Learning from MOM's principles : Le Cam's approach

Guillaume Lecué[a], Matthieu Lerasle[b]

[a] *CREST, CNRS, Université Paris Saclay*
[b] *Laboratoire de Mathmatiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris Saclay*

Abstract

New robust estimators are introduced, derived from median-of-means principle and Le Cam's aggregation of tests. Minimax sparse rates of convergence are obtained with exponential probability, under weak moment's assumptions and possible contamination of the dataset. These derive from general risk bounds of the following informal structure

$$\max\left(\text{minimax rate in the i.i.d. setup}, \frac{\text{number of outliers}}{\text{number of observations}}\right) .$$

In this result, the number of outliers may be as large as *(number of data)* $\times$*(minimax rate)* without affecting the rates. As an example, minimax rates $s\log(ed/s)/N$ of recovery of s-sparse vectors in $\mathbb{R}^d$ holding with exponentially large probability, are deduced for median-of-means versions of the LASSO when the noise has q_0 moments for some $q_0 > 2$, the entries of the design matrix have $C_0\log(ed)$ moments and the dataset is corrupted by up to $C_1 s\log(ed/s)$ outliers.

Keywords: robust statistics, statistical learning, high dimensional statistics.
2010 MSC: 62G35, 62G08.

1. Introduction

Consider the problem of estimating minimizers of the integrated square-loss over a convex class of functions : $f^* \in \operatorname{argmin}_{f\in F} P(Y-f(X))^2$ based on a data set $(X_i,Y_i)_{i=1,\dots,N}$, where the outputs Y are real-valued and the inputs X take values in any measurable space $\mathcal{X}$.

Let P_N denote the empirical distribution based on the sample $(X_i,Y_i)_{i=1,\dots,N}$ and let reg : $F \to \mathbb{R}_+$ denote a regularizing function, regularized versions (RERM) of Empirical Risk Minimizers (ERM) Vapnik (1998); Vapnik and Chervonenkis (1974) are defined by

$$\hat{f}_N^{\mathrm{RERM}} \in \operatorname*{argmin}_{f\in F}\{P_N(Y-f(X))^2 + \mathrm{reg}(f)\} .$$

URL: guillaume.lecue@ensae.fr (Guillaume Lecué), matthieu.lerasle@math.u-psud.fr (Matthieu Lerasle)

These estimators are optimal in i.i.d. subgaussian setups but suffer several drawbacks when data are heavy-tailed or corrupted by "outliers", see Catoni (2012); Huber and Ronchetti (2009), a situation that occurs with declarative data on internet, for storage/transfer issues or when one applies some compression algorithm. They may also be observations met in biology as in the classical *eQTL (Expression Quantitative Trait Loci and The Phenogen Database)* from Saba et al. (2008).

To overcome the problem, the most common strategy is to replace the square-loss in the definition of RERM. Huber (1964) proposed a loss that interpolates between square and absolute loss to produce an estimator between the unbiased (but non robust) empirical mean and the (more robust but biased) empirical median. Huber's estimators have been intensively studied asymptotically by Huber (1964); Huber and Ronchetti (2009), non-asymptotic results have also been obtained more recently by Chichignoud and Lederer (2014); Mendelson (2015b); Fan et al. (2017) for example. An alternative approach has been proposed by Catoni (2012) and used in learning frameworks such as least-squares regression by Audibert and Catoni (2011) and for more general loss functions by Brownlees et al. (2015).

Another line of research to build robust estimators and robust selection procedures was initiated by Le Cam (1973, 1986) and further developed by Birgé (1984); Birgé (2006), Baraud (2011) and Baraud et al. (2017). It is based on *comparisons* or *tests* between elements of F. More precisely, the approach builds on tests statistics $T_N(g,f)$ comparing f and g. These tests define the sets $\mathcal{B}_{T_N}(f)$ of all g's that have been preferred to f and the final estimator $\hat{f}$ is a minimizer of the diameter of $\mathcal{B}_{T_N}(f)$. The measure of diameter is directly related to statistical performance one seeks for the estimator. These methods mostly focus on Hellinger loss and are generally considered difficult to compute, see however Baraud et al. (2014); Sart (2014).

In a related but different approach, Lugosi and Mendelson (2017) have recently introduced "median-of-means tournaments" where Median-of-means estimators of Alon et al. (1999); Jerrum et al. (1986); Nemirovsky and Yudin (1983) are used to compare elements of F. A "champion" is an element $\hat{f}$ that is better than all g sufficiently far from it. As L^2-distances cannot be computed, these diameters are evaluated using an estimator of distances that only provides a reliable estimate when one of the functions is f^*, which is sufficient to bound the risk of champions.

This paper studies estimators derived from Le Cam's procedure based on regularized median-of-means (MOM) tests (see Section 4.1). The main advantage of MOM's tests over Le Cam's original ones is that they allow for more classical loss functions than Hellinger loss. This idea is illustrated on the square-loss. Compared to Huber or Catoni's losses, this approach allows to control easily the risk of our estimators by using classical tools from empirical process theory and to tackle the problem of "aggressive" outliers.

The radii of the sets $\mathcal{B}_{T_N}(f)$ are computed for regularization and L_P^2 norms. The regularization norm is chosen in advance by the statistician to promote

sparsity or smoothness and can be used to build estimators. However, it is not sufficient to derive estimators with small L_P^2 risk and this is why it is necessary to evaluate L_P^2-diameters. The L_P^2-norm is unknown in general since it depends on the distribution of X and cannot be properly estimated by the $L_{P_N}^2$-empirical metric without subgaussian properties of the design vector X. Fortunately, the L_P^2-norm can be estimated by median-of-means estimators. The estimators used in this paper differ from those of Lugosi and Mendelson (2017) but also provide reliable estimators only to distances to oracle. To handle simultaneously both regularization and L_P^2 norms, an extension of Le Cam's original idea is also required. Our first result shows that our new estimators are well localized w.r.t. both regularization and L_P^2 norms.

The closest work is certainly that of Lugosi and Mendelson (2017) even if some important differences can be noticed. Here, the approach does not require a partition of the dataset into three parts and the robustness to corrupted datasets is stressed. This is an important feature in practice even if this robustness follows from a simple argument. In addition, MOM estimators rely on a data splitting into K blocks and this parameter drives the resulting statistical performance of the estimator (cf. Devroye et al. (2016)). Our main results show that optimal rates can be achieved when K is chosen using parameters that depend on both the number of outliers and the oracle f^*, such as its sparsity, which are unknown quantities. To bypass this problem, the strategy of Lepski (1991) is used as in Devroye et al. (2016) to select K adaptively and get a fully data-driven procedure with optimal performance. We also refer to Birgé (1984) where a closely related construction is also proposed in the proof of Theorem 1. This latter adaptive step is not performed in Lugosi and Mendelson (2017) and the problem of regularization is also not studied.

There are four important features in our approach. First, all results are proved under weak $L^{2+\epsilon}$ moment assumptions on the noise, which is an almost minimal condition for the problem to make sense and the class F is only assumed to satisfy a weak "L_2/L_1" comparison. Second, performance of the estimators are not affected by the presence of complete outliers, as long as their number remains comparable to *(number of observations)*×*(rates of convergence)*. Third, all results are non-asymptotic and the regression function $x \mapsto \mathbb{E}[Y|X = x]$ is never assumed to belong to the class F. In particular, the noise $Y - f^*(X)$ can be correlated with X. Finally, even "informative data", those that are not "outliers", are not requested to be i.i.d. $\sim P$, but only to have close first and second moments for all $f \in F - \{f^*\}$. Nevertheless, the estimators are shown to behave as the ERM when data are i.i.d. $\sim P$, $\mathbb{E}[Y|X = \cdot] \in F$, the noise $\zeta = Y - f^*(X)$ and the class F are Gaussian and the noise is independent from the design.

From a mathematical point of view, our results are based on a slight extension of the Small Ball Method (SBM) of Koltchinskii and Mendelson (2015); Mendelson (2014a). Indeed, the SBM was initially introduced to prove lower bounds on nonnegative empirical processes. The method is extended here to control centered empirical processes. All other arguments are standard and the approach is thus easily reproducible in other statistical learning frameworks.

The paper is organized as follows. Section 2 briefly presents the general setting. Section 3 presents Le Cam's construction of estimators based on tests. The construction of estimators and the main assumptions are gathered in Section 4. Our main theorems are stated in Section 5 and proved in Section 6.

Notation. For any real number x, let $\lfloor x \rfloor$ denote the largest integer smaller than x and let $[x] = \{1, \dots, \lfloor x \rfloor\}$ if $x \geq 1$. For any finite set A, let $|A|$ denote its cardinality. All along the paper, $(c_i)_{i\in\mathbb{N}}$ denote absolute constants which may vary from line to line and $\theta_.$, with various subscripts, denote real valued parameters introduced in the assumptions. Finally, for any set $\mathcal{G}$ for which it makes sense, for any $g \in \mathcal{G}$, $c \geq 0$ and $\mathcal{C} \subset \mathcal{G}$,

$$g + c\mathcal{C} = c\mathcal{C} + g = \{h : \exists g' \in \mathcal{C} \text{ such that } h = g + cg'\} \ .$$

Let also $g + \mathcal{G} = g + 1\mathcal{G}$. We also denote by $I(g \in \mathcal{C})$ the indicator function of the set $\mathcal{C}$ which equals to 1 when $g \in \mathcal{C}$ and 0 otherwise.

2. Setting

Let $\mathcal{X}$ denote a measurable space and let $(X, Y), (X_i, Y_i)_{i\in[N]}$ denote random variables taking values in $\mathcal{X}\times\mathbb{R}$, with respective distributions $P, (P_i)_{i\in[N]}$. Given a probability distribution Q, let L^2_Q denote the space of all functions f from $\mathcal{X}$ to $\mathbb{R}$ such that $\|f\|_{L^2_Q} < \infty$ where $\|f\|_{L^2_Q} = \left(Qf^2\right)^{1/2}$. Let $F \subset L^2_P$ denote a convex class of functions $f : \mathcal{X} \to \mathbb{R}$. Assume that $PY^2 < \infty$ and let, for all $f \in F$,

$$R(f) = P\big[(Y - f(X))^2\big], \quad f^* \in \operatorname*{argmin}_{f\in F} R(f) \text{ and } \zeta = Y - f^*(X) \ .$$

Let $\|\cdot\|$ denote a norm defined onto a linear subspace E of L^2_P containing F.

3. Learning from tests

3.1. General Principle

This section details a construction of Le Cam's to build estimators from increment estimators. By definition of the oracle f^*, one has

$$f^* = \operatorname*{argmin}_{f\in F} R(f) = \operatorname*{argmin}_{f\in F} \sup_{g\in F} \{R(f) - R(g)\} \ .$$

As $T_{\mathrm{id}}(g, f) = R(f) - R(g)$ depends on P, it has to be estimated by test statistics $T(g, f, (X_i, Y_i)_{i\in[N]}) \equiv T_N(g, f)$ which are real random variables such that

$$T_N(f, g) + T_N(g, f) = 0 \ . \tag{1}$$

These statistics are used to *compare* f to g, simply by saying that g T_N-*beats* f iff $T_N(g, f) \geq 0$. In this paper, the statistics $T_N(g, f)$ are median-of-means estimators of $R(f) - R(g)$ (cf. (9) in Section 4.1).

Le Cam's construction. Let $(T_N(g,f))_{f,g\in F}$ denote a collection of test statistics and let $d(\cdot,\cdot)$ denote a pseudo-distance on F, that may be data-dependent, measuring (or related to) the risk. For all $f \in F$, let

$$\mathcal{B}_{T_N}(f) = \{g \in F : T_N(g,f) \geq 0\}$$

be the set of all functions $g \in F$ that beat f. By (1), either $f \in \mathcal{B}_{T_N}(g)$ or $g \in \mathcal{B}_{T_N}(f)$ (both happen if $T_N(f,g) = 0$), hence $d(f,g) \leq C_{T_N}(f) \vee C_{T_N}(g)$. In particular, for all $f \in F$,

$$d(f,f^*) \leq C_{T_N}(f) \vee C_{T_N}(f^*) \ . \tag{2}$$

Eq (2) suggests to define the estimator

$$\hat{f}_{T_N} \in \operatorname*{argmin}_{f\in F} C_{T_N}(f) = \operatorname*{argmin}_{f\in F} \sup_{g\in\mathcal{B}_{T_N}(f)} d(f,g) \ . \tag{3}$$

This estimator satisfies, from Eq (2),

$$d(\hat{f}_{T_N}, f^*) \leq C_{T_N}(f^*) \ . \tag{4}$$

Risk bounds for $\hat{f}_{T_N}$ follow from (4), isometry properties of d as a distance to oracle and upper bounds on the radii of $\mathcal{B}_{T_N}(f^*)$.

Remark 1. *More generally, one can compare only the elements of a subset $\mathcal{F} \subset F$, typically a maximal ϵ-net by introducing for all $f \in \mathcal{F}$, the set*

$$\mathcal{B}_{T_N}(f,\mathcal{F}) = \{g \in \mathcal{F} : T_N(g,f) \geq 0\} \tag{5}$$

and then by minimizing the diameter of $\mathcal{B}_{T_N}(f,\mathcal{F})$ over $\mathcal{F}$. This usually improves the rates of convergence for constant deviation results when there is a gap in Sudakov's inequality of the localized sets of F, cf. (Lecué and Mendelson, 2013, Section 5).

Dealing with regularization : the link function. Statistical performance of estimators and the radius of $\mathcal{B}_{T_N}(f^*)$ can be measured by two norms: the regularization norm $\|\cdot\|$ and $\|.\|_{L_P^2}$. As one distance only is used in (3), we extend Le Cam's construction to handle two metrics simultaneously.

Assume first that $d(f,g) = \|f-g\|_{L_P^2}$ can be computed for all $f,g \in F$ (this is the case if the distribution of the design is known). Remark that

$$C_{T_N}(f) = \sup_{g\in\mathcal{B}_{T_N}(f)} \|f-g\| = \min\left\{\rho \geq 0 : \sup_{g\in\mathcal{B}_{T_N}(f)} \|g-f\| \leq \rho\right\}$$

can be used to get estimation results w.r.t. the regularization norm $\|\cdot\|$. But if one also wants estimation results w.r.t. the L_P^2-norm then the main point is to design a link function $r(\cdot)$ between the two metrics. In a nutshell, the value $r(\rho)$ is the L_P^2-minimax rate of convergence in a ball of radius ρ for the regularization norm (cf. (10) in Section 4.3 for a formal definition). Then one can define

$$C_{T_N}^{(2)}(f) = \min\left\{\rho \geq 0 : \sup_{g\in\mathcal{B}_{T_N}(f)} \|g-f\| \leq \rho \text{ and } \sup_{g\in\mathcal{B}_{T_N}(f)} d(f,g) \leq r(\rho)\right\} \ .$$

Theorem 1 shows that a minimizer $\hat{f}^{(2)}$ of $C^{(2)}_{T_N}$ has both $\|\hat{f}^{(2)} - f^*\|$ and $d(\hat{f}^{(2)}, f^*)$ properly controlled.

Dealing with unknown norms : the isometry property. In general, L^2_P-distances cannot be directly computed and have to be estimated. It is estimated by MOM estimators $d_N(f,g)$ of all $d(f,g)$ (cf. Section 4.4) that are relevant as distances to the oracle, see Lemma 3. The final estimator is any minimizer of

$$C''_{T_N}(f) = \min\left\{\rho \geq 0 : \sup_{g\in\mathcal{B}_{T_N}(f)} \|g-f\| \leq \rho \text{ and } \sup_{g\in\mathcal{B}_{T_N}(f)} d_N(f,g) \leq r(\rho)\right\} \ .$$

4. Regularized MOM estimators

4.1. Quantile of means processes and median-of-means tests

Start with a few notations, see van der Vaart and Wellner (1996). For all $\alpha \in [0,1]$, $\ell \geq 1$ and $z \in \mathbb{R}^\ell$, let $\mathcal{Q}_\alpha(z)$ denote the set of α-quantiles of z. For any non-empty subset $B \subset [N]$ and a function $f : \mathcal{X}\times\mathbb{R} \to \mathbb{R}$, let

$$P_B f = \frac{1}{|B|}\sum_{i\in B} f(X_i, Y_i) \ .$$

Let $K \in [N]$ and let $(B_1, \ldots, B_K)$ denote an equipartition of $[N]$ into bins of size $|B_k| = N/K$ (assuming for simplicity that K divides N). For any $\alpha \in [0,1]$ and any $f : \mathcal{X}\times\mathbb{R} \to \mathbb{R}$, the set of α-quantiles of empirical means is denoted by

$$\mathcal{Q}_{\alpha,K}(f) = \mathcal{Q}_\alpha\left((P_{B_k} f)_{k\in[K]}\right) \ .$$

With a slight abuse of notations, we shall repeatedly denote by $Q_{\alpha,K}(f)$ any element in $\mathcal{Q}_{\alpha,K}(f)$ and write $Q_{\alpha,K}(f) = u$ if $u \in \mathcal{Q}_{\alpha,K}(f)$, $Q_{\alpha,K}(f) \geq u$ if there is at least $(1-\alpha)K$ blocks B_k such that $P_{B_k} f \geqslant u$, $Q_{\alpha,K}(f) \leq u$ if there is more than αK blocks B_k such that $P_{B_k} f \leqslant u$, and $Q_{\alpha,K}(f) + Q_{\alpha',K}(f')$ any element in the Minkowski sum $\mathcal{Q}_{\alpha,K}(f) + \mathcal{Q}_{\alpha',K}(f')$. Let also $\mathrm{MOM}_K(f) = Q_{1/2,K}(f)$ denote an empirical median of the empirical means on the blocks B_k. Empirical quantiles satisfy for any $c \geq 0$, $f, f' : \mathcal{X}\times\mathbb{R} \to \mathbb{R}$ and $\alpha \in [0,1]$,

$$\mathcal{Q}_{\alpha,K}(cf) = c\mathcal{Q}_{\alpha,K}(f) \ , \tag{6}$$

$$\mathcal{Q}_{\alpha,K}(-f) = -\mathcal{Q}_{1-\alpha,K}(f) \ , \tag{7}$$

$$\sup\left\{\mathcal{Q}_{1/4,K}(f) + \mathcal{Q}_{1/4,K}(f')\right\} \leq \inf \mathcal{Q}_{1/2,K}(f+f') \ . \tag{8}$$

With some abuse of notations, these properties will be written respectively

$$Q_{\alpha,K}(cf) = cQ_{\alpha,K}(f), \qquad Q_{\alpha,K}(-f) = -Q_{1-\alpha,K}(f) \ ,$$
$$Q_{1/4,K}(f) + Q_{1/4,K}(f') \leq \mathrm{MOM}_K\left[f+f'\right] \ .$$

The *regularization* parameter $\lambda > 0$ is introduced to balance between data adequacy and regularization. The (quadratic) loss and regularized (quadratic) loss are respectively defined on $F \times \mathcal{X} \times \mathbb{R}$ as the real valued functions such that

$$\ell_f(x,y) = (y - f(x))^2, \quad \ell_f^\lambda = \ell_f + \lambda \|f\|, \qquad \forall (f,x,y) \in F \times \mathcal{X} \times \mathbb{R} \ .$$

To compare/test functions f and g in F, median-of-means tests between f and g are now defined by

$$T_{K,\lambda}(g,f) = \mathrm{MOM}_K \left[\ell_f^\lambda - \ell_g^\lambda\right] = \mathrm{MOM}_K \left[\ell_f - \ell_g\right] + \lambda(\|f\| - \|g\|) \ . \quad (9)$$

From (7), $T_{K,\lambda}$ satisfies (1) and is a tests statistic in the sense of Section 3.

4.2. Main assumptions

Assume that $[N] = \mathcal{O} \cup \mathcal{I}$ and that the "outliers" $(X_i, Y_i)_{i \in \mathcal{O}}$ bring no information on f^*. Inliers, or informative data $(X_i, Y_i)_{i \in \mathcal{I}}$ are supposed to satisfy the following assumptions.

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

$$\|f - f^*\|_{L^2_{P_i}} \leq \theta_r \|f - f^*\|_{L^2_P} \ .$$

Of course, Assumption 1 holds in the i.i.d. framework, with $\theta_r = 1$ and $\mathcal{I} = [N]$. The second assumption bounds the correlation between the noise function $\zeta : (y,x) \in \mathbb{R} \times \mathcal{X} \to y - f^*(x)$ and the design on the shifted class $F - f^*$ in L^2_Q for all $Q \in \{P, (P_i)_{i \in \mathcal{I}}\}$.

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

$$var_Q(\zeta(f - f^*)) = Q\left[\zeta^2 (f - f^*)^2 - [Q(\zeta(f - f^*))]^2\right] \leq \theta_m^2 \|f - f^*\|^2_{L^2_P} \ .$$

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

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

Examples of functions and distributions satisfying Assumptions 2 and 3 can be found in Appendix .1.

4.3. Complexity parameters and the link function

This section defines the link function $r(\cdot)$ required in the extension of Le Cam's approach. For any $\rho \geq 0$ and any $f \in E$, let

$$B(f,\rho) = \{g \in E : \|f - g\| \leq \rho\}, \qquad S(f,\rho) = \{g \in E : \|g - f\| = \rho\} \ .$$

Definition 1. *Let $(\epsilon_i)_{i\in\mathcal{I}}$ be independent Rademacher random variables, independent from $(X_i, Y_i)_{i\in\mathcal{I}}$ and let $\mathcal{J} = \{J \subset \mathcal{I}, |J| \geq |\mathcal{I}|/2\}$. For any $\gamma_Q, \gamma_M > 0$ and $\rho > 0$ let $F_{f^\star,\rho,r} = \{f \in F \cap B(f^\star, \rho) : \|f - f^\star\|_{L_P^2} \leq r\}$, let*

$$\mathfrak{Q}_{f^\star,\rho}^{\gamma_Q} = \left\{ r > 0 : \forall J \in \mathcal{J},\ \mathbb{E} \sup_{f\in F_{f^\star,\rho,r}} \left| \sum_{i\in J} \epsilon_i (f - f^\star)(X_i) \right| \leq \gamma_Q |J| r \right\} \ ,$$

$$\mathfrak{M}_{f^\star,\rho}^{\gamma_M} = \left\{ r > 0 : \forall J \in \mathcal{J},\ \mathbb{E} \sup_{f\in F_{f^\star,\rho,r}} \left| \sum_{i\in J} \epsilon_i (Y_i - f^\star(X_i))(f - f^\star)(X_i) \right| \leq \gamma_M |J| r^2 \right\}$$

and denote by

$$r_Q(\rho, \gamma_Q) = \sup_{f^\star\in F} \{\inf \mathfrak{Q}_{f^\star,\rho}^{\gamma_Q}\}, \qquad r_M(\rho, \gamma_M) = \sup_{f^\star\in F} \{\inf \mathfrak{M}_{f^\star,\rho}^{\gamma_M}\} \ .$$

*The **link function** is any continuous and non-decreasing function $r : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $\rho > 0$*

$$r(\rho) = r(\rho, \gamma_Q, \gamma_M) \geq \max(r_Q(\rho, \gamma_Q), r_M(\rho, \gamma_M)) \ . \tag{10}$$

4.4. The estimators

Let $(T_{K,\lambda}(g, f))_{f,g\in F}$ denote the family of tests defined in (9). For every function $f \in F$, let $\mathcal{B}_{K,\lambda}(f) = \{g \in F : T_{K,\lambda}(g, f) \geq 0\}$ denote the set of all functions $g \in F$ that beat f. As explained in Section 3, these sets are measured with two metrics. First, let

$$R_{K,\lambda}^{\text{reg}}(f) = \sup_{g\in\mathcal{B}_{K,\lambda}(f)} \{\|g - f\|\} \text{ and } \hat{f}_{K,\lambda}^{(1)} \in \arg\min_{f\in F} R_{K,\lambda}^{reg}(f) \ .$$

Next, let

$$R_{K,\lambda}^{(2)}(f) = \sup_{g\in\mathcal{B}_{K,\lambda}(f)} \{\text{MOM}_K [|g - f|]\} .$$

The second criterion is given by

$$C_{K,\lambda}^{(2)}(f) = \inf \left\{ \rho \geq 0 : R_{K,\lambda}^{\text{reg}}(f) \leq \rho \text{ and } R_{K,\lambda}^{(2)}(f) \leq 85\theta_r r(\rho) \right\} \ ,$$

where $r(\cdot)$ is a link function as defined in Definition 1. The associated estimator is then given by

$$\hat{f}_{K,\lambda}^{(2)} \in \operatorname*{argmin}_{f\in F} C_{K,\lambda}^{(2)}(f) \ .$$

For the definition of $\hat{f}_{K,\lambda}^{(1)}$ and $\hat{f}_{K,\lambda}^{(2)}$, if the argmin does not exist, one can choose an approximate $1/n$-minimizer without altering statistical performance of these estimators.

4.5. The sparsity equation

From the quadratic / multiplier decomposition of the excess quadratic loss:

$$T_{K,\lambda}(f^*, f) = \mathrm{MOM}_K[(f - f^*)^2 - 2\zeta(f - f^*)] + \lambda(\|f\| - \|f^*\|) \ . \tag{11}$$

Let $f \in F$ and $\rho = \|f - f^*\|$. When ρ is large and $\|f - f^*\|_{L_P^2}$ is small, one can prove $T_{K,\lambda}(f^*, f) > 0$ only thanks to the regularization term $\lambda(\|f\| - \|f^*\|)$ in (11). This is why a lower bound on the regularization term is usefull.

Recall that the subdifferential of $\|\cdot\|$ in $f \in F$ is the set

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

where $(E^*, \|\cdot\|^*)$ is the dual normed space of $(E, \|\cdot\|)$ (and E is the linear space containing F onto which $\|\cdot\|$ is defined). For all $\rho > 0$, let H_ρ denote the set

$$H_\rho = \{f \in F : \|f - f^*\| = \rho, \ \|f - f^*\|_{L_P^2} \leq r(\rho)\}$$

where $r(\cdot)$ is the *link function* from Definition 1. Let $\Gamma_{f^*}(\rho)$ denote the union of all subdifferentials of $\|\cdot\|$ at functions "close" to f^*

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

Intuitively, every norm promotes the "sparsity" defined by large (in the dual sphere) subdifferentials of this norm. Sparse functions f^{**} are useful in our context because a large lower bound on $\|f\| - \|f^{**}\|$ (and so for $\|f\| - \|f^*\|$ when $\|f^{**} - f^*\|$ is small enough) can be derived when the vector $f - f^{**}$ is in the right direction. This intuition is formalized in the sparsity equation. Let

$$\forall \rho > 0, \qquad \Delta(\rho) = \inf_{f \in H_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(f - f^*) \ .$$

By definition, $\Delta(\rho) - \rho/20$ is a uniform lower bound on $\|f\| - \|f^*\|$ if $f \in H_\rho$. Thus, $\|f\| - \|f^*\| \gtrsim \rho$, when $\Delta(\rho) \gtrsim \rho$ which is the *sparsity equation* as introduced in Lecué and Mendelson (2016a).

Definition 2. *A radius $\rho > 0$ satisfies the **sparsity equation** if $\Delta(\rho) \geq 4\rho/5$.*

If ρ^* satisfies the sparsity equation, so do all $\rho \geq \rho^*$. Therefore, one can define

$$\rho^* = \inf\left(\rho > 0 : \Delta(\rho) \geq \frac{4\rho}{5}\right). \tag{12}$$

5. Main results

5.1. Basic risk bounds

Theorem 1 gathers estimation error bounds satisfied by the estimators $\hat{f}_{K,\lambda}^{(j)}$ for $j = 1, 2$ defined in Section 4.4.

Theorem 1. *Grant Assumptions 1, 2 and 3 and let r_Q, r_M anr r denote the functions introduced in Definition 1 for*

$$\gamma_Q = \min\left(\frac{1}{661\theta_0}, \frac{1}{1764\theta_r}\right), \gamma_M = \frac{\epsilon}{168} \text{ and } \epsilon = \frac{3}{331\theta_0^2}.$$

Let ρ^* be defined in (12) and let K^* denote the smallest integer such that

$$K^* \geq \max\left(\frac{8|\mathcal{O}|}{7}, \frac{N\epsilon^2 r^2(\rho^*)}{336\theta_m^2}\right) .$$

For all $K \geq 1$, let ρ_K be a solution of $r^2(\rho_K) = [336\theta_m^2/\epsilon^2]\sqrt{K/N}$. Assume that for every $i \in \mathcal{I}$, $K \in [K^*, N]$ and $f \in F \cap B(f^*, \rho_K)$,

$$2(P_i - P)\zeta(f - f^*) \leq \epsilon \max\left(\frac{336\theta_m^2}{\epsilon^2}\frac{K}{N}, r_M^2(\rho_K, \gamma_M), \|f - f^*\|_{L_P^2}^2\right) . \quad (13)$$

For all $K \in [K^*, N/(84\theta_r^2\theta_0^2)]$, with probability larger than $1 - 4\exp(-K/1008)$, one has

$$\left\|\hat{f}_{K,\lambda}^{(1)} - f^*\right\| \leq \rho_K ,$$

and

$$\left\|\hat{f}_{K,\lambda}^{(2)} - f^*\right\| \leq \rho_K, \quad \left\|\hat{f}_{K,\lambda}^{(2)} - f^*\right\|_{L_P^2} \leq 340\theta_0\theta_r r(\rho_K)$$

when the regularization parameter satisfies

$$\frac{20\epsilon}{7}\frac{r^2(\rho_K)}{\rho_K} < \lambda < \frac{10}{331\theta_0^2}\frac{r^2(\rho_K)}{\rho_K} .$$

To the best of our knowledge, Theorem 1 provides the first theoretical guarantees for any estimator in this setting. Recall that the dataset may be corrupted by adversarial outliers, that informative data may be heavy-tailed and that their distribution P_i for $i \in \mathcal{I}$ only induce L^2 and L^1 geometries over $F - f^*$ equivalent to that of P. In particular, the distributions P_i may be different from P. When $K = K^*$, the rates in Theorem 1, for L_2 estimation (to the square) of f^* is $r^2(\rho_{K^*})$ which is $\sim \max(|\mathcal{O}|/N, r^2(\rho^*))$, as announced in the abstract. When the number of outliers $|\mathcal{O}|$ is less than $Nr^2(\rho^*)$, this rate is the minimax rate of the RERM in i.i.d. subgaussian frameworks with independent noise Lecué and Mendelson (2013).

In Theorem 1, K can be as small as the infimum between the number of outliers and N times the minimax rate of convergence. Hence, if the optimal rate is known, as in Lugosi and Mendelson (2017), Theorem 1 shows that Le Cam's estimators with $K = K^*$ achieve the same performance as the champions in this paper.

Assumption 1 and (13) are automatically satisfied when for all $i \in \mathcal{I}, P_i = P$. Theorem 1 goes beyond the i.i.d. setup, relaxing the i.d. assumptions into proximity assumptions between $L_{P_i}^2$ and L_P^2 geometries.

5.2. Adaptive choice of K by Lepski's method

The main drawback of Theorem 1 is that optimal rates are only achieved when $K \approx K^*$. Since K^* is unknown, it cannot be used in general.

Let $K_1 = K^*$ and $K_2 = N/(84\theta_0^2\theta_r^2)$ be defined as in Theorem 1. For any integer $K \in [K_1, K_2]$, let ρ_K and λ be defined as in Theorem 1 and for $j = 1, 2$ denote by $\hat{f}_K^{(j)} = \hat{f}_{K,\lambda}^{(j)}$ for this choice of λ. For all $f \in F$, let

$$\hat{B}_K^{(2)}(f) = \left\{ g \in F : \mathrm{MOM}_K\left[|g - f|\right] \leq 28900\theta_r^2\theta_0 r(\rho_K) \right\} \ .$$

Now, let

$$R_K^{(1)} = B(\hat{f}_K^{(1)}, \rho_K), \quad R_K^{(2)} = B(\hat{f}_K^{(2)}, \rho_K) \cap \hat{B}_K^{(2)}(\hat{f}_K^{(2)})$$

and for every $j = 1, 2$, let

$$\hat{K}^{(j)} = \inf\left\{ K \in [K_2] : \bigcap_{J=K}^{K_2} R_J^{(j)} \neq \emptyset \right\} \ .$$

Finally, define adaptive (to K) estimators via Lepski's method: for $j = 1, 2$, $\hat{f}_{LE}^{(j)} \in \bigcap_{J=\hat{K}^{(j)}}^{K_2} R_J^{(j)}$.

Theorem 2. *Grant assumptions and notations of Theorem 1. There exist absolute constants $(c_i)_{1\leq i\leq 2}$ such that the estimators $\hat{f}_{LE}^{(j)}$ for $j = 1, 2$ satisfy for every $K \in [K^*, N/(84\theta_0^2\theta_r^2)]$, with probability at least $1 - c_1\exp(-c_2 K)$,*

$$\left\| \hat{f}_{LE}^{(1)} - f^* \right\| \leq 2\rho_K \ ,$$

and

$$\left\| \hat{f}_{LE}^{(2)} - f^* \right\| \leq 2\rho_K, \quad \left\| \hat{f}_{LE}^{(2)} - f^* \right\|_{L_P^2} \leq 680\theta_r\theta_0 r(2\rho_K) \ .$$

In particular, for $K = K^*$, if the following regularity assumption holds: there exists an absolute constant c_3 such that for all $\rho > 0$, $r(2\rho) \leq c_3 r(\rho)$, with probability at least

$$1 - c_1\exp\left(-c_4 N \max\left(\frac{|\mathcal{O}|}{N}, \frac{r^2(\rho^*)}{\theta_0^4\theta_m^2}\right)\right)$$

then,

$$\left\| \hat{f}_{LE}^{(2)} - f^* \right\|_{L_P^2} \leq c_5 \max\left(\theta_0^4\theta_m^2\frac{|\mathcal{O}|}{N}, r^2(\rho^*)\right) \ .$$

Assume that all $(X_i, Y_i), i \in [N]$ are distributed according to (X, Y^{f^*}), where $f^* \in F$, $Y^{f^*} = f^*(X) + \zeta$ and ζ is a centered Gaussian variable with variance σ, independent of X. Assume that F is L-subgaussian: for every $f \in F$ and $p \geq 2$, $\|f\|_{L^p} \leq L\sqrt{p}\,\|f\|_{L^2}$. Then, (Lecué and Mendelson, 2013, Theorem A′) proves that if $\tilde{f}_N$ is an estimator such that for every $f^* \in F$ and every $r > 0$,

with probability at least $1 - c_0 \exp(-\sigma^{-1} r^2 N / c_0)$, $\left\| \tilde{f}_N - f^* \right\|_{L_P^2} \leq \zeta_N$, then necessarily

$$\zeta_N \gtrsim \min\left(r, \operatorname{diam}(F, L_P^2)\right). \tag{14}$$

When $Y^{f^*} = f^*(X) + \zeta$, $c \sim 1/\theta_m \sim 1/\sigma$. Applying this result to $r = r(\rho_K)$ for some given $K \geq K^*$ shows no procedure can estimate f^* in L_P^2 uniformly over F with confidence at least $1 - c_0 \exp(-K/c_0)$ at a rate better than $r(\rho_K)$ (we implicitly assumed that $r(\rho_K) \leq \operatorname{diam}(F, L_P^2)$ since $r(\rho_K)$ can obviously be replaced by $r(\rho_K) \wedge \operatorname{diam}(F, L_P^2)$ in all results). Moreover, this rate is minimax since (Lecué and Mendelson, 2013, Theorem A) also shows that the ERM over $\rho_K B$, $\hat{f}_N^{ERM} \in \operatorname{argmin}_{f \in \rho_K B} P_N \ell_f$, satisfies $\left\| \hat{f}_N^{ERM} - f^* \right\|_{L^2} \lesssim r(\rho_K)$ with probability at least $1 - c_0 \exp(-\sigma^{-1} r^2(\rho_K) N / c_0)$ when $\sigma \gtrsim r_Q(\rho_K)$.

Theorem 2 shows that $\hat{f}_{LE}$ achieves similar rates of convergence with the same exponentially high confidence in the relaxed setting of this paper. Compared to Lugosi and Mendelson (2017), the Lepski method chooses automatically the tuning parameter K, which yields to exact minimax rates without knowledge of this rate for the construction of the estimators and for a possibly corrupted database.

5.3. Application to MOM Lasso

As a proof of concept, theoretical properties are illustrated in the example of sparse-recovery in high-dimensional spaces using ℓ_1-regularization, cf. Bühlmann and van de Geer (2011); Giraud (2015). The interested reader can check that it also applies to other procedures like Slope (cf. Bogdan et al. (2015); Su and Candès (2015)) and trace-norm regularization as well as kernel methods, for instance, by using the results in Lecué and Mendelson (2016a,b).

Recall this classical setup. For every $t = (t_j)_1^d \in \mathbb{R}^d$ and $1 \leq p \leq +\infty$, let

$$F = \{\langle \cdot, t \rangle : t \in \mathbb{R}^d\} \quad \text{and} \quad \left\| \langle \cdot, t \rangle \right\| = \|t\|_1, \text{ where } \|t\|_p = \left(\sum_{j=1}^d |t_j|^p \right)^{1/p}.$$

Let t^* be defined by $f^* = \langle \cdot, t^* \rangle$. Let $(e_1, \ldots, e_d)$ be the canonical basis of $\mathbb{R}^d$ and let B_p^d (resp. S_p^{d-1}) denote the unit ball (resp. sphere) associated to $\|\cdot\|_p$. To simplify presentation of this example, assume that $P = P_i$ for all $i \in \mathcal{I}$ and write L^q for L_P^q, that is $(X_i, Y_i)_{i \in \mathcal{I}}$ are i.i.d. $\sim P$. The following result is then a corollary of Theorem 2.

Theorem 3. *Assume that* $\mathbb{E}\langle X, t \rangle^2 = \|t\|_2^2$ *for all* $t \in \mathbb{R}^d$ *and*

- *o) there exist* $s \in [N]$ *such that* $N \geq c_1 s \log(ed/s)$ *and* $v \in \mathbb{R}^d$ *such that* $\|t^* - v\|_1 \leq \sigma s \sqrt{\log(ed/s)/N}/20$ *and* $|\operatorname{supp}(v)| \leq s$.
- *i')* $|\mathcal{I}| \geq N/2$ *and* $|\mathcal{O}| \leq c_1 s \log(ed/s)$,
- *ii)* $\zeta = Y - \langle X, t^* \rangle \in L_{q_0}$ *for some* $q_0 > 2$
- *iii')* *for every* $1 \leq p \leq c_0 \log(ed)$, $\left\| \langle X, e_j \rangle \right\|_{L_p} \leq L\sqrt{p}$,

iv') *there exists* θ_0 *such that* $\left\|\langle X,t\rangle\right\|_{L^1} \leq \theta_0 \|t\|_2$, *for all* $t \in \mathbb{R}^d$,

v) *there exists* θ_m *such that* $\mathrm{var}(\zeta\langle X,t\rangle) \leq \theta_m^2 \|t\|_2^2$, *for all* $t \in \mathbb{R}^d$.

The MOM-LASSO estimator $\hat{t}_{LE}$ *defined by* $\hat{f}_{LE} = \langle \hat{t}_{LE}, \cdot \rangle$ *satisfies, with probability at least* $1 - c_2 \exp(-c_3 s \log(ed/s))$, *for every* $1 \leq p \leq 2$,

$$\left\|\hat{t}_{LE} - t^*\right\|_p \leq c_4(L, \theta_m) \|\zeta\|_{L_{q_0}} s^{1/p} \sqrt{\frac{1}{N} \log\left(\frac{ed}{s}\right)} \ .$$

Theoretical properties of MOM LASSO outperform those of LASSO, cf. for example [Theorem 1.4 in Lecué and Mendelson (2016a)], in several ways.

- Estimation rates achieved by MOM-LASSO are the actual minimax rates $s\log(ed/s)/N$, see Bellec et al. (2016), while classical LASSO estimators achieve the rate $s\log(ed)/N$. This improvement is possible thanks to the adaptation step in MOM-LASSO.
- the probability deviation for LASSO is polynomial – $1/N^{(q_0/2-1)}$ – it is exponentially small for MOM LASSO. Exponential rates for LASSO hold only if the noise ζ is subgaussian ($\|\zeta\|_{L_p} \leq C\sqrt{p}\|\zeta\|_{L_2}$ for all $p \geq 2$).
- MOM LASSO is insensitive to data corruption by up to s times $\log(ed/s)$ outliers while only one outlier can break performance of LASSO.
- All assumptions on X are weaker for MOM LASSO than for LASSO.

6. Proofs

In all the proof section, $\mathbb{P}$ denotes the distribution of $(X_1, \ldots, X_N)$ and $\mathbb{E}$ the corresponding expectation. For any non-empty subset $B \subset [N]$ and any $f : \mathcal{X} \to \mathbb{R}$ for which it makes sense, let $\overline{P}_B f = \frac{1}{|B|}\sum_{i\in B} P_i f$. For any $f \in L_P^2$ and $r \geq 0$, let

$$B_2(f,r) = \{g \in L_P^2 : \|f-g\|_{L_P^2} \leq r\}, \quad S_2(f,r) = \{g \in L_P^2 : \|f-g\|_{L_P^2} = r\} \ .$$

Let $\mathcal{K}$ denote the set of indices of blocks B_k containing only informative data:

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

The are mainly three stochastic results necessary to analyze regularized Le Cam's test estimators. The first and the second bound the quadratic and multiplier "MOM processes" and the third one is an isometric result. Similar ingredient were used to analyze ERM in Lecué and Mendelson (2013), see Lemma 2.6 and its following remark and Equation (2.4). Proofs of these stochastic ingredients are based on the small ball method of Koltchinskii and Mendelson (2015). In parallel to our work, similar results appeared in Lemmas 5.1 and 5.5 in Lugosi and Mendelson (2017) in the i.i.d. setup for the study of tournament estimators, see also Mendelson (2017).

6.1. Lower Bound on the quadratic MOM process

Lemma 1. *Grant Assumptions 1 and 3. Fix* $\eta \in (0,1)$, $\rho > 0$ *and let* $\alpha, \gamma_Q, \gamma, x \in (0,1)$ *be such that* $\gamma\left(1-\alpha-x-32\theta_0\gamma_Q\right) \geq 1-\eta$. *Let* $K \in [|\mathcal{O}|/(1-\gamma), N\alpha/(2\theta_0\theta_r)^2]$.

There exists an event $\Omega_Q(K,\rho)$ *such that* $\mathbb{P}\left(\Omega_Q(K,\rho)\right) \geq 1-\exp(-K\gamma x^2/2)$ *on which for all* $f \in B(f^*,\rho)$ *if* $\|f-f^*\|_{L_P^2} \geq r_Q(\rho,\gamma_Q)$ *then*

$$Q_{\eta,K}(|f-f^*|) \geq \frac{1}{4\theta_0}\|f-f^*\|_{L_P^2} \text{ and } Q_{\eta,K}((f-f^*)^2) \geq \frac{1}{(4\theta_0)^2}\|f-f^*\|_{L_P^2}^2 \ .$$

Proof. For all $f \in F - \{f^*\}$, let $n_f = (f-f^*)/\|f-f^*\|_{L_P^2}$. For $i \in \mathcal{I}$, $P_i|n_f| \geq \theta_0^{-1}$ by Assumption 3 and $P_i n_f^2 \leq \theta_r^2$ by Assumption 1. By Markov's inequality, for all $k \in \mathcal{K}$,

$$\mathbb{P}\left(|P_{B_k}|n_f| - \overline{P}_{B_k}|n_f|| > \frac{\theta_r}{\sqrt{\alpha|B_k|}}\right) \leq \alpha$$

and so

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

Since $K \leq N\alpha/(2\theta_0\theta_r)^2$ then $|B_k| = N/K \geq \alpha/(2\theta_0\theta_r)^2$ and thus

$$\mathbb{P}\left(P_{B_k}|n_f| \geq \frac{1}{2\theta_0}\right) \geq 1-\alpha \ . \tag{15}$$

Let ϕ denote the function defined by $\phi(t) = (t-1)I(1 \leq t \leq 2) + I(t \geq 2)$ for all $t \in \mathbb{R}_+$ and, for all $f \in F - \{f^*\}$, let $Z(f) = \sum_{k\in[K]} I(4\theta_0 P_{B_k}|n_f| \geq 1)$. Since $I(t \geq 1) \geq \phi(t)$ for any $t \geq 0$ then $Z(f) \geq \sum_{k\in\mathcal{K}} \phi\left(4\theta_0 P_{B_k}|n_f|\right)$. Since $\phi(t) \geq I(t \geq 2)$ for all $t \geq 0$, it follows from (15) that

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

Therefore, for all $f \in F$,

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

Let $\mathcal{F} = \{f \in B(f^*,\rho) : \|f-f^*\|_{L_P^2} \geq r_Q(\rho,\gamma_Q)\}$. By the bounded difference inequality (cf. (McDiarmid, 1989, Lemma 1.2) or (Boucheron et al., 2013, Theorem 6.2), there exists an event $\Omega(x)$ such that $\mathbb{P}(\Omega(x)) \geq 1-\exp(-x^2|\mathcal{K}|/2)$, on which

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

By the Giné-Zynn symmetrization argument (Boucheron et al., 2013, Lemma 11.4),

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

where $(\epsilon_k)_{k\in\mathcal{K}}$ are independent Rademacher variables independent of the data. Moreover, ϕ is 1-Lipschitz and $\phi(0)=0$. By the contraction principle (cf. Equation (4.20) in (Ledoux and Talagrand, 1991, Theorem 4.12))

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\epsilon_k\phi\left(4\theta_0 P_{B_k}|n_f|\right)\right| \leq 8\theta_0\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\epsilon_k P_{B_k}|n_f|\right| \quad .$$

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

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

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

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

Therefore

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{k\in\mathcal{K}}\epsilon_k P_{B_k}|n_f|\right| \leq \frac{4K}{N}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in\cup_{k\in\mathcal{K}}B_k}\epsilon_i n_f(X_i)\right| \quad .$$

It follows from the convexity of F that for all $f\in\mathcal{F}$, $r_Q(\rho,\gamma_Q)n_f\in F-f^*$ and it also belongs to the L_P^2 sphere of radius $r_Q(\rho,\gamma_Q)$. Therefore, by definition of $r_Q:=r_Q(\rho,\gamma_Q)$ and for $J=\cup_{k\in\mathcal{K}}B_k$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i\in J}\epsilon_i n_f(X_i)\right| = \frac{1}{r_Q}\mathbb{E}\sup_{f\in F\cap S_2(f^*,r_Q)}\left|\sum_{i\in J}\epsilon_i(f-f^*)(X_i)\right| \leq \gamma_Q\frac{|\mathcal{K}|N}{K} \quad .$$

In conclusion, on $\Omega(x)$, all $f \in \mathcal{F}$ are such that

$$Z(f) \geq |\mathcal{K}|\,(1-\alpha-x-32\theta_0\gamma_Q) \geq (1-\eta)K \ .$$

In other words, on $\Omega(x)$, for all $f \in \mathcal{F}$, there exist at least $(1-\eta)K$ blocks B_k such that $P_{B_k}|n_f| \geq (4\theta_0)^{-1}$. For any of these blocks B_k, $P_{B_k} n_f^2 \geq (P_{B_k}|n_f|)^2$, hence, on $\Omega(x)$, $Q_{\eta,K}[|n_f|] \geq (4\theta_0)^{-1}$ and $Q_{\eta,K}[n_f^2] \geq (4\theta_0)^{-2}$. ∎

6.2. Upper Bound on the multiplier MOM process

Lemma 2. *Grant Assumption 2. Fix $\eta \in (0,1)$, $\rho \in (0,+\infty]$, and let $\alpha, \gamma_M, \gamma, x$ and ϵ be positive absolute constants such that $\gamma\,(1-\alpha-x-8\gamma_M/\epsilon) \geq 1-\eta$. Let $K \in [|\mathcal{O}|/(1-\gamma), N]$. There exists an event $\Omega_M(K,\rho)$ such that $\mathbb{P}(\Omega_M(K,\rho)) \geq 1-\exp(-\gamma K x^2/2)$ and on $\Omega_M(K,\rho)$, for all $f \in B(f^*,\rho)$ there is at least $(1-\eta)K$ blocks B_k with $k \in \mathcal{K}$ such that*

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

Proof. For all $k \in [K]$ and $f \in F$, define $W_k(f) = 2(P_{B_k}-\overline{P}_{B_k})\,(\zeta(f-f^*))$ and

$$\gamma_k(f) = \epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho,\gamma_M), \|f-f^*\|_{L_P^2}^2\right) \ .$$

Let $f \in F$ and $k \in \mathcal{K}$. It follows from Markov's inequality and Assumption 2 that

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

Denote $J = \cup_{k\in\mathcal{K}} B_k$ and remark that $J \in \mathcal{J}$ as defined in Definition 1. Let $r_M := r_M(\rho,\gamma_M)$ for simplicity. We have

$$\begin{aligned}
\mathbb{E}\sup_{f\in B(f^*,\rho)}\sum_{k\in\mathcal{K}}\epsilon_k\frac{W_k(f)}{\gamma_k(f)} &\leq 2\mathbb{E}\sup_{f\in B(f^*,\rho)}\left|\sum_{k\in\mathcal{K}}\frac{\epsilon_k(P_{B_k}-\overline{P}_{B_k})(\zeta(f-f^*))}{\epsilon\max(r_M^2,\|f-f^*\|_{L_P^2}^2)}\right| \\
&\leq \frac{2}{\epsilon r_M^2}\mathbb{E}\left[\sup_{f\in B(f^*,\rho)\setminus B_2(f^*,r_M)}\left|\sum_{k\in\mathcal{K}}\epsilon_k(P_{B_k}-\overline{P}_{B_k})\left(\zeta r_M\frac{f-f^*}{\|f-f^*\|_{L_P^2}}\right)\right|\right. \\
&\qquad\left. \vee \sup_{f\in B(f^*,\rho)\cap B_2(f^*,r_M)}\left|\sum_{k\in\mathcal{K}}\epsilon_k(P_{B_k}-\overline{P}_{B_k})\,(\zeta(f-f^*))\right|\right] \\
&\leq \frac{2}{\epsilon r_M^2}\mathbb{E}\sup_{f\in B(f^*,\rho)\cap B_2(f^*,r_M)}\left|\sum_{k\in\mathcal{K}}\epsilon_k(P_{B_k}-\overline{P}_{B_k})\,(\zeta(f-f^*))\right| \ ,
\end{aligned}$$

where in the last but one inequality we used that F is convex and the same argument as in the proof of Lemma 1. Moreover, since the random variables $((\zeta_i(f-f^*)(X_i) - P_i\zeta(f-f^*)) : i \in \mathcal{I})$ are centered and independent, the symmetrization argument applies and, by definition of r_M,

$$\mathbb{E}\sup_{f\in B(f^*,\rho)}\sum_{k\in\mathcal{K}}\epsilon_k\frac{W_k(f)}{\gamma_k(f)} \le \frac{4K}{\epsilon r_M^2 N}\mathbb{E}\sup_{f\in B(f^*,\rho)\cap B_2(f^*,r_M)}\left|\sum_{i\in J}\epsilon_i\zeta_i(f-f^*)(X_i)\right|$$
$$\le \frac{4K}{\epsilon N}\gamma_M|\mathcal{K}|\frac{N}{K} = \frac{4\gamma_M}{\epsilon}|\mathcal{K}| \ . \tag{17}$$

Now, let $\psi(t) = (2t-1)I(1/2 \le t \le 1) + I(t \ge 1)$ for all $t \ge 0$ and note that ψ is 2-Lipschitz, $\psi(0) = 0$ and satisfies $I(t \ge 1) \le \psi(t) \le I(t \ge 1/2)$ for all $t \ge 0$. Therefore, all $f \in B(f^*, \rho)$ satisfies

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

where we used (16) in the last inequality. The bounded difference inequality ensures that, for all $x > 0$, there exists an event $\Omega(x)$ satisfying $\mathbb{P}(\Omega(x)) \ge 1 - \exp(-x^2|\mathcal{K}|/2)$ on which

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

Furthermore, it follows from the symmetrization argument that

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

and, from the contraction principle and (17), that

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

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

$$\begin{aligned}\sum_{k\in\mathcal{K}}I\left(|W_k(f)|<\gamma_k(f)\right)&\geq\left(1-\alpha-x-8\gamma_M/\epsilon\right)|\mathcal{K}|\\&\geq K\gamma\left(1-\alpha-x-8\gamma_M/\epsilon\right)\geq(1-\eta)K\ .\end{aligned}$$

■

6.3. An isometry property of $MOM_K[\cdot]$ processes

Besides the controls of the quadratic and multiplier MOM processes presented in Lemmas 1 and 2 respectively, the estimation error bounds for the MOM estimators rely on the following isometry property of the MOM processus $f\in F\to \mathrm{MOM}_K[|f-f^*|]$.

Lemma 3. *[Isometry property of the $MOM_K[\cdot]$ process] Grant Assumptions 1 and 3. Fix $\eta\in(0,1)$, $\rho>0$ and let α,γ_Q,γ,x denote absolute constants in $(0,1)$ such that $\gamma\left(1-\alpha-x-4\theta_r\gamma_Q/\alpha\right)\geq 1-\eta$. Let $K\in[|\mathcal{O}|/(1-\gamma),N\alpha/(2\theta_0\theta_r)^2]$. There exists an event $\Omega_{iso}(K,\rho)\subset\Omega_Q(K,\rho)$ such that $\mathbb{P}(\Omega_{iso}(K,\rho))\geq 1-2\exp\left(-\gamma x^2K/2\right)$ and on the event $\Omega_{iso}(K,\rho)$, for all $f\in B(f^*,\rho)$,*

$$Q_{1-\eta,K}|f-f^*|\leq\theta_r\|f-f^*\|_{L_P^2}+\frac{4\theta_r}{\alpha}\max\left(r_Q(\rho,\gamma_Q),\|f-f^*\|_{L_P^2}\right)$$

and if $\|f-f^*\|_{L_P^2}\geq r_Q(\rho,\gamma_Q)$ then $Q_{\eta,K}|f-f^*|\geq(1/(4\theta_0))\|f-f^*\|_{L_P^2}$.

In particular, for $\eta=1/2$, on the event $\Omega_{iso}(K,\rho)$, for all $f\in B(f^*,\rho)$, if $\|f-f^*\|_{L_P^2}\geq r_Q(\rho,\gamma_Q)$, then

$$\frac{1}{4\theta_0}\|f-f^*\|_{L_P^2}\leq MOM_K[|f-f^*|]\leq\theta_r\left(1+\frac{4}{\alpha}\right)\|f-f^*\|_{L_P^2}\,.\tag{18}$$

Proof. It follows from Lemma 1 that on the event $\Omega_Q(K,\rho)$ for all $f\in B(f^*,\rho)$, if $\|f-f^*\|_{L_P^2}\geq r_Q(\rho,\gamma_Q)$ then $Q_{\eta,K}|f-f^*|\geq(1/(4\theta_0)\|f-f^*\|_{L_P^2}$. This yields the "lower bound" result in (18).

For the upper bound of the isomorphic result, we essentially repeat the proof of Lemma 2. Let us just highlight the main differences. We will use the same notation as in the proof of Lemma 2 except that for all $f\in F$, we define

$$W_k(f)=(P_{B_k}-\overline{P}_{B_k})|f-f^*|\text{ and }\gamma_k(f)=\frac{4\theta_r}{\alpha}\max\left(r_Q(\rho,\gamma_Q),\|f-f^*\|_{L_P^2}\right).$$

It follows from Chebyshev's inequality and Assumption 1 that

$$\mathbb{P}\left[2|W_k(f)|\geq\gamma_k(f)\right]\leq\frac{4\overline{P}_{B_k}|f-f^*|}{\gamma_k(f)}\leq\frac{4\theta_r\|f-f^*\|_{L_P^2}}{\gamma_k(f)}\leq\alpha.$$

Moreover, by convexity of F, we have, for $r_Q := r_Q(\rho, \gamma_Q)$,

$$(\star) := \mathbb{E} \sup_{f \in B(f^*,\rho)} \sum_{k \in \mathcal{K}} \epsilon_k \frac{W_k(f)}{\gamma_k(f)}$$
$$\leq \frac{4\theta_r}{\alpha r_Q} \mathbb{E} \sup_{f \in B(f^*,\rho) \cap S_2(f^*, r_Q)} \left| \sum_{k \in \mathcal{K}} \epsilon_k (P_{B_k} - \overline{P}_{B_k}) |f - f^*| \right|$$

and then using a symmetrization argument, we obtain that

$$(\star) \leq \frac{4\theta_r K}{\alpha r_Q N} \mathbb{E} \sup_{f \in B(f^*,\rho) \cap S_2(f^*, r_Q)} \left| \sum_{i \in J} \epsilon_i (f - f^*)(X_i) \right| \leq \frac{4\theta_r \gamma_Q |\mathcal{K}|}{\alpha}.$$

Finally, using the same argument as in the proof of Lemma 2, for all $x > 0$ there exists an event $\Omega(x)$ such that $\mathbb{P}(\Omega(x)) \geq 1 - \exp(-x^2|\mathcal{K}|/2)$, on which for all $f \in B(f^*, \rho)$,

$$\sum_{k \in \mathcal{K}} I(|W_k(f)| \leq \gamma_k(f)) \geq |\mathcal{K}|(1 - \alpha - x - 4\theta_r \gamma_Q/\alpha) \geq (1 - \eta)|\mathcal{K}|.$$

In particular, on the event $\Omega(x)$, for all $f \in B(f^*, \rho)$ there are more than $(1 - \eta)K$ blocks B_k for which, $P_{B_k}|f - f^*| \leq \overline{P}_{B_k}|f - f^*| + \gamma_k(f)$. Now, the result follows from Assumption 1 since $\overline{P}_{B_k}|f - f^*| \leq \theta_r \|f - f^*\|_{L_P^2}$. ∎

6.4. Conclusion to the proof of Theorem 1

The proof relies on the following proposition.

Proposition 1. *Grant conditions of Theorem 1. Let $\gamma_Q = 1/(661\theta_0)$, $\gamma_M = \epsilon/168$ for some $\epsilon < 7/(662\theta_0^2)$ and the regularization parameter be such that*

$$\frac{20\epsilon r^2(\rho_K)}{7\rho_K} < \lambda < \frac{10 r^2(\rho_K)}{331\theta_0^2 \rho_K}.$$

The event $\Omega_0(K) = \Omega_Q(K, \rho_K) \cap \Omega_M(K, \rho_K)$ is such that $\mathbb{P}(\Omega_0(K)) \geq 1 - 2\exp(-K/1008)$ and on $\Omega_0(K)$ for all $f \in F$ if $\|f - f^*\|_{L_P^2} \geq r(\rho_K)$ or $\|f - f^*\| \geq \rho_K$ then

$$MOM_K[\ell_f - \ell_{f^*}] + \lambda(\|f\| - \|f^*\|) > 0 \ .$$

Proof. Using (6), (7) and (8) together with the quadratic / multiplier decomposition of the excess quadratic loss yields that for all $f \in F$,

$$\begin{aligned} \mathrm{MOM}_K[\ell_f - \ell_{f^*}] &= \mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] \\ &\geq Q_{1/4,K}[(f - f^*)^2] - 2Q_{3/4,K}[\zeta(f - f^*)] \ . \end{aligned} \tag{19}$$

Note that $\gamma(1 - \alpha - x - 32\theta_0\gamma_Q) \geq 1 - \eta$ when one chooses

$$\eta = \frac{1}{4}, \gamma = \frac{7}{8}, \alpha = \frac{1}{21}, x = \frac{1}{21}, \gamma_Q = \frac{1}{661\theta_0}, \gamma_M = \frac{\epsilon}{168} \text{ and } \epsilon \leq \frac{1}{64\theta_0^2}. \tag{20}$$

For this choice of constants, Lemma 1 applies and for $\rho = \rho_K$ we get that there exists an event $\Omega_Q(K, \rho_K)$ with probability larger than $1 - \exp(-K/1008)$ and on that event, for all $f \in B(f^*, \rho_K)$, if $\|f - f^*\|_{L_P^2} \geq r_Q(\rho_K, \gamma_Q)$ then

$$Q_{1/4,K}[(f - f^*)^2] \geq \frac{1}{(4\theta_0)^2} \|f - f^*\|_{L_P^2}^2 \quad . \tag{21}$$

Moreover, for the choice of parameters as in (20), we also have $\gamma(1 - \alpha - x - 8\gamma_M/\epsilon) \geq 1 - \eta$, hence Lemma 2 applies and for $\rho = \rho_K$ we get that there exists an event $\Omega_M(K, \rho_K)$ with probability larger than $1 - \exp(-K/1008)$ and on that event, for all $f \in B(f^*, \rho_K)$ there are more than $3K/4$ blocks B_k with $k \in \mathcal{K}$ such that

$$|2(P_{B_k} - \overline{P}_{B_k})\zeta(f - f^*)| \leq \epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho_K, \gamma_M), \|f - f^*\|_{L_P^2}^2\right) \quad .$$

Combining the last result with Assumption (13) together with the fact that $P\zeta(f - f^*) \leq 0$ for all $f \in F$ (because of the convexity of F), it follows that on the event $\Omega_M(K, \rho_K)$, for all $f \in B(f^*, \rho_K)$,

$$2Q_{3/4,K}[\zeta(f - f^*)] \leq 2\epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho_K, \gamma_M), \|f - f^*\|_{L_P^2}^2\right) \quad . \tag{22}$$

Let us now prove that on the event $\Omega_M(K, \rho_K) \cap \Omega_Q(K, \rho_K)$, one has for all $f \in B(f^*, \rho_K)$,

$$\mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] \geq -2\epsilon r^2(\rho_K) \quad . \tag{23}$$

Assume that $\Omega_M(K, \rho_K) \cap \Omega_Q(K, \rho_K)$ holds and let $f \in B(f^*, \rho_K)$. First assume that $\|f - f^*\|_{L_P^2} \geq r^2(\rho_K)$. Then, it follows from (19), (21) and (22), the choice of ϵ in (20) and the definition of ρ_K that

$$\mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] \geq \left(\frac{1}{(4\theta_0)^2} - 2\epsilon\right)\|f - f^*\|_{L_P^2}^2 \geq \frac{\|f - f^*\|_{L_P^2}^2}{32\theta_0^2}. \tag{24}$$

Now, if $\|f - f^*\|_{L_P^2} \leq r^2(\rho_K)$ then it follows from (19), (22) and the definition of ρ_K that

$$\mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] \geq -2\epsilon r^2(\rho_K)$$

and (23) follows.

Conclusion of the proof when the regularization distance is small (i.e. $\|f - f^*\| \leq \rho_K$) and the L_P^2-distance is large (i.e. $\|f - f^*\|_{L_P^2} \geq r(\rho_K)$). Let $f \in F$ be such that $\|f - f^*\| \leq \rho_K$ and $\|f - f^*\|_{L_P^2} \geq r(\rho_K)$. It follows from the triangular inequality that $\|f\| - \|f^*\| \geq -\|f - f^*\| \geq -\rho_K$. Combining this together with (24), it follows that

$$\mathrm{MOM}_K\left[\ell_f - \ell_{f^*}\right] + \lambda(\|f\| - \|f^*\|) \geq \frac{\|f - f^*\|_{L_P^2}^2}{32\theta_0^2} - \lambda\rho_K \geq \frac{r^2(\rho_K)}{32\theta_0^2} - \lambda\rho_K > 0$$

when $\lambda < r^2(\rho_K)/(32\theta_0^2\rho_K)$.

Conclusion of the proof when the regularization distance is large (i.e. $\|f-f^*\| \geq \rho_K$): the homogeneity argument.

Lemma 4. *For all $f \in F$ such that $\|f - f^*\| \geq \rho_K$*

$$\|f\| - \|f^*\| \geq \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) - \frac{\rho_K}{10} \ .$$

Proof. For every $f^{**} \in f^* + (\rho_K/20)B$ and every $z^* \in (\partial \|\cdot\|)_{f^{**}}$,

$$\begin{aligned}\|f\| - \|f^*\| &\geq \|f\| - \|f^{**}\| - \|f^{**} - f^*\| \geq z^*(f - f^{**}) - \frac{\rho_K}{20} \\ &= z^*(f - f^*) - z^*(f^{**} - f^*) - \frac{\rho_K}{20} \geq z^*(f - f^*) - \frac{\rho_K}{10} \ .\end{aligned}$$

■

Lemma 5. *Assume that, for all $f \in F \cap S(f^*, \rho_K)$,*

$$MOM_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) > \lambda \frac{\rho_K}{10} \ . \tag{25}$$

Then (25) holds for all $f \in F$ such that $\|f - f^*\| \geq \rho_K$.

Proof. Let $f \in F$ be such that $\|f - f^*\| \geq \rho_K$. Define $g = f^* + \rho_K \frac{f - f^*}{\|f - f^*\|}$ and remark that $\|g - f^*\|_{L_P^2} = \rho_K$ and that, by convexity of F, $g \in F$. It follows from (25) that for $\kappa = \|f - f^*\|/\rho_K \geq 1$, one has

$$\begin{aligned}&\mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \\ &= \mathrm{MOM}_K\left[\kappa^2(g - f^*)^2 - 2\kappa\zeta(g - f^*)\right] + \lambda\kappa \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g - f^*) \\ &\geq \kappa \left(\mathrm{MOM}_K\left[(g - f^*)^2 - 2\zeta(g - f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g - f^*) \right) \\ &> \kappa\lambda\frac{\rho_K}{10} \geq \lambda\frac{\rho_K}{10} \ .\end{aligned}$$

■

Let $f \in F$ be such that $\|f - f^*\| \geq \rho_K$. By Lemma 4,

$$\begin{aligned}&\mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] + \lambda(\|f\| - \|f^*\|) \\ &\geq \mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) - \lambda\frac{\rho_K}{10} \ .\end{aligned}$$

Therefore, it will follow from Lemma 5 that

$$\mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] + \lambda(\|f\| - \|f^*\|) > 0$$

if we can prove that for all $g \in F$ such that $\|g - f^*\| = \rho_K$ one has

$$\mathrm{MOM}_K\left[(g-f^*)^2 - 2\zeta(g-f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) > \lambda \frac{\rho_K}{10} \ . \tag{26}$$

Let us now prove that (26) holds. Let $g \in F$ be such that $\|g - f^*\| = \rho_K$. First assume that $\|g - f^*\|_{L_P^2} \leq r(\rho_K)$ so that $g \in H_{\rho_K}$. By definition $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g - f^*) \geq \Delta(\rho_K)$ and, since $\rho_K \geq \rho^*$, ρ_K satisfies the sparsity equation and thus, $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g - f^*) \geq 4\rho_K/5$. Therefore, thanks to (23), when $\lambda > 20\epsilon r^2(\rho_K)/(7\rho_K)$, one has

$$\begin{aligned}\mathrm{MOM}_K\left[(g-f^*)^2 - 2\zeta(g-f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) \\ \geq -2\epsilon r^2(\rho_K) + \lambda \frac{4}{5}\rho_K > \lambda \frac{\rho_K}{10} \ .\end{aligned}$$

Finally assume that $\|g - f^*\|_{L_P^2} \geq r(\rho_K)$. Since $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \geq -\|f - f^*\| = -\rho_K$, it follows from (24) that

$$\begin{aligned}&\mathrm{MOM}_K\left[(g-f^*)^2 - 2\zeta(g-f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) \\ &\geq \frac{1}{32\theta_0^2}\|g - f^*\|_{L_P^2}^2 - \lambda\rho_K \geq \frac{r^2(\rho_K)}{32\theta_0^2} - \lambda\rho_K > \lambda\frac{\rho_K}{10}\end{aligned}$$

when $\lambda < 10r^2(\rho_K)/(331\theta_0^2\rho_K)$. ∎

End of the proof of Theorem 1. On the event $\Omega_0(K)$ of Proposition 1, $\mathcal{B}_{K,\lambda}(f^*)$ is included in the ball $B(f^*, \rho_K)$, therefore, by definition of $\hat{f}^{(1)}_{K,\lambda}$ (cf. (4)),

$$\left\|\hat{f}^{(1)}_{K,\lambda} - f^*\right\| \leq C^{(1)}_{K,\lambda}(f^*) \leq \rho_K \ .$$

Again, by Proposition 1, on the same event $\Omega_0(K)$, $\mathcal{B}_{K,\lambda}(f^*) \subset B(f^*, \rho_K) \cap B_2(f^*, r(\rho_K))$, hence, on $\Omega_0(K) \cap \Omega_{iso}(K)$, where $\Omega_{iso}(K)$ is an event defined in Lemma 3, for all $f \in \mathcal{B}_{K,\lambda}(f^*)$,

$$\mathrm{MOM}_K\left[|f - f^*|\right] \leq 85\theta_r \|f - f^*\|_{L_P^2} \leq 85\theta_r r(\rho_K)$$

where $\alpha = 1/21$ according to (20). Therefore, $C^{(2)}_{K,\lambda}(f^*) \leq \rho_K$, which implies that $\left\|\hat{f}^{(2)}_{K,\lambda} - f^*\right\| \leq \rho_K$ (cf. (4)) and that $C^{(2)}_{K,\lambda}(\hat{f}^{(2)}_{K,\lambda}) \leq \rho_K$ and therefore, by Lemma 3, on $\Omega_0(K) \cap \Omega_{iso}(K)$, either $\left\|\hat{f}^{(2)}_{K,\lambda} - f^*\right\|_{L_P^2} \leq r_Q(\rho_K, \gamma_K)$ and so $\left\|\hat{f}^{(2)}_{K,\lambda} - f^*\right\|_{L_P^2} \leq 340\theta_0\theta_r r(\rho_K)$ or $\left\|\hat{f}^{(2)}_{K,\lambda} - f^*\right\|_{L_P^2} \geq r_Q(\rho_K, \gamma_K)$ and so

$$\left\|\hat{f}^{(2)}_{K,\lambda} - f^*\right\|_{L_P^2} \leq 4\theta_0 \mathrm{MOM}_K\left[|\hat{f}^{(2)}_{K,\lambda} - f^*|\right] \leq 340\theta_0\theta_r r(\rho_K) \ .$$

6.5. Conclusion to the proof of Theorem 2

First, it follows from Theorem 1 that for all $K \in [K_1, K_2]$, with probability at least $1 - c_0 \exp(-c_1 K)$, for both $j = 1, 2$, $f^* \in \cap_{J=K}^{K_2} R_K^{(j)}$, so $\widehat{K}^{(j)} \leq K$, which implies that both f^* and $\widehat{f}_{\mathrm{LE}}^{(j)}$ belong to $B(\widehat{f}_{K,\lambda}^{(j)}, \rho_K)$, therefore, $\left\| f^* - \widehat{f}_{\mathrm{LE}}^{(j)} \right\| \leq 2\rho_K$. Second, for the L_P^2-estimation error bound of $\widehat{f}_{\mathrm{LE}}^{(2)}$, denote by $r_J = 340\theta_r\theta_0 r(\rho_J)$ the bound on the L_P^2 risk of the estimator $\hat{f}_J^{(2)}$ obtained in Theorem 1. Let $K \in [K_1, K_2]$. It follows from Lemma 3 for $\rho = 2\rho_J, J \geq K$ that there exists absolute constants c_1, c_2 and an event Ω_{iso} such that $\mathbb{P}(\Omega_{iso}) \geq 1 - c_1 \exp(-c_2 K)$ and, on the event Ω_{iso}, for all $J \geq K$, $\eta \in \{1/4, 1/2, 3/4\}$ and $f \in B(f^*, 2\rho_J)$,

$$\text{if } \|f - f^*\|_{L_P^2} \geq r_Q(2\rho_J, \gamma_Q), \qquad Q_{\eta,J}(|f - f^*|) \begin{cases} \geq \frac{1}{4\theta_0} \|f - f^*\|_{L_P^2} \\ \leq 85\theta_r \|f - f^*\|_{L_P^2} \end{cases} .$$

Let Ω be the event defined as the following intersection:

$$\Omega = \bigcap_{J=K}^{K_2} \left\{ \left\| \widehat{f}_J^{(2)} - f^* \right\| \leq \rho_J \text{ and } \left\| \widehat{f}_J^{(2)} - f^* \right\|_{L_P^2} \leq r_J \right\} \bigcap \Omega(K) \bigcap \Omega_{iso} .$$

It follows from Theorem 1 that $\mathbb{P}(\Omega) \geq 1 - c_3 \exp(-c_4 K)$. Moreover, on Ω, for all $J \geq K$,

$$Q_{3/4,J}\left(|f^* - \widehat{f}_J^{(2)}| \right) \leq 85\theta_r r_J .$$

So, in particular, $f^* \in \cap_{J=K}^{K_2} \left\{ f \in B(\widehat{f}_J^{(2)}, \rho_J) : \mathrm{MOM}_J \left[|f - \widehat{f}_J^{(2)}| \right] \leq 85\theta_r r_J \right\}$. By definition of $\hat{K}^{(2)}$, this implies that $\hat{K}^{(2)} \leq K$ on Ω. Therefore, on Ω,

$$\hat{f}_{LE}^{(2)} \in \cap_{J=K}^{K_2} \left\{ f \in B(f^*, 2\rho_J) : \mathrm{MOM}_J \left[|f - \widehat{f}_J^{(2)}| \right] \leq 85\theta_r r_J \right\} .$$

In particular,

$$\mathrm{MOM}_K \left[|\hat{f}_{LE}^{(2)} - \widehat{f}_K^{(2)}| \right] \leq 85\theta_r r_K .$$

Now on Ω_{iso}, one has for all $f \in B(f^*, 2\rho_K)$, if $\|f - f^*\|_{L_P^2} \geq r_Q(2\rho_K, \gamma_Q)$ then

$$Q_{1/4,J}[|f - f^*|] \geq \frac{1}{4\theta_0} \|f - f^*\|_{L_P^2} .$$

Therefore on Ω_{iso}, one has either $\left\| \hat{f}_{LE}^{(2)} - f^* \right\|_{L_P^2} \leq r_Q(2\rho_K, \gamma_Q)$ or $\left\| \hat{f}_{LE}^{(2)} - f^* \right\|_{L_P^2} \geq r_Q(2\rho_K, \gamma_Q)$ and in the latter case,

$$\begin{aligned} \left\| \hat{f}_{LE}^{(2)} - f^* \right\|_{L_P^2} &\leq 4\theta_0 Q_{1/4,K}[|\hat{f}_{LE}^{(2)} - f^*|] \\ &\leq 4\theta_0 \left(\mathrm{MOM}_K \left[|\hat{f}_{LE}^{(2)} - \widehat{f}_K^{(2)}| \right] + Q_{3/4,K}(|\widehat{f}_K^{(2)} - f^*|) \right) \\ &\leq 680\theta_0\theta_r r_K . \end{aligned}$$

■

6.6. Proof of Theorem 3

Gaussian mean widths of any $V \subset \mathbb{R}^d$ are defined by

$$\ell^*(V) = \mathbb{E}\left\{ \sup_{(v_j)\in V} \sum_{j=1}^d g_j v_j \right\}, \quad \text{where} \quad (g_1,\dots,g_d) \sim \mathcal{N}_d(0, I_d) \ . \tag{27}$$

The dual norm of the ℓ_1^d-norm is 1-unconditional with respect to the canonical basis of $\mathbb{R}^d$ (Mendelson, 2016, Definition 1.4). Therefore, (Mendelson, 2016, Theorem 1.6) applies, for every $\rho > 0$,

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

Local Gaussian mean widths $\ell^*(\rho B_1^d \cap r B_2^d)$ are bounded from above in (Lecué and Mendelson, 2016a, Lemma 5.3) and computations of r_M and r_Q follow

$$r_M^2(\rho) \lesssim_{L,q_0,\gamma_M} \begin{cases} \sigma^2 \frac{d}{N} & \text{if } \rho^2 N \ge \sigma^2 d^2 \\ \rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)} & \text{otherwise} \end{cases} ,$$
$$r_Q^2(\rho) \begin{cases} = 0 & \text{if } N \gtrsim_{L,\gamma_Q} d \\ \lesssim_{L,\gamma_Q} \frac{\rho^2}{N}\log\left(\frac{c(L,\gamma_Q)d}{N}\right) & \text{otherwise} \end{cases} .$$

Therefore, a link function is explicitly given by

$$r^2(\rho) \sim_{L,q_0,\gamma_Q,\gamma_M} \begin{cases} \max\left(\rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\sigma^2 d}{N}\right) & \text{if } N \gtrsim_L d \\ \max\left(\rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\rho^2}{N}\log\left(\frac{d}{N}\right)\right) & \text{otherwise} \end{cases} . \tag{28}$$

The sparsity equation has been solved in this example in (Lecué and Mendelson, 2016a, Lemma 4.2), recall this result.

Lemma 6. *If there exists $v \in \mathbb{R}^d$ such that $v \in t^* + (\rho/20)B_1^d$ and $|\mathrm{supp}(v)| \le c\rho^2/r^2(\rho)$ then*

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

Compute finally ρ_K and $\lambda \sim r^2(\rho_K)/\rho_K$. The equation $K = cr(\rho_K)^2 N$ is solved by

$$\rho_K \sim_{L,q_0} \frac{K}{\sigma}\sqrt{\frac{1}{N}\log^{-1}\left(\frac{\sigma^2 d}{K}\right)} \tag{29}$$

for the $r(\cdot)$ function defined in (28). Therefore,

$$\lambda \sim \frac{r^2(\rho_K)}{\rho_K} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho_K\sqrt{N}}\right)} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma^2 d}{K}\right)} \ . \tag{30}$$

Conclusion of the proof. It follows from Theorem 2, the computation of $r(\rho_K)$ from (28) and ρ_K in (29) that with probability at least $1-c_0\exp(-cr(\rho_K)^2N/\overline{C})$, $\left\|\hat{t}_{LE}-t^*\right\|_1 \leq \rho_{K^*}$ and $\left\|\hat{t}_{LE}-t^*\right\|_2 \lesssim r(\rho_K)$. The result follows since $\rho_{K^*} \sim \rho^* \sim_{L,q_0} \sigma s\sqrt{\frac{1}{N}\log\left(\frac{ed}{s}\right)}$ and $\|v\|_p \leq \|v\|_1^{-1+2/p}\|v\|_2^{2-2/p}$ for all $v \in \mathbb{R}^d$ and $1 \leq p \leq 2$.

Alon, N., Matias, Y., Szegedy, M., 1999. The space complexity of approximating the frequency moments. J. Comput. System Sci. 58 (1, part 2), 137–147, twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).
URL http://dx.doi.org/10.1006/jcss.1997.1545

Audibert, J.-Y., Catoni, O., 2011. Robust linear least squares regression. Ann. Statist. 39 (5), 2766–2794.
URL http://dx.doi.org/10.1214/11-AOS918

Baraud, Y., 2011. Estimator selection with respect to Hellinger-type risks. Probab. Theory Related Fields 151 (1-2), 353–401.
URL http://dx.doi.org/10.1007/s00440-010-0302-y

Baraud, Y., Birgé, L., 2009. Estimating the intensity of a random measure by histogram type estimators. Probab. Theory Related Fields 143 (1-2), 239–284.
URL http://dx.doi.org/10.1007/s00440-007-0126-6

Baraud, Y., Birgé, L., 2016. Rho-estimators for shape restricted density estimation. Stochastic Process. Appl. 126 (12), 3888–3912.
URL http://dx.doi.org/10.1016/j.spa.2016.04.013

Baraud, Y., Birgé, L., Sart, M., 2017. A new method for estimation and model selection : ρ-estimation. Invent. Math. 207 (2), 435–517.

Baraud, Y., Giraud, C., Huet, S., 2014. Estimator selection in the Gaussian setting. Ann. Inst. Henri Poincaré Probab. Stat. 50 (3), 1092–1119.
URL http://dx.doi.org/10.1214/13-AIHP539

Bellec, P., Lecué, G., Tsybakov, A., 2016. Slope meets lasso: Improved oracle bounds and optimality. Tech. rep., CREST, CNRS, Université Paris Saclay.

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

Birgé, L., 2006. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Ann. Inst. H. Poincaré Probab. Statist. 42 (3), 273–325.
URL http://dx.doi.org/10.1016/j.anihpb.2005.04.004

Birgé, L., 2013. Robust tests for model selection. In: From probability to statistics and back: high-dimensional models and processes. Vol. 9 of Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, pp. 47–64.
URL http://dx.doi.org/10.1214/12-IMSCOLL905

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., Candès, E. J., 2015. SLOPE—Adaptive variable selection via convex optimization. Ann. Appl. Stat. 9 (3), 1103–1140.

Boucheron, S., Lugosi, G., Massart, P., 2013. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, iSBN 978-0-19-953525-5.

Brownlees, C., Joly, E., Lugosi, G., 2015. Empirical risk minimization for heavy-tailed losses. Ann. Statist. 43 (6), 2507–2536.
URL http://dx.doi.org/10.1214/15-AOS1350

Bühlmann, P., van de Geer, S., 2011. Statistics for high-dimensional data. Springer Series in Statistics. Springer, Heidelberg, methods, theory and applications.
URL http://dx.doi.org/10.1007/978-3-642-20192-9

Catoni, O., 2012. Challenging the empirical mean and empirical variance: a deviation study. Ann. Inst. Henri Poincaré Probab. Stat. 48 (4), 1148–1185.
URL http://dx.doi.org/10.1214/11-AIHP454

Chichignoud, M., Lederer, J., 2014. A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression. Bernoulli 20 (3), 1560–1599.
URL http://dx.doi.org/10.3150/13-BEJ533

de la Peña, V. H., Giné, E., 1999. Decoupling. Probability and its Applications (New York). Springer-Verlag, New York.
URL http://dx.doi.org/10.1007/978-1-4612-0537-1

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., 2016. Sub-Gaussian mean estimators. Ann. Statist. 44 (6), 2695–2725.
URL http://dx.doi.org/10.1214/16-AOS1440

Fan, J., Li, Q., Wang, Y., 2017. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J. R. Stat. Soc. Ser. B. Stat. Methodol. 79 (1), 247–265.
URL http://dx.doi.org/10.1111/rssb.12166

Giraud, C., 2015. Introduction to high-dimensional statistics. Vol. 139 of Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL.

Huber, P. J., 1964. Robust estimation of a location parameter. Ann. Math. Statist. 35, 73–101.

Huber, P. J., Ronchetti, E. M., 2009. Robust Statistics. Wiley.

Jerrum, M. R., Valiant, L. G., Vazirani, V. V., 1986. Random generation of combinatorial structures from a uniform distribution. Theoret. Comput. Sci. 43 (2-3), 169–188.
URL http://dx.doi.org/10.1016/0304-3975(86)90174-X

Koltchinskii, V., Mendelson, S., 2015. Bounding the smallest singular value of a random matrix without concentration. Int. Math. Res. Not. IMRN (23), 12991–13008.
URL http://dx.doi.org/10.1093/imrn/rnv096

Le Cam, L., 1973. Convergence of estimates under dimensionality restrictions. Ann. Statist. 1, 38–53.
URL http://links.jstor.org/sici?sici=0090-5364(197301)1:1<38:COEUDR>2.0.CO;2-V&origin=MSN

Le Cam, L., 1986. Asymptotic methods in statistical decision theory. Springer Series in Statistics. Springer-Verlag, New York.
URL http://dx.doi.org/10.1007/978-1-4612-4946-7

Lecué, G., Mendelson, S., 2013. Learning subgaussian classes: Upper and minimax bounds. Tech. rep., CNRS, Ecole polytechnique and Technion.

Lecué, G., Mendelson, S., 2014. Sparse recovery under weak moment assumptions. Tech. rep., CNRS, Ecole Polytechnique and Technion, to appear in Journal of the European Mathematical Society.

Lecué, G., Mendelson, S., 2016a. Regularization and the small-ball method i: sparse recovery. Tech. rep., CNRS, ENSAE and Technion, I.I.T.

Lecué, G., Mendelson, S., 2016b. Regularization and the small-ball method ii: complexity dependent error rates. Tech. rep., CNRS, ENSAE and Technion, I.I.T.

Ledoux, M., Talagrand, M., 1991. Probability in Banach spaces. Vol. 23 of Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Springer-Verlag, Berlin, isoperimetry and processes.

Lepski, O. V., 1991. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. Teor. Veroyatnost. i Primenen. 36 (4), 645–659.
URL http://dx.doi.org/10.1137/1136085

Lugosi, G., Mendelson, S., 2017. Risk minimization by median-of-means tournaments. Preprint available on ArXive:1608.00757.

McDiarmid, C., 1989. On the method of bounded differences. In: Surveys in combinatorics, 1989 (Norwich, 1989). Vol. 141 of London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, pp. 148–188.

Mendelson, S., 2014a. Learning without concentration. In: Proceedings of the 27th annual conference on Learning Theory COLT14. pp. pp 25–39.

Mendelson, S., 2014b. A remark on the diameter of random sections of convex bodies. In: Geometric aspects of functional analysis. Vol. 2116 of Lecture Notes in Math. Springer, Cham, pp. 395–404.

Mendelson, S., 2015a. Learning without concentration. J. ACM 62 (3), Art. 21, 25.
URL http://dx.doi.org/10.1145/2699439

Mendelson, S., 2015b. Learning without concentration for a general loss function. Tech. rep., Technion and ANU, Canberra.

Mendelson, S., 2016. On multiplier processes under weak moment assumptions. Tech. rep., Technion.

Mendelson, S., 2017. On aggregation for heavy-tailed classes. Probab. Theory Related Fields 168 (3-4), 641–674.
URL https://doi.org/10.1007/s00440-016-0720-6

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

Rudelson, M., Vershynin, R., 2014. Small ball probabilities for linear images of high dimensional distributions. Tech. rep., University of Michigan, international Mathematics Research Notices, to appear. [arXiv:1402.4492].

Saba, L., Hoffman, P. L., Hornbaker, C., Bhave, S. V., Tabakoff, B., 2008. Expression quantitative trait loci and the phenogen database 31 (3).

Sart, M., 2014. Estimation of the transition density of a Markov chain. Ann. Inst. Henri Poincaré Probab. Stat. 50 (3), 1028–1068.
URL http://dx.doi.org/10.1214/13-AIHP551

Su, W., Candès, E. J., 2015. Slope is adaptive to unknown sparsity and asymptotically minimax. Tech. rep., Stanford University, to appear in The Annals of Statistics.

van der Vaart, A. W., Wellner, J. A., 1996. Weak convergence and empirical processes, with applications to statistics. Springer Series in Statistics. Springer-Verlag, New York.

Vapnik, V. N., 1998. Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, a Wiley-Interscience Publication.

Vapnik, V. N., Chervonenkis, A. Y., 1974. Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya. Izdat. "Nauka", Moscow.

Appendix .1. Discussion of the main assumptions

Let us give some examples where Assumption 2 holds. If the noise random variable $\zeta(Y,X)$ (resp. $\zeta(Y_i,X_i)$ for $i\in\mathcal{I}$) has a variance conditionally to X (resp. X_i for $i\in\mathcal{I}$) that is uniformly bounded then Assumption 2 holds. This is the case when $\zeta(Y,X)$ (resp. $\zeta(Y_i,X_i)$ for $i\in\mathcal{I}$) is independent of X (resp. X_i for $i\in\mathcal{I}$) and has finite L^2-moment with $\theta_m = \max_{Q\in P,\{P_i\}_{i\in\mathcal{I}}}\|\zeta\|_{L^2_Q}$. It also holds without independence under higher moment conditions. For example, assume $\sigma = \max_{Q\in P,\{P_i\}_{i\in\mathcal{I}}}\|\zeta\|_{L^4_Q} < \infty$ and, for every $f\in F$, $\|f-f^*\|_{L^4_Q}\le\theta_1\|f-f^*\|_{L^2_P}$ then by Cauchy-Schwarz inequality, $\sqrt{\mathrm{var}_Q(\zeta(f-f^*))}\le\|\zeta(f-f^*)\|_{L^2_Q}\le\|\zeta\|_{L^4_Q}\|f-f^*\|_{L^4_Q}\le\theta_1\sigma\|f-f^*\|_{L^2_P}$ and so Assumption 2 holds for $\theta_m=\theta_1\sigma$.

By Cauchy-Schwarz inequality, $\|f-f^*\|_{L^1_{P_i}}\le\|f-f^*\|_{L^2_{P_i}}$ for all $f\in F$ and $i\in\mathcal{I}$. Therefore, Assumptions 1 and 3 together imply that all norms $L^2_P, L^2_{P_i}, L^1_{P_i}, i\in\mathcal{I}$ are equivalent over $F-f^*$. Note also that Assumption 3 is related to the small ball property (cf. Koltchinskii and Mendelson (2015); Mendelson (2014a)) as shown by Proposition 2 bellow. The small ball property has been recently used in Learning theory and signal processing. We refer to Koltchinskii and Mendelson (2015); Lecué and Mendelson (2014); Mendelson (2015b, 2014b, 2015a); Rudelson and Vershynin (2014) for examples of distributions satisfying this assumption.

Proposition 2. *Let Z be a real-valued random variable.*

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

Proof. If $\mathbb{P}(|Z|\ge\kappa_0\|Z\|_2)\ge u_0$ then

$$\|Z\|_1\ge\int_{|z|\ge\kappa_0\|Z\|_2}|z|P_Z(dz)\ge u_0\kappa_0\|Z\|_2 \quad ,$$

where P_Z denotes the distribution of Z. Conversely, if $\|Z\|_2\le\theta_0\|Z\|_1$, the Paley-Zigmund's argument (de la Peña and Giné, 1999, Proposition 3.3.1) shows that, if $p=\mathbb{P}(|Z|\ge\kappa_0\|Z\|_2)$,

$$\begin{aligned}\|Z\|_2\le\theta_0\|Z\|_1 &= \theta_0\left(\mathbb{E}[|Z|I(|Z|\le\kappa_0\|Z\|_2)]+\mathbb{E}[|Z|I(|Z|\ge\kappa_0\|Z\|_2)]\right)\\ &\le\theta_0\|Z\|_2(\kappa_0+\sqrt{p}) \quad .\end{aligned}$$

As one can assume that $\|Z\|_2\neq0$, $p\ge(\theta_0^{-1}-\kappa_0)^2$. ∎

Appendix .2. Examples of Le Cam's tests estimators

Le Cam's approach has been used by Birgé to define T-estimators (cf. Baraud and Birgé (2009); Birgé (2006, 2013)) and by Baraud, Birgé and Sart to define ρ-estimators (cf. Baraud and Birgé (2016); Baraud et al. (2017)). Baraud (2011); Baraud et al. (2014) also built efficient estimator selection procedures with this approach. It also extends many common procedures in statistical learning theory, as shown by the following examples.

Example 1 : Empirical minimizers. Assume $T_N(g,f) = \ell_N(f) - \ell_N(g)$ for some random function $\ell_N : F \to \mathbb{R}$ and denote by $\hat{f} = \arg\min_{f\in F} \ell_N(f)$ a minimizer of the corresponding criterion (provided that it exists and is unique). Then it is easy to check that $\mathcal{B}_{T_N}(\hat{f}) = \{\hat{f}\}$, so its radius is null, while the radius of any other point f is larger than $d(f,\hat{f}) > 0$ (whatever the non-degenerate notion of pseudo-distance used for d). It follows that $\hat{f}$ is the estimator (3). In particular, any possibly penalized empirical risk minimizer

$$\hat{f} = \arg\min_{f\in F}\{P_N\ell_f + \mathrm{reg}(f)\}$$

is obtained by Le Cam's construction with the tests

$$T_N(g,f) = P_N(\ell_f - \ell_g) + \mathrm{reg}(f) - \mathrm{reg}(g) \ .$$

These examples encompass classical empirical risk minimizers of Vapnik (1998) but also their robust versions from Huber (1964); Audibert and Catoni (2011).

Example 2 : median-of-means estimators. Another, perhaps less obvious example is the median-of-means estimator Alon et al. (1999); Jerrum et al. (1986); Nemirovsky and Yudin (1983) of the expectation PZ of a real valued random variable Z. Let $Z_1,\dots,Z_N$ denote a sample and let $B_1,\dots,B_K$ denote a partition of $[N]$ into bins of equal size N/K. The estimator $\mathrm{MOM}_K(Z)$ is the (empirical) median of the vector of empirical means $\left(P_{B_k}Z = |B_k|^{-1}\sum_{i\in B_k} Z_i\right)_{k\in[K]}$. Recall that

$$PZ = \operatorname*{argmin}_{m\in\mathbb{R}} P(Z-m)^2 = \operatorname*{argmin}_{m\in\mathbb{R}} \max_{m'\in\mathbb{R}} P[(Z-m)^2 - (Z-m')^2] \ .$$

Define the MOM test statistic to compare any $m, m' \in \mathbb{R}$ by

$$T_N(m,m') = \mathrm{MOM}_K[(Z-m')^2 - (Z-m)^2] \ .$$

Basic properties of the median (recalled in Eq (6) and (7) of Section 4.1) yield

$$\begin{aligned} T_N(m,m') &= (m')^2 - m^2 + \mathrm{MOM}_K[-2Z(m'-m)] \\ &= (m')^2 - 2m'\mathrm{MOM}_K(Z) - [m^2 - 2m\mathrm{MOM}_K(Z)] \\ &= (m' - \mathrm{MOM}_K(Z))^2 - (m - \mathrm{MOM}_K(Z))^2 \ . \end{aligned}$$

Defining $\ell_N(m) = (m - \mathrm{MOM}_K(Z))^2$, one has

$$T_N(m,m') = \ell_N(m') - \ell_N(m) \ .$$

As in the previous example, Le Cam's estimator based on T_N is therefore the unique minimizer of ℓ_N, that is $\mathrm{MOM}_K(Z)$.

Example 3 : "Champions" of a Tournament. In a related but different approach, Lugosi and Mendelson (2017) introduced median-of-means tournaments, which are also based on median-of-means tests to compare elements in F.