tilde{t}_0 \in \mathbb{R}^p$ are taken at random. In logistic regression, the step sizes η and $\tilde{\eta}$ are usually chosen equal to $\|\mathbb{X}\mathbb{X}^\top\|_{\text{op}}/4N$, where $\mathbb{X}$ is the $N \times p$ matrix with row vectors equal to $X_1^\top, \cdots, X_N^\top$ and $\|\cdot\|_{\text{op}}$ denotes the operator norm. In a corrupted environment, this choice might lead to disastrous performance. This is why η and $\tilde{\eta}$ are computed at each iteration using only data in the median block: let B_k denote the median block at the current step, then one chooses $\eta = \tilde{\eta} = \|\mathbb{X}_{(k)}\mathbb{X}_{(k)}^\top\|_{\text{op}}/4|B_k|$ where $\mathbb{X}_{(k)}$ is the $|B_k| \times p$ matrix with rows given by X_i^T for $i \in B_k$. In practice, K is chosen by robust cross-validation choice as in [24].

In a first approach and according to our theoretical results, the blocks are chosen at the beginning of the algorithm. As illustrated in Figure 2, this first strategy has some limitations. To understand the problem, for all $k = 1, \ldots, K$, let C_k denote the following set

$$C_k = \{t \in \mathbb{R}^p : P_{B_k}\ell_t = \text{Median}\{P_{B_1}\ell_t, \ldots, P_{B_K}\ell_t\}\} \;.$$

If the minimum of $t \to P_{B_k}\ell_t$ lies in C_k, the algorithm typically converges to this minimum if one iteration enters C_k. As a consequence, when the minmax MOM estimator (18) lies in another cell, the algorithm does not converge to this estimator.

To bypass this issue, the partition is changed at every ascent/descent steps of the algorithm, it is chosen uniformly at random among all equipartition of the dataset. This alternative algorithm is described in Algorithm 1. In practice, changing the partition seems to widely accelerate the convergence (see Figure 2).

Input: The number of block K, initial points t_0 and $\tilde{t}_0$ in $\mathbb{R}^p$ and the stopping criterion $\epsilon > 0$
Output: A solution of the minimax problem (18)
1 **while** $\|t_i - \tilde{t}_i\|_2 \geq \epsilon$ **do**
2 Split the data into K disjoint blocks $(B_k)_{k\in\{1,\cdots,K\}}$ of equal sizes chosen at random: $B_1 \cup \cdots \cup B_K = \{1, \cdots, N\}$.
3 Find $k \in [K]$ such that $\mathrm{MOM}_K \ell_{t_i} = P_{B_k}\ell_{t_i}$.
4 Compute $\eta = \tilde{\eta} = \|\mathbb{X}_{(k)}^T \mathbb{X}_{(k)}\|_{op}/4N$.
5 Update $t_{i+1} = t_i - \frac{1}{\eta}\nabla_t(P_{B_k}\ell_t)_{|t=t_i}$ and $\tilde{t}_{i+1} = \tilde{t}_i - \frac{1}{\tilde{\eta}}\nabla_{\tilde{t}}(P_{B_k}\ell_{\tilde{t}})_{|\tilde{t}=\tilde{t}_i}$.
6 **end**

Algorithm 1: Descent-ascent gradient method with blocks of data chosen at random at every steps.

Simulation results are gathered in Figure 2. In these simultations, there is no outlier, $N = 1000$ and $p = 100$ with $(X_i, Y_i)_{i=1}^{1000}$ i.i.d with the same distribution as (X, Y). Minmax MOM estimators (18) are compared with the Logistic Regression algorithm from the scikit-learn library of [36].

The upper pictures compare performance of MOM ascent/descent algorithms with fixed and changing blocks. These pictures give an example where the fixed block algorithm is stuck into local minima and another one where it does not converge. In both cases, the changing blocks version converges to t^*.

Running times of logistic regression (LR) and its MOM version (MOM LR) are compared in the lower picture of Figure 2 in a dataset free from outliers. LR and MOM LR are coded with the same algorithm in this example, meaning that MOM gradient descent-ascent and simple gradient descent are performed with the same descent algorithm. As illustrated in Figure 2, running each step of the gradient descent on one block only and not on the whole dataset accelerates the running time. The larger the dataset, the bigger the benefit is expected.

The resistance to outliers of logistic regression and its minmax MOM alternative are depicted in Figure 1 in the introduction. We added an increasing number of outliers to the dataset. Outliers $\{(X_i, Y_i), i \in \mathcal{O}\}$ in this simulation are such that $X_i \sim \mathcal{LN}(0, 5)$ and $Y_i = -sign(\langle X_i, t\rangle + \epsilon_i)$, with $\epsilon_i \sim \epsilon$ as above. Figure 1 shows that logistic classification is mislead by a single outlier while MOM version maintains reasonable performance with up to 50 outliers (i.e 5% of the database is corrupted).

A byproduct of Algorithm 1 is an outlier detection algorithm. Each data receives a score equal to the number of times it is selected in a median block in the random choice of block version of the algorithm. The first iterations may be misleading: before convergence, the empirical loss at the current point may not reveal the centrality of the data because the current point may be far from t^*. Simulations are run with $N = 100$, $p = 10$ and 5000 iterations and therefore only the score obtained by each data in the last 4000 iterations are displayed. 3 outliers $(X_i, Y_i)_{i\in\{1,2,3\}}$ with $X_i = (10)_{j=1}^p$ and $Y_i = -sign(\langle X_i, t\rangle)$ have been introduced at number 42, 62 and 66. Figure 3 shows that these are not selected once.

7 Conclusion

The paper introduces a new homogeneity argument for learning problems with convex and Lipschitz losses. This argument allows to obtain estimation rates and oracle inequalities for ERM and minmax MOM estimators improving existing results. The ERM requires subgaussian hypotheses on the design and a local Bernstein condition (see Theorem 1), both assumptions can be removed for minmax MOM estimators (see Theorem 5). The local Bernstein conditions provided in this article can be verified in several learning problems. In particular, it allows to derive optimal risk bounds in examples where analysis based on the small ball hypothesis fail. Minmax MOM estimators applied to convex and Lipschitz losses are efficient without assumptions on the outputs Y, under minimal L_2 assumptions on the design and the results are robust to the presence of few outliers in the dataset. A modification of these estimators can be implemented efficiently and confirm all these conclusions.

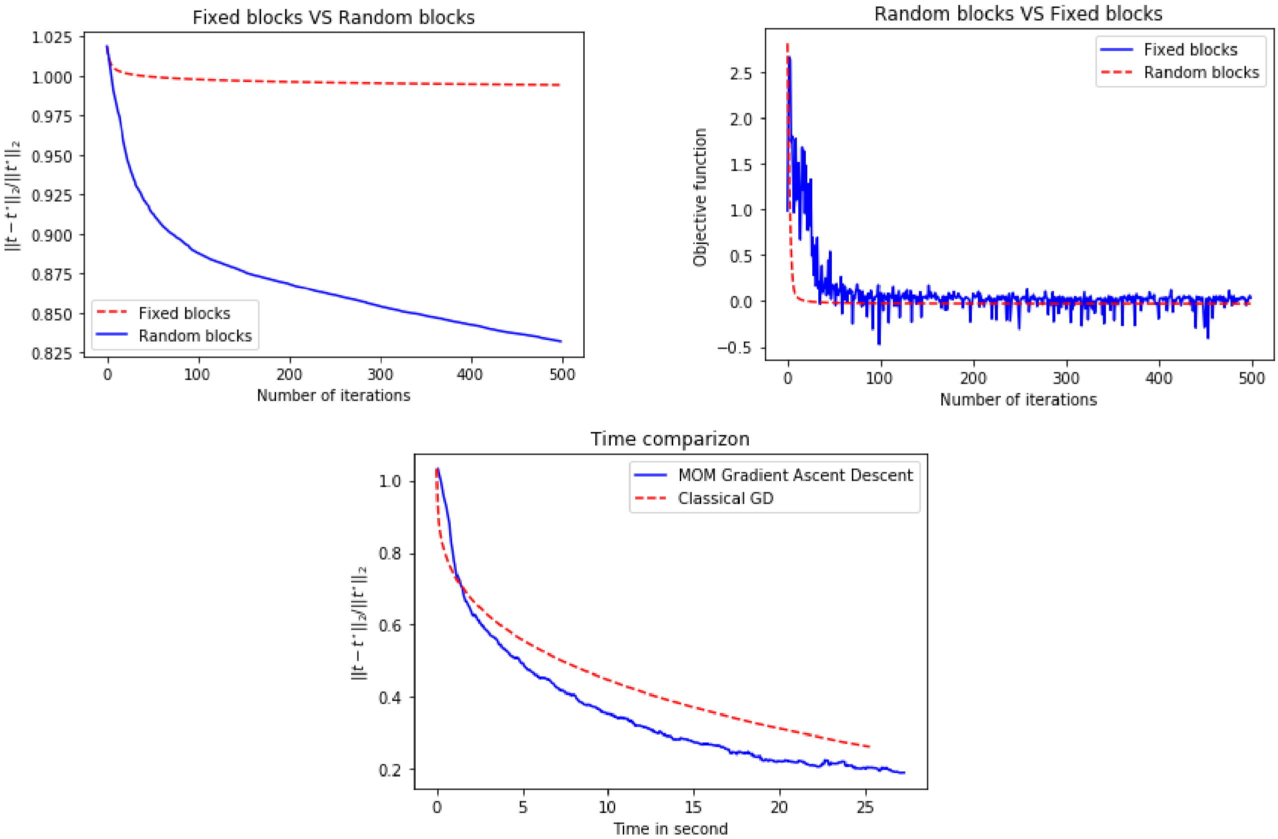

Figure 2: Top left and right: Comparizon of the algorithm with fixed and changing blocks. Bottom: Comparizon of running time between classical gradient descent and algorithm 1. In all simulation $N = 1000$, $p = 100$ and there is no outliers.

A Proof of Theorems 1, 2, 3 and 4

A.1 Proof of Theorem 1

The proof is splitted in two parts. First, we identify an event where the statistical behavior of the regularized estimator $\hat{f}^{ERM}$ can be controlled. Then, we prove that this event holds with probability at least (3). Introduce the following event:

$$\Omega := \left\{ \forall f \in F \cap (f^* + r_2(\theta) B_{L_2}), \quad \left| (P - P_N)\mathcal{L}_f \right| \leq \theta r_2(\theta)^2 \right\}$$

where θ is a parameter appearing in the definition of r_2 in Definition 3.

Proposition 1. *On the event Ω, one has*

$$\|\hat{f}^{ERM} - f^*\|_{L_2} \leq r_2(\theta) \text{ and } P\mathcal{L}_{\hat{f}^{ERM}} \leq \theta r_2^2(\theta).$$

Proof. By construction, $\hat{f}^{ERM}$ satisfies $P_N \mathcal{L}_{\hat{f}^{ERM}} \leq 0$. Therefore, it is sufficient to show that, on Ω, if $\|f - f^*\|_{L_2} > r_2(\theta)$, then $P_N \mathcal{L}_f > 0$. Let $f \in F$ be such that $\|f - f^*\|_{L_2} > r_2(\theta)$. By convexity of F, there exists $f_0 \in F \cap (f^* + r_2(\theta) S_{L_2})$ and $\alpha > 1$ such that

$$f = f^* + \alpha(f_0 - f^*) \ . \tag{20}$$

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

$$\psi_i(u) = \bar{\ell}(u + f^*(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i). \tag{21}$$

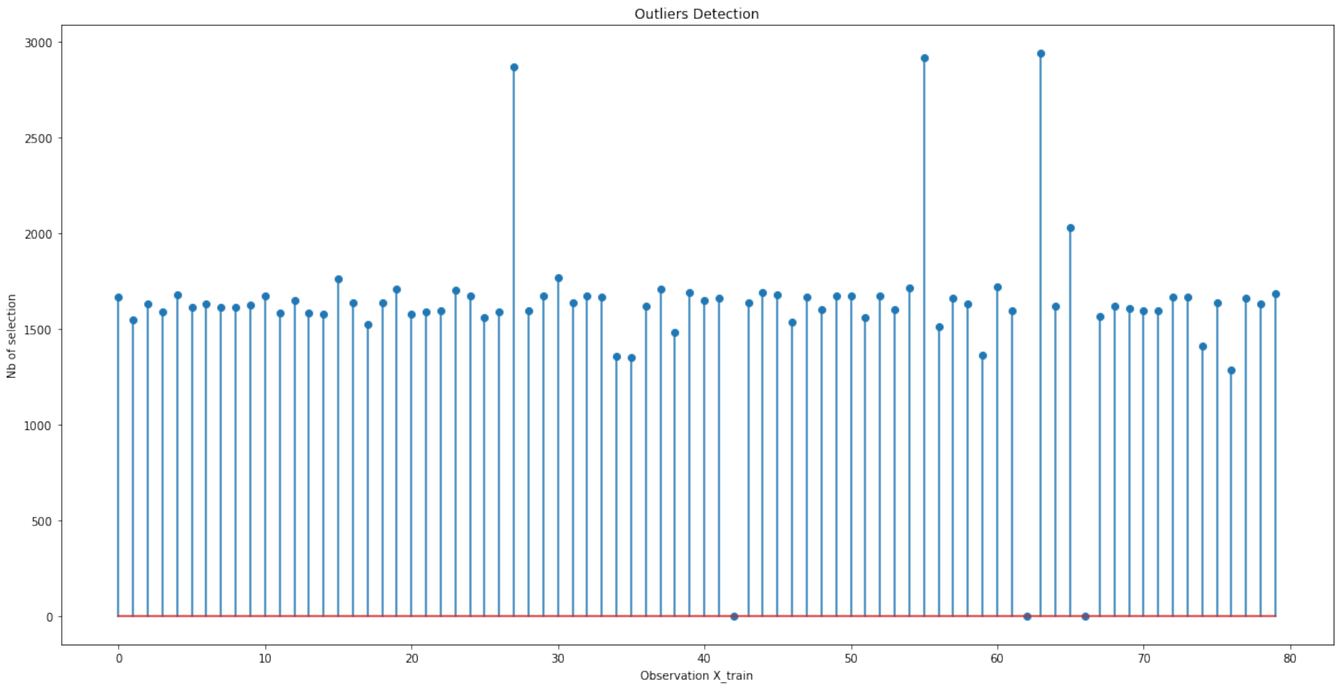

Figure 3: Outliers Detection Procedure for $N = 100$, $p = 10$ and outliers are $i = 42, 62, 66$

The functions ψ_i are such that $\psi_i(0) = 0$, they are convex because $\bar{\ell}$ is, in particular $\alpha\psi_i(u) \le \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha \ge 1$ and $\psi_i(f(X_i) - f^*(X_i)) = \bar{\ell}(f(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i)$ so that the following holds:

$$
\begin{aligned}
P_N \mathcal{L}_f &= \frac{\alpha}{N} \sum_{i=1}^{N} \frac{\psi_i\big(f(X_i) - f^*(X_i)\big)}{\alpha} = \frac{\alpha}{N} \sum_{i=1}^{N} \frac{\psi_i(\alpha(f_0(X_i) - f^*(X_i)))}{\alpha} \\
&\ge \frac{\alpha}{N} \sum_{i=1}^{N} \psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_N \mathcal{L}_{f_0}. \qquad (22)
\end{aligned}
$$

Until the end of the proof, the event Ω is assumed to hold. Since $f_0 \in F \cap (f^* + r_2(\theta) B_{L_2})$, $P_N \mathcal{L}_{f_0} \ge P\mathcal{L}_{f_0} - \theta r_2^2(\theta)$. Moreover, by Assumption 4, $P\mathcal{L}_{f_0} \ge A^{-1}\|f_0 - f^*\|_{L_2}^2 = A^{-1} r_2^2(\theta)$, thus

$$
P_N \mathcal{L}_{f_0} \ge (A^{-1} - \theta) r_2^2(\theta). \qquad (23)
$$

From Eq. (22) and (23), $P_N \mathcal{L}_f > 0$ since $A^{-1} > \theta$. Therefore, $\|\hat{f}^{ERM} - f^*\|_{L_2} \le r_2^2(\theta)$. This proves the L_2-bound.

Now, as $\|\hat{f}^{ERM} - f^*\|_{L_2} \le r_2^2(\theta)$, $|(P - P_N)\mathcal{L}_{\hat{f}^{ERM}}| \le \theta r_2^2(\theta)$. Since $P_N \mathcal{L}_{\hat{f}^{ERM}} \le 0$,

$$
P\mathcal{L}_{\hat{f}^{ERM}} = P_N \mathcal{L}_{\hat{f}^{ERM}} + (P - P_N)\mathcal{L}_{\hat{f}^{ERM}} \le \theta r_2^2(\theta) \ .
$$

This show the excess risk bound. ■

Proposition 1 shows that $\hat{f}^{ERM}$ has the risk bounds given in Theorem 1 on the event Ω. To show that Ω holds with probability (3), recall the following results from [2].

Lemma 2. *[2] [Lemma 8.1] Grant Assumptions 1 and 3. Let $F' \subset F$ with finite L_2-diameter $d_{L_2}(F')$. For every $u > 0$, with probability at least $1 - 2\exp(-u^2)$,*

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

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

$$\sup_{f \in F \cap (f^* + r_2(\theta) B_{L_2})} \left|(P - P_N)\mathcal{L}_f\right| \leq \sup_{f,g \in F \cap (f^* + r_2(\theta) B_{L_2})} \left|(P - P_N)(\mathcal{L}_f - \mathcal{L}_g)\right|$$
$$\leq \frac{16L}{\sqrt{N}} \big(w((F - f^*) \cap r_2(\theta)B_{L_2}) + u d_{L_2}((F - f^*) \cap r_2(\theta)B_{L_2})\big)$$

where $d_{L_2}((F - f^*) \cap r_2(\theta)B_{L_2}) \leq r_2(\theta)$. By definition of the complexity parameter (see Eq. (3)), for $u = \theta\sqrt{N} r_2(\theta)/(64L)$, with probability at least

$$1 - 2\exp\big(-\theta^2 N r_2^2(\theta)/(16^3 L^2)\big) \ , \tag{24}$$

for every f in $F \cap (f^* + r_2(\theta)B_{L_2})$,

$$\left|(P - P_N)\mathcal{L}_f\right| \leq \theta r_2^2(\theta). \tag{25}$$

Together with Proposition 1, this concludes the proof of Theorem 1.

A.2 Proof of Theorem 2

The proof is splitted in two parts. First, we identify an event Ω_K where the statistical properties of $\hat{f}$ from Theorem 2 can be established. Next, we prove that this event holds with probability (8). Let α, θ and γ be positive numbers to be chosen later. Define

$$C_{K,r} = \max\left(\frac{4L^2 K}{\theta^2 \alpha N}, \tilde{r}_2^2(\gamma)\right)$$

where the exact form of α, θ and γ are given in Equation (34). Set the event Ω_K to be such that

$$\Omega_K = \left\{\forall f \in F \cap \left(f^* + \sqrt{C_{K,r}} B_{L_2}\right), \exists J \subset \{1, \dots, K\} : |J| > K/2 \text{ and } \forall k \in J, |(P_{B_k} - P)\mathcal{L}_f| \leq \theta C_{K,r}\right\} . \tag{26}$$

A.2.1 Deterministic argument

The goal of this section is to show that, on the event Ω_K, $\|\hat{f} - f^*\|_{L_2}^2 \leq C_{K,r}$ and $P\mathcal{L}_{\hat{f}} \leq 2\theta C_{K,r}$.

Lemma 3. *If there exists $\eta > 0$ such that*

$$\sup_{f \in F \backslash (f^* + \sqrt{C_{K,r}} B_{L_2})} MOM_K\big(\ell_{f^*} - \ell_f\big) < -\eta \quad \textit{and} \quad \sup_{f \in F \cap (f^* + \sqrt{C_{K,r}} B_{L_2})} MOM_K\big(\ell_{f^*} - \ell_f\big) \leq \eta \ , \tag{27}$$

then $\|\hat{f} - f^*\|_{L_2}^2 \leq C_{K,r}$.

Proof. Assume that (27) holds, then

$$\inf_{f \in F \backslash (f^* + \sqrt{C_{K,r}} B_{L_2})} \text{MOM}_K[\ell_f - \ell_{f^*}] > \eta \ . \tag{28}$$

Moreover, if $T_K(f) = \sup_{g \in F} \text{MOM}_K[\ell_f - \ell_g]$ for all $f \in F$, then

$$T_K(f^*) = \sup_{f \in F \cap (f^* + \sqrt{C_{K,r}} B_{L_2})} \text{MOM}_K[\ell_{f^*} - \ell_f] \vee \sup_{f \in F \backslash (f^* + \sqrt{C_{K,r}} B_{L_2})} \text{MOM}_K[\ell_{f^*} - \ell_f] \leqslant \eta \ . \tag{29}$$

By definition of $\hat{f}$ and (29), $T_K(\hat{f}) \leqslant T_K(f^*) \leqslant \eta$. Moreover, by (28), any $f \in F \backslash \big(f^* + \sqrt{C_{K,r}} B_{L_2}\big)$ satisfies $T_K(f) \geqslant \text{MOM}_K[\ell_f - \ell_{f^*}] > \eta$. Therefore $\hat{f} \in F \cap (f^* + \sqrt{C_{K,r}} B_{L_2})$. ■

Lemma 4. *Grant Assumption 6 and assume that* $\theta - A^{-1} < -\theta$. *On the event* Ω_K, (27) *holds with* $\eta = \theta C_{K,r}$.

Proof. Let $f \in F$ be such that $\|f - f^*\|_{L_2} > C_{K,r}$. By convexity of F, there exists $f_0 \in F \cap \big(f^* + \sqrt{C_{K,r}} S_{L_2}\big)$ and $\alpha > 1$ such that $f = f^* + \alpha(f_0 - f^*)$. For all $i \in \{1, \dots, N\}$, let $\psi_i : \mathbb{R} \to \mathbb{R}$ be defined for all $u \in \mathbb{R}$ by

$$\psi_i(u) = \bar{\ell}(u + f^*(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i). \tag{30}$$

The functions ψ_i are convex because $\bar{\ell}$ is and such that $\psi_i(0) = 0$, so $\alpha\psi_i(u) \le \psi_i(\alpha u)$ for all $u \in \mathbb{R}$ and $\alpha \ge 1$. As $\psi_i(f(X_i) - f^*(X_i)) = \bar{\ell}(f(X_i), Y_i) - \bar{\ell}(f^*(X_i), Y_i)$, for any block B_k,

$$\begin{aligned} P_{B_K}\mathcal{L}_f &= \frac{\alpha}{|B_k|}\sum_{i \in B_k} \frac{\psi_i\big(f(X_i) - f^*(X_i)\big)}{\alpha} = \frac{\alpha}{|B_k|}\sum_{i \in B_k} \frac{\psi_i(\alpha(f_0(X_i) - f^*(X_i)))}{\alpha} \\ &\ge \frac{\alpha}{|B_k|}\sum_{i \in B_k} \psi_i((f_0(X_i) - f^*(X_i))) = \alpha P_{B_k}\mathcal{L}_{f_0}. \end{aligned} \tag{31}$$

As $f_0 \in F \cap (f^* + \sqrt{C_{K,r}} B_{L_2})$, on Ω_K, there are strictly more than $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_{f_0} \ge P\mathcal{L}_{f_0} - \theta C_{K,r}$. Moreover, from Assumption 6, $P\mathcal{L}_{f_0} \ge A^{-1}\|f_0 - f^*\|_{L_2}^2 = A^{-1} C_{K,r}$. Therefore, on strictly more than $K/2$ blocks B_k,

$$P_{B_k}\mathcal{L}_{f_0} \ge (A^{-1} - \theta) C_{K,r}. \tag{32}$$

From Eq. (31) and (32), there are strictly more than $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_f \ge (A^{-1} - \theta)C_{K,r}$. Therefore, on Ω_K, as $(\theta - A^{-1}) < -\theta$,

$$\sup_{f \in F \setminus \left(f^* + \sqrt{C_{K,r}} B_{L_2}\right)} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) < (\theta - A^{-1})C_{K,r} < -\theta C_{K,r} \ .$$

In addition, on the event Ω_K, for all $f \in F \cap (f^* + \sqrt{C_{K,r}} B_{L_2})$, there are strictly more than $K/2$ blocks B_k where $|(P_{B_k} - P)\mathcal{L}_f| \le \theta C_{K,r}$. Therefore

$$\mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \le \theta C_{K,r} - P\mathcal{L}_f \le \theta C_{K,r}.$$

■

Lemma 5. *Grant Assumption 6 and assume that* $\theta - A^{-1} < -\theta$. *On the event* Ω_K, $P\mathcal{L}_{\hat{f}} \le 2\theta C_{K,r}$.

Proof. Assume that Ω_K holds. From Lemmas 3 and 4, $\|\hat{f} - f^*\|_{L_2} \le \sqrt{C_{K,r}}$. Therefore, on strictly more than $K/2$ blocks B_k, $P\mathcal{L}_{\hat{f}} \le P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r}$. In addition, by definition of $\hat{f}$ and (29) (for $\eta = \theta C_{K,r}$),

$$MOM_K(\ell_{\hat{f}} - \ell_{f^*}) \le \sup_{f \in F} MOM_K(\ell_{f^*} - \ell_f) \le \theta C_{K,r}.$$

As a consequence, there exist at least $K/2$ blocks B_k where $P_{B_k}\mathcal{L}_{\hat{f}} \le \theta C_{K,r}$. Therefore, there exists at least one block B_k where both $P\mathcal{L}_{\hat{f}} \le P_{B_k}\mathcal{L}_{\hat{f}} + \theta C_{K,r}$ and $P_{B_k}\mathcal{L}_{\hat{f}} \le \theta C_{K,r}$. Hence $P\mathcal{L}_{\hat{f}} \le 2\theta C_{K,r}$. ■

A.2.2 Stochastic argument

This section shows that Ω_K holds with probability at least (8).

Proposition 2. *Grant Assumptions 1, 2, 5 and 6 and assume that* $(1-\beta)K \ge |\mathcal{O}|$. *Let* $x > 0$ *and assume that* $\beta(1 - \alpha - x - 8\gamma L/\theta) > 1/2$. *Then* Ω_K *holds with probability larger than* $1 - \exp(-x^2\beta K/2)$.

Proof. Let $\mathcal{F} = F \cap (f^* + \sqrt{C_{K,r}} B_{L_2})$ and set $\phi : t \in \mathbb{R} \to I\{t \geq 2\} + (t-1)I\{1 \leq t \leq 2\}$ so, for all $t \in \mathbb{R}$, $I\{t \geq 2\} \leq \phi(t) \leq I\{t \geq 1\}$. Let $W_k = ((X_i, Y_i))_{i \in B_k}$, $G_f(W_k) = (P_{B_k} - P)\mathcal{L}_f$. Let

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

Let $\mathcal{K}$ denote the set of indices of blocks which have not been corrupted by outliers, $\mathcal{K} = \{k \in \{1, \cdots, K\} : B_k \subset \mathcal{I}\}$ and let $f \in \mathcal{F}$. Basic algebraic manipulations show that

$$z(f) \geq |\mathcal{K}| - \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \Big(\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \Big) - \sum_{k \in \mathcal{K}} \mathbb{E}\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \ .$$

By Assumptions 1 and 5, using that $C_{K,r}^2 \geq \|f - f^*\|_{L_2}^2 \left[(4L^2 K)/(\theta^2 \alpha N)\right]$,

$$\begin{aligned}
\mathbb{E}\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) &\leq \mathbb{P}\Big(|G_f(W_k)| \geq \frac{\theta C_{K,r}}{2} \Big) \leq \frac{4}{\theta^2 C_{K,r}^2} \mathbb{E} G_f(W_k)^2 = \frac{4}{\theta^2 C_{K,r}^2} \mathbb{V}ar(P_{B_k} \mathcal{L}_f) \\
&\leq \frac{4K^2}{\theta^2 C_{K,r}^2 N^2} \sum_{i \in B_k} \mathbb{E}[\mathcal{L}_f^2(X_i, Y_i)] \leq \frac{4L^2 K}{\theta^2 C_{K,r}^2 N} \|f - f^*\|_{L_2}^2 \leq \alpha \ .
\end{aligned}$$

Therefore,

$$z(f) \geq |\mathcal{K}|(1-\alpha) - \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \Big(\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \Big) \ . \tag{33}$$

Using Mc Diarmid's inequality [11, Theorem 6.2], for all $x > 0$, with probability larger than $1 - \exp(-x^2|\mathcal{K}|/2)$,

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \Big(\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \Big) \\
&\qquad \leq x|\mathcal{K}| + \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \Big(\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1}\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \Big) \ .
\end{aligned}$$

Let $\epsilon_1, \ldots, \epsilon_K$ denote independent Rademacher variables independent of the $(X_i, Y_i), i \in \mathcal{I}$. By Giné-Zinn symmetrization argument,

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \Big(\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) - \mathbb{E}\phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \Big) \\
&\qquad\qquad \leq x|\mathcal{K}| + 2\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \epsilon_k \phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|)
\end{aligned}$$

As ϕ is 1-Lipschitz with $\phi(0) = 0$, using the contraction lemma [28, Chapter 4],

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \epsilon_k \phi(2\theta^{-1} C_{K,r}^{-1} |G_f(W_k)|) \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \epsilon_k \frac{G_f(W_k)}{\theta C_{K,r}} = 2\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \epsilon_k \frac{(P_{B_k} - P)\mathcal{L}_f}{\theta C_{K,r}}.$$

Let $(\sigma_i : i \in \cup_{k \in \mathcal{K}} B_k)$ be a family of independent Rademacher variables independent of $(\epsilon_k)_{k \in \mathcal{K}}$ and $(X_i, Y_i)_{i \in \mathcal{I}}$. It follows from the Giné-Zinn symmetrization argument that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{k \in \mathcal{K}} \epsilon_k \frac{(P_{B_k} - P)\mathcal{L}_f}{C_{K,r}} \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \frac{K}{N} \sum_{i \in \cup_{k \in \mathcal{K}} B_k} \sigma_i \frac{\mathcal{L}_f(X_i, Y_i)}{C_{K,r}}.$$

By the Lipschitz property of the loss, the contraction principle applies and

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{\mathcal{L}_f(X_i,Y_i)}{C_{K,r}}\leq L\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r}}\ .$$

To bound from above the right-hand side in the last inequality, consider two cases 1) $C_{K,r}=\tilde{r}_2^2(\gamma)$ or 2) $C_{K,r}=4L^2K/(\alpha\theta^2N)$. In the first case, by definition of the complexity parameter $\tilde{r}_2(\gamma)$ in (6),

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{C_{K,r}}=\mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\leq r_2(\gamma)}\frac{1}{\tilde{r}_2^2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|\leq\frac{\gamma|\mathcal{K}|N}{K}.$$

In the second case,

$$\begin{aligned}&\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\frac{\sigma_i(f-f^*)(X_i)}{C_{K,r}}\\&\leq\mathbb{E}\left[\sup_{\substack{f\in F:\\ \|f-f^*\|_{L_2}\leq\tilde{r}_2(\gamma)}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\frac{\sigma_i(f-f^*)(X_i)}{\tilde{r}_2^2(\gamma)}\right|\vee\sup_{\substack{f\in F:\\ \tilde{r}_2(\gamma)\leq\|f-f^*\|_{L_2}\leq\sqrt{\frac{4L^2K}{\alpha\theta^2N}}}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\frac{4L^2K}{\alpha\theta^2N}}\right|\right]\ .\end{aligned}$$

Let $f\in F$ be such that $\tilde{r}_2(\gamma)\leq\|f-f^*\|_{L_2}\leq\sqrt{[4L^2K]/[\alpha\theta^2N]}$; by convexity of F, there exists $f_0\in F$ such that $\|f_0-f^*\|_{L_2}=\tilde{r}_2(\gamma)$ and $f=f^*+\alpha(f_0-f^*)$ with $\alpha=\|f-f^*\|_{L_2}/\tilde{r}_2(\gamma)\geq 1$. Therefore,

$$\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\frac{4L^2K}{\alpha\theta^2N}}\right|\leq\frac{1}{\tilde{r}_2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\|f-f^*\|_{L_2}}\right|=\frac{1}{\tilde{r}_2^2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f_0-f^*)(X_i)\right|$$

and so

$$\sup_{\substack{f\in F:\\ \tilde{r}_2(\gamma)\leq\|f-f^*\|_{L_2}\leq\sqrt{\frac{4L^2K}{\alpha\theta^2N}}}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i\frac{(f-f^*)(X_i)}{\frac{4L^2K}{\alpha\theta^2N}}\right|\leq\frac{1}{\tilde{r}_2^2(\gamma)}\sup_{\substack{f\in F:\\ \|f-f^*\|_{L_2}=\tilde{r}_2(\gamma)}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right|.$$

By definition of $\tilde{r}_2(\gamma)$, it follows that

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

Therefore, as $|\mathcal{K}|\geq K-|\mathcal{O}|\geq\beta K$, with probability larger than $1-\exp(-x^2\beta K/2)$, for all $f\in F$ such that $\|f-f^*\|_{L_2}\leq\sqrt{C_{K,r}}$,

$$z(f)\geq|\mathcal{K}|\left(1-\alpha-x-\frac{8\gamma L}{\theta}\right)>\frac{K}{2}. \tag{34}$$

■

A.2.3 End of the proof of Theorem 2

Theorem 2 follows from Lemmas 3, 4, 5 and Proposition 2 for the choice of constant

$$\theta=1/(3A)\quad\alpha=1/24,\quad x=1/24,\quad\beta=4/7\text{ and }\gamma=1/(575AL).$$

A.3 Proof of Theorem 3

Let $K \in \left[7|\mathcal{O}|/3, N\right]$ and consider the event Ω_K defined in (26). It follows from the proof of Lemmas 3 and 4 that $T_K(f^*) \leq \theta C_{K,r}$ on Ω_K. Setting $\theta = 1/(3A)$, on $\cap_{J=K}^N \Omega_J$, $f^* \in \hat{R}_J$ for all $J = K, \dots, N$, so $\cap_{J=K}^N \hat{R}_J \neq \emptyset$. By definition of $\hat{K}$, it follows that $\hat{K} \leq K$ and by definition of $\tilde{f}$, $\tilde{f} \in \hat{R}_K$ which means that $T_K(\tilde{f}) \leq \theta C_{K,r}$. It is proved in Lemmas 3 and 4 that on Ω_K, if $f \in F$ satisfies $\|f - f^*\|_{L_2} \geq \sqrt{C_{K,r}}$ then $T_K(f) > \theta C_{K,r}$. Therefore, $\left\|\tilde{f} - f^*\right\|_{L_2} \leq \sqrt{C_{K,r}}$. On Ω_K, since $\left\|\tilde{f} - f^*\right\|_{L_2} \leq \sqrt{C_{K,r}}$, $P\mathcal{L}_{\tilde{f}} \leq 2\theta C_{K,r}$. Hence, on $\cap_{J=K}^N \Omega_J$, the conclusions of Theorem 3 hold. Finally, by Proposition 2,

$$\mathbb{P}\left[\cap_{J=K}^N \Omega_J\right] \geq 1 - \sum_{J=K}^{N} \exp(-K/2016) \geq 1 - 4\exp(-K/2016).$$

A.4 Proof of Theorem 4

The proof of Theorem 4 follows the same path as the one of Theorem 2. We only sketch the different arguments needed because of the localization by the excess loss and the lack of Bernstein condition.

Define the event Ω'_K in the same way as Ω_K in (26) where $C_{K,r}$ is replaced by $\bar{r}_2^2(\gamma)$ and the L_2 localization is replaced by the "excess loss localization":

$$\Omega'_K = \left\{ \forall f \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}, \exists J \subset \{1, \dots, K\} : |J| > K/2 \text{ and } \forall k \in J, |(P_{B_k} - P)\mathcal{L}_f| \leq (1/4)\bar{r}_2^2(\gamma) \right\} \tag{35}$$

where $(\mathcal{L}_F)_{\bar{r}_2^2(\gamma)} = \{f \in F : P\mathcal{L}_f \leq \bar{r}_2^2(\gamma)\}$. Our first goal is to show that on the event Ω'_K, $P\mathcal{L}_{\hat{f}} \leq (1/4)\bar{r}_2^2(\gamma)$. We will then handle $\mathbb{P}[\Omega'_K]$.

Lemma 6. *Grant Assumptions 1 and 2. For every $r \geq 0$, the set $(\mathcal{L}_F)_r := \{f \in F : P\mathcal{L}_f \leq r\}$ is convex and relatively closed to F in $L_1(\mu)$. Moreover, if $f \in F$ is such that $P\mathcal{L}_f > r$ then there exists $f_0 \in F$ and $(P\mathcal{L}_f/r) \geq \alpha > 1$ such that $(f - f^*) = \alpha(f_0 - f^*)$ and $P\mathcal{L}_{f_0} = r$.*

Proof. Let f and g be in $(\mathcal{L}_F)_r$ and $0 \leq \alpha \leq 1$. We have $\alpha f + (1-\alpha)g \in F$ because F is convex and for all $x \in \mathcal{X}$ and $y \in \mathbb{R}$, using the convexity of $u \to \bar{\ell}(u + f^*(x), y)$, we have

$$\begin{aligned}
&\ell_{\alpha f + (1-\alpha)g}(x,y) - \ell_{f^*}(x,y) = \bar{\ell}(\alpha(f - f^*)(x) + (1-\alpha)(g - f^*)(x) + f^*(x), y) - \bar{\ell}(f^*(x), y) \\
&\leq \alpha\big(\bar{\ell}((f - f^*)(x) + f^*(x), y) - \bar{\ell}(f^*(x), y)\big) + (1-\alpha)\big(\bar{\ell}((g - f^*)(x) + f^*(x), y) - \bar{\ell}(f^*(x), y)\big) \\
&= \alpha(\ell_f - \ell_{f^*}) + (1-\alpha)(\ell_g - \ell_{f^*})
\end{aligned}$$

and so $P\mathcal{L}_{\alpha f + (1-\alpha)g} \leq \alpha P\mathcal{L}_f + (1-\alpha)P\mathcal{L}_g$. Given that $P\mathcal{L}_f, P\mathcal{L}_g \leq r$ we also have $P\mathcal{L}_{\alpha f + (1-\alpha)g} \leq r$. Therefore, $\alpha f + (1-\alpha)g \in (\mathcal{L}_F)_r$ and $(\mathcal{L}_F)_r$ is convex.

For all $f, g \in F$, $|P\mathcal{L}_f - P\mathcal{L}_g| \leq \|f - f^*\|_{L_1(\mu)}$ so that $f \in F \to P\mathcal{L}_f$ is continuous onto F in $L_1(\mu)$ and therefore its level sets, such as $(\mathcal{L}_F)_r$, are relatively closed to F in $L_1(\mu)$.

Finally, let $f \in F$ be such that $P\mathcal{L}_f > r$. Define $\alpha_0 = \sup\{\alpha \geq 0 : f^* + \alpha(f - f^*) \in (\mathcal{L}_F)_r\}$. Note that $P\mathcal{L}_{f^* + \alpha(f - f^*)} \leq \alpha P\mathcal{L}_f = r$ for $\alpha = r/P\mathcal{L}_f$ so that $\alpha_0 \geq r/P\mathcal{L}_f$. Since $(\mathcal{L}_F)_r$ is relatively closed to F in $L_1(\mu)$, we have $f^* + \alpha_0(f - f^*) \in (\mathcal{L}_F)_r$ and in particular $\alpha_0 < 1$ otherwise, by convexity of $(\mathcal{L}_F)_r$, we would have $f \in (\mathcal{L}_F)_r$. Moreover, by maximality of α_0, $f_0 = f^* + \alpha_0(f - f^*)$ is such that $P\mathcal{L}_{f_0} = r$ and the results follows for $\alpha = \alpha_0^{-1}$. ∎

Lemma 7. *Grant Assumptions 1 and 2. On the event Ω'_K, $P\mathcal{L}_{\hat{f}} \leq \bar{r}_2^2(\gamma)$.*

Proof. Let $f \in F$ be such that $P\mathcal{L}_r > \bar{r}_2^2(\gamma)$. It follows from Lemma 6 that there exists $\alpha \geq 1$ and $f_0 \in F$ such that $P\mathcal{L}_{f_0} = \bar{r}_2^2(\gamma)$ and $f - f^* = \alpha(f_0 - f^*)$. According to (31), we have for every $k \in \{1, \dots, K\}$, $P_{B_k}\mathcal{L}_f \geq \alpha P_{B_k}\mathcal{L}_{f_0}$. Since $f_0 \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}$, on the event Ω'_K, there are strictly more than $K/2$ blocks B_k

such that $P_{B_k}\mathcal{L}_{f_0} \geq P\mathcal{L}_{f_0} - (1/4)\bar{r}_2^2(\gamma) = (3/4)\bar{r}_2^2(\gamma)$ and so $P_{B_k}\mathcal{L}_f \geq (3/4)\bar{r}_2^2(\gamma)$. As a consequence, we have

$$\sup_{f\in F\setminus(\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (-3/4)\bar{r}_2^2(\gamma) \ . \tag{36}$$

Moreover, on the event Ω'_K, for all $f \in (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}$, there are strictly more than $K/2$ blocks B_k such that $P_{B_k}(-\mathcal{L}_f) \leq (1/4)\bar{r}_2^2(\gamma) - P\mathcal{L}_f \leq (1/4)\bar{r}_2^2(\gamma)$. Therefore,

$$\sup_{f\in(\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}} \mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (1/4)\bar{r}_2^2(\gamma) \ . \tag{37}$$

We conclude from (36) and (37) that $\sup_{f\in F}\mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (1/4)\bar{r}_2^2(\gamma)$ and that every $f \in F$ such that $P\mathcal{L}_f > \bar{r}_2^2(\gamma)$ satisfies $\mathrm{MOM}_K\big(\ell_f - \ell_{f^*}\big) \geq (3/4)\bar{r}_2^2(\gamma)$. But, by definition of $\hat{f}$, we have

$$\mathrm{MOM}_K\big(\ell_{\hat{f}} - \ell_{f^*}\big) \leq \sup_{f\in F}\mathrm{MOM}_K\big(\ell_{f^*} - \ell_f\big) \leq (1/4)\bar{r}_2^2(\gamma) \ .$$

Therefore, we necessarily have $P\mathcal{L}_{\hat{f}} \leq \bar{r}_2^2(\gamma)$. ∎

Now, we prove that Ω'_K is an exponentially large event using similar argument as in Proposition 2.

Proposition 3. *Grant Assumptions 1, 2 and 7 and assume that* $(1-\beta)K \geq |\mathcal{O}|$ *and* $\beta(1-1/12-32\gamma L) > 1/2$. *Then* Ω'_K *holds with probability larger than* $1-\exp(-\beta K/1152)$.

Sketch of proof. The proof of Proposition 3 follows the same line as the one of Proposition 2. Let us precise the main differences. We set $\mathcal{F}' = (\mathcal{L}_F)_{\bar{r}_2^2(\gamma)}$ and for all $f \in \mathcal{F}'$, $z'(f) = \sum_{k=1}^K I\{|G_f(W_k)| \leq (1/4)\bar{r}_2^2(\gamma)\}$ where $G_f(W_k)$ is the same quantity as in the proof of Proposition 3. Let us consider the contraction ϕ introduced in Proposition 3. By definition of $\bar{r}_2^2(\gamma)$ and $V_K(\cdot)$, we have

$$\begin{aligned}
\mathbb{E}\phi(8(\bar{r}_2^2(\gamma))^{-1}|G_f(W_k)|) &\leq \mathbb{P}\left(|G_f(W_k)| \geq \frac{\bar{r}_2^2(\gamma)}{8}\right) \leq \frac{64}{(\bar{r}_2^2(\gamma))^2}\mathbb{E}G_f(W_k)^2 = \frac{64}{(\bar{r}_2^2(\gamma))^2}\mathbb{V}ar(P_{B_k}\mathcal{L}_f)\\
&\leq \frac{64K^2}{(\bar{r}_2^2(\gamma))^2N^2}\sum_{i\in B_k}\mathbb{V}ar_{P_i}(\mathcal{L}_f) \leq \frac{64K}{(\bar{r}_2^2(\gamma))^2N}\sup\{\mathbb{V}ar_{P_i}(\mathcal{L}_f) : f\in\mathcal{F}', i\in\mathcal{I}\}\\
&\leq \frac{64K}{(\bar{r}_2^2(\gamma))^2N}\sup\{\mathbb{V}ar_{P_i}(\mathcal{L}_f) : P\mathcal{L}_f \leq \bar{r}_2^2(\gamma), i\in\mathcal{I}\} \leq \frac{1}{24} \ .
\end{aligned}$$

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 2, we obtain with probability larger than $1-\exp(-|\mathcal{K}|/1152)$, for all $f \in \mathcal{F}'$,

$$z(f) \geq |\mathcal{K}|(1-1/12) - \frac{32LK}{N}\mathbb{E}\sup_{f\in\mathcal{F}'}\frac{1}{\bar{r}_2^2(\gamma)}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| . \tag{38}$$

Now, it remains to use the definition of $\bar{r}_2^2(\gamma)$ to bound the expected supremum in the right-hand side of (38) to get

$$\mathbb{E}\sup_{f\in\mathcal{F}'}\frac{1}{\bar{r}_2(\gamma)^2}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\sigma_i(f-f^*)(X_i)\right| \leq \frac{\gamma|\mathcal{K}|N}{K}. \tag{39}$$

Proof of Theorem 4. The proof of Theorem 4 follows from Lemma 7 and Proposition 3 for $\beta = 4/7$ and $\gamma = 1/(768L)$.

B Proof of Lemma 1

Proof. We have

$$\frac{1}{\sqrt{N}}\mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\le r}\sum_{i=1}^N \sigma_i(f-f^*)(X_i)=\mathbb{E}\sup_{t\in\mathbb{R}^p:\mathbb{E}<t,X>^2\le r^2}<t,\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_i X_i> \quad .$$

Let $\Sigma=\mathbb{E}X^TX$ denote the covariance matrix of X and consider its SVD, $\Sigma=QDQ^T$ where $Q=[Q_1|\cdots|Q_p]\in\mathbb{R}^{p\times p}$ is an orthogonal matrix and D is a diagonal $p\times p$ matrix with non-negative entries. For all $t\in\mathbb{R}^p$, we have $\mathbb{E}\big\langle X,t\big\rangle^2=t^T\Sigma t=\sum_{j=1}^p d_j\big\langle t,Q_j\big\rangle^2$. Then

$$\begin{aligned}
&\mathbb{E}\sup_{t\in\mathbb{R}^p:\sqrt{\mathbb{E}<t,X>^2}\le r}<t,\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i>=\mathbb{E}\sup_{t\in\mathbb{R}^p:\sqrt{\mathbb{E}<t,X>^2}\le r}\Big\langle\sum_{j=1}^p<t,Q_j>Q_j,\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\Big\rangle\\
&=\mathbb{E}\sup_{t\in\mathbb{R}^p:\sqrt{\sum_{j=1}^p d_j<t,Q_j>^2}\le r}\ \sum_{j=1:d_j\neq 0}^p\sqrt{d_j}<t,Q_j>\Big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\Big\rangle\\
&\le r\mathbb{E}\sqrt{\sum_{j=1:d_j\neq0}^p\Big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\Big\rangle^2}\le r\sqrt{\mathbb{E}\sum_{j=1:d_j\neq0}^p\Big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\Big\rangle^2}\quad .
\end{aligned}$$

Moreover, for any j such that $d_j\neq 0$,

$$\begin{aligned}
\mathbb{E}\Big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\Big\rangle^2&=\mathbb{E}\frac{1}{N}\sum_{k,l=1}^N\sigma_l\sigma_k\big\langle\frac{Q_j}{\sqrt{d_j}},X_k\big\rangle\big\langle\frac{Q_j}{\sqrt{d_j}},X_l\big\rangle=\frac{1}{N}\sum_{k=1}^N\mathbb{E}\big\langle\frac{Q_j}{\sqrt{d_j}},X_k\big\rangle^2\\
&=\frac{1}{N}\sum_{k=1}^N\Big(\frac{Q_j}{\sqrt{d_j}}\Big)^T\mathbb{E}X_k^TX_k\Big(\frac{Q_j}{\sqrt{d_j}}\Big)=\frac{1}{N}\sum_{k=1}^N\Big(\frac{Q_j}{\sqrt{d_j}}\Big)^T\Sigma\Big(\frac{Q_j}{\sqrt{d_j}}\Big)
\end{aligned}$$

By orthonormality, $Q^TQ_j=e_j$ and $Q_j^TQ=e_j^T$, then, for any j such that $d_j\neq0$,

$$\mathbb{E}\Big\langle\frac{Q_j}{\sqrt{d_j}},\frac{1}{\sqrt{N}}\sum_{i=1}^N\sigma_iX_i\Big\rangle^2=\frac{1}{N}\sum_{k=1}^N\frac{1}{d_j}e_j^TDe_j=1\quad .$$

Finally, we obtain

$$\frac{1}{\sqrt{N}}\mathbb{E}\sup_{f\in F:\|f-f^*\|_{L_2}\le r}\sum_{i=1}^N\sigma_i(f-f^*)(X_i)\le r\sqrt{\sum_{j=1}^p\mathbf{1}_{\{d_j\neq0\}}}=r\sqrt{\text{Rank}\Sigma)}$$

and therefore the fixed point $\tilde r_2(\gamma)$ is such that

$$\begin{aligned}
\tilde r_2(\gamma)=&\inf\Big\{r>0,\forall J\in\mathcal{I}:|J|\ge N/2,\ \ \mathbb{E}\sup_{t\in\mathbb{R}^p:\sqrt{\mathbb{E}<t-t^*,X>^2}\le r}\sum_{i\in J}\sigma_i<X_i,t-t^*>\le r^2|J|\gamma\Big\}\\
&\le\inf\Big\{r>0,\forall J\in\mathcal{I}:|J|\ge N/2,\quad r\sqrt{\text{Rank}(\Sigma)}\le r^2\sqrt{|J|}\gamma\Big\}\le\sqrt{\frac{\text{Rank}(\Sigma)}{2\gamma^2N}}\quad .
\end{aligned}$$

■

C Proofs of the results of Section 5

C.1 Proof of Theorem 6

Let $f \in F$ be such that $\|f - f^*\|_{L_2} \leq r$. For all $x \in \mathcal{X}$ denote by $F_{Y|X=x}$ the conditional c.d.f. of Y given $X = x$. We have

$$\mathbb{E}\Big[\ell_f(X,Y)|X=x\Big] = (\tau - 1)\int_{y \leq f(x)} (y - f(x))F_{Y|X=x}(\mathrm{d}y) + \tau \int_{y > f(x)} (y - f(x))F_{Y|X=x}(\mathrm{d}y)$$
$$= \int_{y > f(x)} (y - f(x))F_{Y|X=x}(\mathrm{d}y) + (\tau - 1)\int_{\mathbb{R}} (y - f(x))F_{Y|X=x}(\mathrm{d}y) \ .$$

By Fubini's theorem,

$$\int_{z \geq f(x)} (1 - F_{Y|X=x}(z))\mathrm{d}z = \int_{z \geq f(x)} \Big(1 - P(Y \leq z|X=x)\Big)\mathrm{d}z = \int_{z \geq f(x)} \mathbb{E}[I\{Y > z\}|X=x]\mathrm{d}z$$
$$= \int\int I\{y > z \geq f(x)\}f_{Y|X=x}(y)\mathrm{d}y\mathrm{d}z = \int I\{y > f(x)\}(y - f(x))f_{Y|X=x}(y)\mathrm{d}y$$
$$= \int_{z > f(x)} (z - f(x))F_{Y|X=x}(\mathrm{d}z) \ .$$

Therefore,

$$\mathbb{E}\Big[\ell_f(X,Y)|X=x\Big] = \int_{y \geq f(x)} (1 - F_{Y|X=x}(y))\mathrm{d}y + (\tau - 1)\Big(\int_{\mathbb{R}} yF_{Y|X=x}(\mathrm{d}y) - f(x)\Big)$$
$$= g(x, f(x)) + (\tau - 1)\int_{\mathbb{R}} yF_{Y|X=x}(\mathrm{d}y) \ .$$

where $g(x,a) = \int_{y \geq a}(1 - F_{Y|X=x}(y))\mathrm{d}y + (1-\tau)a$. It follows that

$$P\mathcal{L}_f = \mathbb{E}[g(X, f(X)) - g(X, f^*(X))] \ . \qquad (40)$$

Given that for all $x \in \mathcal{X}$, $\partial_a g(x, f^*(x)) = 0$ and $\partial_a^2 g(x,a) = f_{Y|X=x}(a)$ for all $a \in \mathbb{R}$, a Taylor expansion yields, for some z in $\big[\min(f(x), f^*(x)), \max(f(x), f^*(x))\big]$,

$$g(x, f(x)) - g(x, f^*(x)) = \frac{\big(f(x) - f^*(x)\big)^2}{2} f_{Y|X=x}(z) \ .$$

Consider $A = \{x \in \mathcal{X}, |f(x) - f^*(x)| \leq 2(C')^2 r\}$. Given that $\|f - f^*\|_{L_2} \leq r$, by Markov's inequality, $P(X \in A) \geqslant 1 - 1/(4(C')^4)$. Moreover, for any $x \in A$, by Assumption 10, $g(x, f(x)) - g(x, f^*(x)) \geqslant \alpha\frac{(f(x) - f^*(x))^2}{2}$. Plugging this bound in (40) yields

$$\frac{2P\mathcal{L}_f}{\alpha} \geqslant \mathbb{E}[I_A(X)(f(X) - f^*(X))^2] = \|f - f^*\|_{L_2}^2 - \mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \ . \qquad (41)$$

By Cauchy-Schwarz and Markov's inequalities,

$$\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \sqrt{\mathbb{E}[I_{A^c}(X)]\mathbb{E}[(f(X) - f^*(X))^4]} \leqslant \frac{\|f - f^*\|_{L_4}^2}{2(C')^2} \ .$$

By Assumption 9, it follows that $\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \frac{\|f - f^*\|_{L_2}^2}{2}$ and we conclude with (41).

C.2 Proof of Theorem 7

Let $f \in F$ be such that $\|f - f^*\|_{L_2} \leq r$. Write first that

$$P\mathcal{L}_f = \mathbb{E}_X\mathbb{E}\Big[\rho_H(Y - f(x)) - \rho_H(Y - f^*(x))|X = x\Big] = \mathbb{E}\big[g(X, f(X)) - g(X, f^*(X))\big]$$

where for all $x \in \mathcal{X}$ and $a \in \mathbb{R}$, $g(x, a) = \mathbb{E}[\rho_H(Y - a)|X = x]$. Let $F_{Y|X=x}$ denote the c.d.f. of Y given $X = x$. Using that for all $x \in \mathcal{X}, \partial_2 g(x, f^*(x)) = 0$ a second Taylor expansion yields that there exists $z \in [\min(f(x), f^*(x)), \max(f(x), f^*(x))]$ such that $g(x, f(x)) - g(x, f^*(x)) = (1/2)(f(x) - f^*(x))^2 g''(z)$. Moreover, for all $a \in \mathbb{R}$,

$$\partial_2 g(x, a) = -\int_{a-\delta}^{a+\delta} (y - a \mathrm{d}F_{Y|X=x}(y) + \delta \int_{-\infty}^{a-\delta} \mathrm{d}F_{Y|X=x}(y) - \delta \int_{a+\delta}^{+\infty} \mathrm{d}F_{Y|X=x}(y) = \int_{a-\delta}^{a+\delta} F_{Y|X=x}(y)\mathrm{d}y - \delta$$

and so $\partial_a^2 g(x, a) = F_{Y|X=x}(a + \delta) - F_{Y|X=x}(a - \delta)$.

Now, let $A = \{x \in \mathcal{X} : |f(x) - f^*(x)| \leq 2(C')^2 r\}$. It follows from Assumption 10 that $P\mathcal{L}_f \geq (\alpha/2)\mathbb{E}[(f(X) - f^*(X))^2 I_A(X)]$. Since $\|f - f^*\|_{L_2} \leq r$, by Markov's inequality, $P(X \in A) \geqslant 1 - 1/(4(C')^4)$ and so by Cauchy-Schwarz, we obtain

$$\mathbb{E}[I_{A^c}(X)(f(X) - f^*(X))^2] \leqslant \frac{\|f - f^*\|_{L_4}^2}{2(C')^2} \leqslant \frac{\|f - f^*\|_{L_2}^2}{2}$$

where we used Assumption 9 in the last inequality. Finally, we obtain $P\mathcal{L}_f \geqslant (\alpha/4)\|f - f^*\|_{L_2}^2$.

C.3 Proof of Theorem 9

Let $f \in F$ be such that $\|f - f^*\|_{L_2} \leq r$. Write first that $P\mathcal{L}_f = \mathbb{E}_X\mathbb{E}\Big[g(X, f(X)) - g(X, f^*(X))\Big]$ where for all $x \in \mathcal{X}$ and $a \in \mathbb{R}$, $g(x, a) = \eta(x)\log(1 + \exp(-a)) + (1 - \eta(x))\log(1 + \exp(a))$. Using that for all $x \in \mathcal{X}, \partial_2 g(x, f^*(x)) = 0$ and a second Taylor expansion yield

$$g(x, f(x)) - g(x, f^*(x)) = \partial_2^2 g(x, z_x)\frac{(f(x) - f^*(x))^2}{2} = \frac{e^{-z_x}}{(1 + e^{-z_x})^2}\frac{(f(x) - f^*(x))^2}{2}$$

for some $z_x \in [\min(f(x), f^*(x)), \max(f(x), f^*(x))]$.

Now, let $\delta > 0$ to be optimized later and $A = \{x \in \mathcal{X} : |f^*(x)| \leq c_0, |f(x) - f^*(x)| \leq r/\delta\}$. By Markov's inequality, $\mathbb{P}(X \in A) \geq 1 - \delta^2 - e^{-|c_0|}$. Therefore, we get

$$P\mathcal{L}_f \geq c_1(r, \delta, c_0)\mathbb{E}[I_A(X)(f(X) - f^*(X))^2]$$

for $c_1(c, \delta, c_0) = e^{-(r/\delta + c_0)}/(1 + e^{-(r/\delta + c_0)})^2$. We conclude with Assumption 9 and the same analysis as in the two previous proofs.

References

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

[2] P. Alquier, V. Cottet, and G. Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *arXiv preprint arXiv:1702.01402*, 2017.

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

[4] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 5:19–53, 2011.

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

[6] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[7] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[8] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

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

[10] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[11] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[12] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

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

[15] Andreas Elsener and Sara van de Geer. Robust low-rank matrix estimation. *arXiv preprint arXiv:1603.09071*, 2016.

[16] Qiyang Han and Jon A. Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. 2017.

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

[18] 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.

[19] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[20] Vladimir Koltchinskii. Empirical and rademacher processes. In *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, pages 17–32. Springer, 2011.

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

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

[23] G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. *preprint in arXiv:1701.01961*, 2017.

[24] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.

[25] Guillaume Lecué et al. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

[26] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[27] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via mom minimization. *arXiv:1808.03106*, 2018.

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

[29] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in JEMS*, 2016.

[30] Gabor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Preprint available on arXiv:1701.04112*, 2017.

[31] Gabor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *To appear in Ann. Statist. arXiv:1702.00482*, 2017.

[32] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

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

[34] Shahar Mendelson. On multiplier processes under weak moment assumptions. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 301–318. Springer, Cham, 2017.

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

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

[38] 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. Modern methods and classical problems.

[39] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[40] Sara van de Geer. *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics.* Springer, [Cham], 2016. 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].

[41] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[42] Vladimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[43] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.

[44] W.-X. Zhou, K. Bose, J. Fan, and H. Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. 2018.