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

Guillaume Lecué[a], Matthieu Lerasle[b]

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

Abstract

We obtain estimation error rates for estimators obtained by aggregation of regularized median-of-means tests, following a construction of Le Cam. The results hold with exponentially large probability, under only weak moments assumptions on data. Any norm may be used for regularization. When it has some sparsity inducing power we recover sparse rates of convergence. The procedure is robust since a large part of data may be corrupted, these outliers have nothing to do with the oracle we want to reconstruct. Our general risk bound is of order

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

In particular, the number of outliers may be as large as *(number of data)* $\times$*(minimax rate)* without affecting this rate. The other data do not have to be identically distributed but should only have equivalent L^1 and L^2 moments. For example, the minimax rate $s\log(ed/s)/N$ of recovery of a s-sparse vector in $\mathbb{R}^d$ is achieved with exponentially large probability by a median-of-means version of the LASSO when the noise has q_0 moments for some $q_0 > 2$, the entries of the design matrix should have $C_0 \log(ed)$ moments and the dataset can be corrupted 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 $(X_i, Y_i)_{i=1,\dots,N}$. The labels Y and Y_i's are real-valued while the inputs X and X_i's take values in an abstract measurable space $\mathcal{X}$.

Empirical Risk Minimizers (ERM) of Vapnik (1998); Vapnik and Chervonenkis (1974) and later on, their regularized versions replace the unknown distribution P in the definition of f^* by the empirical distribution P_N based on

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

the sample $(X_i, Y_i)_{i=1,\dots,N}$. Given a function reg : $F \to \mathbb{R}_+$, this produces regularized ERM defined by

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

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). These issues are critical in many modern applications such as high-frequency trading, where heavy-tailed data are quite common or in various areas of biology such as micro-array analysis or neuroscience where data are sometimes still nasty after being preprocessed. To overcome the problem, various methods have been proposed. The most common strategy is to replace the square-loss function to make it less sensitive to outliers. For example, 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 (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 performances 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". Median-of-means estimators of Alon et al. (1999); Jerrum et al. (1986); Nemirovsky and Yudin (1983) compare elements of F. A "champion" is an element $\hat{f}$ such that $\mathcal{B}_{T_N}(\hat{f})$ is smaller than a computable upper bound on the radius of $\mathcal{B}_{T_N}(f^*)$. They prove that the risk of any champion is controlled by this upper bound. An important message of this paper is that Le Cam's estimators are quite common in statistics, in particular in robust statistics. For example, Section 3 shows that any penalized empirical loss function can be obtained by Le Cam's approach and that Le Cam's estimators based on median-of-means tests are champions of median-of-means tournaments.

This paper studies estimators derived from Le Cam's procedure based on regularized median-of-means (MOM) tests (see Section 4.1). Our estimators are therefore particular instances of champions of MOM's tournaments and another motivation is to push further the analysis of this particular champion. 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, it also allows to tackle the problem of "aggressive" outliers.

The closest work is certainly that of Lugosi and Mendelson (2017), but we believe that our paper contains substantial improvements. We stress the intimate relationship between their estimator and Le Cam general construction and use this parallel to propose a much simpler estimator. Our risk bounds are always better and we extend their results to possibly corrupted data-sets.

To investigate robustness properties of median-of-means estimators, we partition the dataset into two parts. One is made of outliers data. They are indexed by $\mathcal{O} \subset [N]$ of cardinality $|\mathcal{O}| = K_o$. **On those data, absolutely nothing is assumed** : they may not be independent, have distributions P_i totally different from P, with no moment at all, etc.. These are typically data polluting datasets like in the case of declarative data on internet or when something went wrong during the storage, compression or transfer which resulted in complete non sense data. 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). Many other examples of datasets containing outliers could be provided, this includes frauds detection and terrorist activity as examples. Of course, outliers are not flagged in advance and the statistician is given no a priori information on which data is an outlier or not. The other part of the dataset is made of data on which the MOM estimator rely on to estimate the oracle f^*. There should be enough information in those data so that the estimation of f^* is possible, even in the presence of outliers provided they remain in a "decent proportion". We therefore call the non-outliers, the *informative data*, those that bring information on f^*. We denote by $\mathcal{I} \subset [N]$ the set indexing these data. We therefore end up with a partition of $[N]$ as $[N] = \mathcal{I} \cup \mathcal{O}$ which, again, is not known from the statistician.

The radii of the sets $\mathcal{B}_{T_N}(f)$ are computed for regularization and L^2_P norms. The regularization norm is chosen in advance by the statistician to promote sparsity or smoothness. It can be used freely in our procedure, but it doesn't ensure a small L^2_P risk for the estimator. The L^2_P-norm is unknown in general since it depends on the distribution of X. Furthermore, the classical $L^2_{P_N}$-empirical metric fails to estimate the L^2_P metric without subgaussian properties of the design vector X. Fortunately, it can be replaced by a median-of-means metric. To handle simultaneously both regularization and L^2_P norms, we will also slightly extend Le Cam's principle. Our first important result shows that the resulting estimator is well localized w.r.t. both regularization and L^2_P norms.

Median-of-means estimators rely on a data splitting into K blocks and this parameter drives the resulting statistical performances (cf. Devroye et al. (2016)). To achieve optimal rates, K should be ultimately chosen using parameters that depend on the oracle f^* like its sparsity which is not in general available to the statistician. 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.

There are four important features in our approach. First, all results are proved under weak $L^{2+\epsilon}$ moment assumptions on the noise. This is an almost minimal condition for the problem to make sense. The class F is only assumed to satisfy a weak "L_2/L_1" comparison. Second, performances 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 well as the ERM when the 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.

Example: sparse-recovery via MOM LASSO. As a proof of concept, theoretical properties are illustrated in the classical example of sparse-recovery in high-dimensional spaces using the ℓ_1-regularization. This example illustrates typical results that follow from our analysis in one of the most classical problem of high dimensional statistics (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. Let X denote a random vector in $\mathbb{R}^d$ such that $\mathbb{E}\big\langle X, t\big\rangle^2 = \|t\|_2^2$ for all $t \in \mathbb{R}^d$ (X is isotropic) and let Y be a real-valued random vector. Let $t^* \in \operatorname{argmin}_{t\in\mathbb{R}^d} \mathbb{E}(Y - \big\langle X, t\big\rangle)^2$. Let $(X_i, Y_i)_{i\in[N]}$ denote independent data corrupted by outliers : no assumption is made on a subset $(X_i, Y_i)_{i\in\mathcal{O}}$ of the dataset. Let $\mathcal{I} = [N]\backslash\mathcal{O}$ denote the indices of *informative data* $(X_i, Y_i)_{i\in\mathcal{I}}$: for all $i \in \mathcal{I}$, (X_i, Y_i) are independent with the same distribution (X, Y). For the sake of simplicity, we only consider the case of i.i.d. informative data in this example. In high-dimensional statistics, $N \leq d$ but t^* has only s ($s < N$) non-zero coordinates. To estimate t^*, the ℓ_1-norm $\|\cdot\|_1$ is used for penalization to promote zero coordinates. The following result holds.

Theorem 1. *[Theorem 1.4 in Lecué and Mendelson (2016a)] Assume t^* is s-sparse, $N \geq c_0 s\log(ed/s)$, X is isotropic and*

- *i) $|\mathcal{I}| = N$ and $|\mathcal{O}| = 0$ (no outliers in the dataset),*
- *ii) $\zeta = Y - \big\langle X, t^*\big\rangle \in L_{q_0}$ for some $q_0 > 2$*
- *iii) there exists $L > 0$ such that for all $t \in \mathbb{R}^d$ and all $p \geq 2$, $\big\|\big\langle X, t\big\rangle\big\|_{L_p} \leq L\sqrt{p}\,\big\|\big\langle X, t\big\rangle\big\|_{L_2}$*
- *iv) there exist $u_0 > 0$ and $\beta_0 > 0$ such that for all $t \in \mathbb{R}^d$,*

$$\mathbb{P}\left[|\big\langle X, t\big\rangle| \geq u_0 \big\|\big\langle X, t\big\rangle\big\|_{L_2}\right] \geq \beta_0 \ .$$

The LASSO estimator, defined by

$$\tilde{t} \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \left(\frac{1}{N} \sum_{i=1}^{N} \big(Y_i - \langle X_i, t\rangle\big)^2 + c_1 \left\| \zeta \right\|_{L_{q_0}} \sqrt{\frac{\log(ed)}{N}} \left\| t \right\|_1 \right)$$

satisfies for every $1 \leq p \leq 2$*,*

$$\left\| \tilde{t} - t^* \right\|_p \leq c_4(L, u_0, \kappa_0) \left\| \zeta \right\|_{L_{q_0}} s^{1/p} \sqrt{\frac{\log(ed)}{N}} \ ,$$

with probability at least

$$1 - \frac{c_2 \log^{q_0} N}{N^{q_0/2-1}} - 2\exp\left(-c_3 s \log(ed/s)\right) \ . \tag{1}$$

This paper shows that Theorem 1 holds for a MOM version of the LASSO estimator under much weaker assumptions, with a better probability estimate than (1). More precisely, the following theorem is proved.

Theorem 2. *Assume that* t^* *is* s*-sparse,* $N \geq c_0 s \log(ed/s)$*,* X *is isotropic and*

i') $|\mathcal{I}| \geq N/2$ *and* $|\mathcal{O}| \leq c_1 s \log(ed/s)$ *(the number of outliers may be proportional to the sparsity times* $\log(ed/s)$*),*

ii) $\zeta = Y - \langle X, t^* \rangle \in L_{q_0}$ *for some* $q_0 > 2$

iii') *for every* $1 \leq p \leq C_0 \log(ed)$*,* $\left\| \langle X, e_j \rangle \right\|_{L_p} \leq L\sqrt{p} \left\| \langle X, e_j \rangle \right\|_{L_2}$ *where* $(e_j)_{j \in [d]}$ *is the canonical basis of* $\mathbb{R}^d$ *and* C_0 *is some absolute constant,*

iv') *there exists* θ_0 *such that* $\left\| \langle X, t \rangle \right\|_{L^1} \leq \theta_0 \left\| \langle X, t \rangle \right\|_{L^2}$*, for all* $t \in \mathbb{R}^d$*,*

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

There exists an estimator $\hat{t}$*, called MOM-LASSO, satisfying for every* $1 \leq p \leq 2$*,*

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

with probability at least

$$1 - c_2 \exp(-c_3 s \log(ed/s)) \ . \tag{2}$$

Theoretical properties of MOM LASSO outperform those of LASSO 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 in (1) is polynomial – $1/N^{(q_0/2-1)}$ in (1) – it is exponentially small for MOM LASSO. Exponential rates for LASSO hold only if ζ is subgaussian ($\left\| \zeta \right\|_{L_p} \leq C\sqrt{p} \left\| \zeta \right\|_{L_2}$ 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 be responsible of a dramatic breakdown of the performances of LASSO.
- All assumptions on X are weaker for MOM LASSO than for LASSO. In particular, condition *v)* holds with $\theta_m = \|\zeta\|_{L_4}$ if for all $t \in \mathbb{R}^d$, $\left\|\left\langle X, t\right\rangle\right\|_{L_4} \leq \theta_0 \left\|\left\langle X, t\right\rangle\right\|_{L_2}$ – which is a much weaker requirement than condition iii) for LASSO.

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) to handle non-i.d. data. SBM is also extended to bound both quadratic and multiplier parts of the quadratic loss. Otherwise, all arguments are standard, which makes the approach very attractive and easily reproducible in other frameworks of statistical learning.

The paper is organized as follows. Section 2 briefly presents the general setting and our main illustrative example. Section 3 presents Le Cam's construction of estimators based on tests. We also show why many learning procedures may be obtained by this approach. 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 $\theta_\cdot$, 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 $\mathcal{C} \subset \mathcal{G}$,

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

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 $f \in F$,

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

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

Example : ℓ_1-regularization of linear functionals. For every $t = (t_j)_1^d \in \mathbb{R}^d$ and $1 \leq p \leq +\infty$, let

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

Let $f^* = \langle \cdot, t^*\rangle \in F$, where

$$t^* \in \operatorname*{argmin}_{t\in\mathbb{R}^d} \left\{P\big(Y - \langle X, t\rangle\big)^2\right\} .$$

Whenever it's necessary, $(e_1, \dots, e_d)$ will denote the canonical basis of $\mathbb{R}^d$ and B_p^d (resp. S_p^{d-1}) will denote the unit ball (resp. sphere) associated to $\|\cdot\|_p$. To ease readability in this example, we focus on rates of convergence, we do not consider the "full" non-i.i.d. setup and assume that $P = P_i$ for all $i \in \mathcal{I}$. We write L^q for L_P^q to shorten notations.

3. Learning from tests

3.1. General Principle

This section details the ideas underlying the construction of a MOM estimator using an extension of Le Cam's approach.

Basic idea. By definition of the oracle f^*, one has

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

As $T_{\text{id}}(g, f) = R(f) - R(g)$ depends on P, we estimate it by test statistics $T(g, f, (X_i, Y_i)_{i\in[N]}) \equiv T_N(g, f)$ that is, real random variables such that

$$T_N(f, g) + T_N(g, f) = 0 . \tag{3}$$

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. (12) 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 measuring (or related to) the risk we want to control. Let for all $f \in F$,

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

be the set of all functions $g \in F$ that beat f. If f is far from f^*, then $\mathcal{B}_{T_N}(f)$ is expected to have a large radius w.r.t. $d(\cdot, \cdot)$. We therefore introduce this radius as a criteria to minimize : for all $f \in F$, let $C_{T_N}(f) = \sup_{g\in\mathcal{B}_{T_N}(f)} d(f, g)$.

By (3), $f \in \mathcal{B}_{T_N}(g)$ or $g \in \mathcal{B}_{T_N}(f)$ (both happen if $T_N(f,g)=0$), hence $d(f,g) \le C_{T_N}(f) \vee C_{T_N}(g)$. In particular, for all $f \in F$,

$$d(f,f^*) \le C_{T_N}(f) \vee C_{T_N}(f^*) \ . \tag{4}$$

Eq (4) suggests to define the estimator

$$\hat{f}_{T_N} \in \operatorname*{argmin}_{f\in F} C_{T_N}(f) = \operatorname*{argmin}_{f\in F} \sup_{g\in\mathcal{B}_{T_N}(f)} d(f,g) \ . \tag{5}$$

This estimator satisfies, from Eq (4),

$$d(\hat{f}_{T_N}, f^*) \le C_{T_N}(f^*) \ . \tag{6}$$

Risk bounds for $\hat{f}_{T_N}$ follow from (6) 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*

$$\mathcal{B}_{T_N}(f,\mathcal{F}) = \{g \in \mathcal{F} : T_N(g,f) \ge 0\} \tag{7}$$

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. Section 5 in Lecué and Mendelson (2013) for more details). These results are not presented because we are interested in exponentially large deviation results for which our results are optimal.

Dealing with regularization : the link function. Statistical performances of estimators and the radius of $\mathcal{B}_{T_N}(f^*)$ can be measured by two norms: the regularization norm $\|\cdot\|$ and $\|.\|_{L^2_P}$. As (5) allows only for one distance d, we propose the following extension of Le Cam approach to handle two metrics.

To introduce this extension, assume first that $d(f,g) = \|f-g\|_{L^2_P}$ can be computed for all $f,g \in F$ (this is the case if the distribution of the design is known). The next paragraph explains how to deal with the more common framework where this distance is unknown. Remark that

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

The main point to extend Le Cam's approach to simultaneously control two norms is to design a link function $r(\cdot)$. In a nutshell, the values $r(\rho)$ is the L^2_P-minimax rate of convergence in a ball of radius ρ for the regularization norm (cf. (13) in Section 4.3 for a formal definition). Then one can define

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

Theorem 3 shows that while a minimizer $\hat{f}^{(1)}$ of C_{T_N} has only a nice risk for $\|\cdot\|$, a minimizer $\hat{f}^{(2)}$ of $C^{(2)}_{T_N}$ has both $\left\|\hat{f}^{(2)} - f^*\right\|$ 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. To deal with this issue, one considers usually the empirical $L^2_{P_N}$ distance and prove that empirical and actual distances are equivalent outside a L^2_P-ball centered in f^* (cf. for instance, remark after Lemma 2.6 in Lecué and Mendelson (2013)). Unfortunately this approach only works under strong concentration property that we want to relax in this paper.

The unknown L^2_P-metric is instead estimated by a median-of-means approach, that is, we use MOM estimators $d_N(f,g)$ of all $d(f,g)$ (cf. Section 4.4). The final estimator is therefore defined as a minimizer of

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

3.2. Examples

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 (5). In particular, any possibly penalized empirical risk minimizer

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

is obtained by Le Cam's construction with the tests

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

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 $\text{MOM}_K(Z)$ 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

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

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 (8) and (9) 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. Lugosi and Mendelson (2017) introduced median-of-means tournaments. More precisely, they used median-of-means tests to compare elements in F. These tests cannot be separated $T_N(f, g) \neq \ell_N(g) - \ell_N(f)$ in general. Lugosi and Mendelson (2017) assume that an upper bound r^* on the radius $C_{T_N}(f^*)$ of $\mathcal{B}_{T_N}(f^*)$ (that holds with exponentially large probability) is known from the statistician and call "champion" any element $\hat{f}$ of F such that $C_{T_N}(\hat{f}) \leq r^*$. It is clear that, by definition the radius $C_{T_N}(\hat{f}_{T_N})$ of $\hat{f}_{T_N}$ is smaller than $C_{T_N}(f^*)$ and therefore smaller than r^*. This means that $\hat{f}_{T_N}$ is a "champion" for this terminology. The main advantage of Le Cam's approach is that r^* (which usually depends on some attribute of the oracle like the sparsity) is not required to *build* the estimator $\hat{f}_{T_N}$.

4. Construction of the regularized MOM estimators

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

This section presents median-of-means (MOM) tests used in this work. Designing a family of tests $(T_N(g, f) : f, g \in F)$ is one of the most important building blocks in Le Cam's approach together with the right choice of the metric measuring the diameters $\mathcal{B}_{T_N}(f)$ for $f \in F$.

Start with a few notations. For all $\alpha \in [0, 1]$, $\ell \geq 1$ and $z \in \mathbb{R}^\ell$, the set of α-quantiles of z is denoted by

$$\mathcal{Q}_\alpha(z) = \left\{ x \in \mathbb{R} : \frac{1}{\ell}\sum_{k=1}^{\ell} I(z_i \leq x) \geq \alpha \quad \text{and} \quad \frac{1}{\ell}\sum_{k=1}^{\ell} I(z_i \geq x) \geq 1 - \alpha \right\} \ .$$

For a non-empty subset $B \subset [N]$ and a function $f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, let

$$P_B f = \frac{1}{|B|}\sum_{i \in B} f(X_i, Y_i) \text{ and } \overline{P}_B f = \frac{1}{|B|}\sum_{i \in B} P_i f \ .$$

Let $K \in [N]$ and let $(B_1, \ldots, B_K)$ denote an equipartition of $[N]$ into bins of size $|B_k| = N/K$. When K does not divide N, at most $K-1$ data can be removed from the dataset. For any real number $\alpha \in [0,1]$ and any function $f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$, the set of α-quantiles of empirical means is denoted by

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

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 $\sup \mathcal{Q}_{\alpha,K}(f) \geq u$, $Q_{\alpha,K}(f) \leq u$ if $\inf \mathcal{Q}_{\alpha,K}(f) \leq u$, 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 $\alpha \in [0,1]$,

$$\mathcal{Q}_{\alpha,K}(cf) = c\mathcal{Q}_{\alpha,K}(f) \enspace , \tag{8}$$

$$\mathcal{Q}_{\alpha,K}(-f) = -\mathcal{Q}_{1-\alpha,K}(f) \enspace , \tag{9}$$

$$\sup\left\{\mathcal{Q}_{1/4,K}(f) + \mathcal{Q}_{1/4,K}(f')\right\} \leq \inf \mathcal{Q}_{1/2,K}(f+f') \enspace , \tag{10}$$

$$\sup \mathcal{Q}_{1/2,K}(f+f') \leq \inf\left\{\mathcal{Q}_{3/4,K}(f) + \mathcal{Q}_{3/4,K}(f')\right\} \enspace . \tag{11}$$

With some abuse of notations, we shall write these properties respectively

$$Q_{\alpha,K}(cf) = cQ_{\alpha,K}(f), \qquad Q_{\alpha,K}(-f) = -Q_{1-\alpha,K}(f) \enspace ,$$
$$Q_{1/4,K}(f) + Q_{1/4,K}(f') \leq \mathrm{MOM}_K\left[f+f'\right] \leq Q_{3/4,K}(f) + Q_{3/4,K}(f') \enspace .$$

A *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

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

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

$$T_{K,\lambda}(g,f) = \mathrm{MOM}_K\left[\ell_f^\lambda - \ell_g^\lambda\right] = \mathrm{MOM}_K\left[\ell_f - \ell_g\right] + \lambda(\|f\| - \|g\|) \enspace . \tag{12}$$

From (9), $T_{K,\lambda}$ satisfies (3) and is a tests statistic in the sense of Section 3.

4.2. Main assumptions

Recall that $[N] = \mathcal{O} \cup \mathcal{I}$ and that $(X_i, Y_i)_{i\in\mathcal{O}}$ is a set of outliers on which we make no assumption so these may be aggressive in any sense one can imagine. The remaining informative data $(X_i, Y_i)_{i\in\mathcal{I}}$ need to bring enough information onto f^*. We therefore need some assumption on the sub-dataset $(X_i, Y_i)_{i\in\mathcal{I}}$ and, in particular, some connexion between the distributions P_i for $i \in \mathcal{I}$ and P. These assumptions are pretty weaksince we only assume essentially that the $L^2_P, L^2_{P_i}$ and $L^1_{P_i}$ geometries are comparable in the following sense.

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

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

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 $f \in F$,*

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

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, for example, 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 $f \in F$, $\|f - f^*\|_{L^4_Q} \leq \theta_1 \|f - f^*\|_{L^2_P}$ then by Cauchy-Schwarz inequality, $\sqrt{\text{var}_Q(\zeta(f - f^*))} \leq \|\zeta(f - f^*)\|_{L^2_Q} \leq \|\zeta\|_{L^4_Q} \|f - f^*\|_{L^4_Q} \leq \theta_1\sigma \|f - f^*\|_{L^2_P}$ and so Assumption 2 holds for $\theta_m = \theta_1\sigma$.

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

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

By Cauchy-Schwarz inequality, $\|f - f^*\|_{L^1_{P_i}} \leq \|f - f^*\|_{L^2_{P_i}}$ 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 1 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 1. *Let Z be a real-valued random variable.*

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

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

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

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

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

As one can assume that $\|Z\|_2 \neq 0$, $p \geq (\theta_0^{-1} - \kappa_0)^2$. ■

4.3. Complexity parameters and the link function

This section defines the link function $r(\cdot)$ making the connections between norms that will be required in the extension of Le Cam's approach to a simultaneous control of two norms (one of the two being unknown). For any $\rho \geq 0$ and any $f \in E$, let

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

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

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

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

and the two fixed point functions

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

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

$$r(\rho) = r(\rho, \gamma_Q, \gamma_M) \geq \max(r_Q(\rho, \gamma_Q), r_M(\rho, \gamma_M)). \tag{13}$$

Note that if the function $\rho \to \max(r_Q(\rho, \gamma_Q), r_M(\rho, \gamma_M))$ is itself continuous and non-decreasing then it can be taken equal to $r(\cdot)$. In the next paragraph, we provide an explicit computation of the functions $r_Q(\cdot)$, $r_M(\cdot)$ and $r(\cdot)$ in the "LASSO case".

Complexity parameters for the ℓ_1-regularization. One can derive $r_Q(\cdot)$ and $r_M(\cdot)$ from Gaussian mean widths defined for any $V \subset \mathbb{R}^d$, by

$$\ell^*(V) = \mathbb{E}\left\{\sup_{(v_j)\in V}\sum_{j=1}^{d} g_j v_j\right\}, \quad \text{where} \quad (g_1,\dots,g_d)\sim\mathcal{N}_d(0,I_d)\ . \tag{14}$$

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 under the following assumption.

Assumption 4. *There exist constants $q_0 > 2$, C_0 and L such that $\zeta \in L^{q_0}$, X is isotropic ($\mathbb{E}\langle X,t\rangle^2 = \|t\|_2^2$ for every $t \in \mathbb{R}^d$) and its coordinates have $C_0 \log d$ subgaussian moments: for every $1 \le j \le d$ and every $1 \le p \le C_0 \log d$, $\left\|\langle X,e_j\rangle\right\|_{L^p} \le L\sqrt{p}\left\|\langle X,e_j\rangle\right\|_{L^2}$.*

Under Assumption 4, if $\sigma = \|\zeta\|_{L^{q_0}}$, (Mendelson, 2016, Theorem 1.6) shows that, for every $\rho > 0$,

$$\mathbb{E}\sup_{v\in\rho B_1^d\cap rB_2^d}\left|\sum_{i\in[N]}\epsilon_i\langle v,X_i\rangle\right| \le c_2\sqrt{N}\ell^*(\rho B_1^d\cap rB_2^d)\ ,$$

$$\mathbb{E}\sup_{v\in\rho B_1^d\cap rB_2^d}\left|\sum_{i\in[N]}\epsilon_i\zeta_i\langle v,X_i\rangle\right| \le c_2\sigma\sqrt{N}\ell^*(\rho B_1^d\cap rB_2^d)\ .$$

Local Gaussian mean widths $\ell^*(\rho B_1^d \cap rB_2^d)$ 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

$$r^2(\rho) \sim_{L,q_0,\gamma_Q,\gamma_M} \begin{cases} \max\left(\rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\sigma^2 d}{N}\right) & \text{if } N \gtrsim_L d \\ \max\left(\rho\sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho\sqrt{N}}\right)}, \frac{\rho^2}{N}\log\left(\frac{d}{N}\right)\right) & \text{otherwise}\end{cases}\ . \tag{15}$$

4.4. *The estimators*

Let $(T_{K,\lambda}(g,f))_{f,g\in F}$ denote the family of tests defined in (12). For every function $f \in F$, let $\mathcal{B}_{K,\lambda}(f) = \{g \in F : T_{K,\lambda}(g,f) \ge 0\}$ denote the set of all functions $g \in F$ that beats f. As explained in Section 3, these sets will be measured by two metrics. First, let

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

Next, let

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

Lemma 4 below proves that, with large probability, $\text{MOM}_K\left[|f - g|\right]$ and $\|f - g\|_{L_P^2}$ are isomorphic distances. The second criterion is then given by

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

where $r(\cdot)$ is a link function as defined in Definition 1. That is a continuous and non-decreasing function such that for all $\rho > 0$, $r(\rho) \geq \max(r_M(\rho, \gamma_M), r_Q(\rho, \gamma_Q))$ where the choice of γ_Q and γ_M is given in Theorem 3 below. The associated estimator is then given by

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

4.5. The sparsity equation

By (6), estimation rates for $\hat{f}_{K,\lambda}^{(2)}$ will be derived from upper bounds on $C_{K,\lambda}^{(2)}(f^*)$. To get these, our strategy is to show that $T_{K,\lambda}(f^*, f) > 0$ for all f such that $\|f - f^*\|$ or $\|f - f^*\|_{L_P^2}$ is large.

Recall that the quadratic / multiplier decomposition of the excess quadratic risk:

$$T_{K,\lambda}(f^*, f) = \text{MOM}_K[(f - f^*)^2 - 2\zeta(f - f^*)] + \lambda(\|f\| - \|f^*\|) \enspace . \tag{16}$$

Let $f \in F$ and $\rho = \|f - f^*\|$. When ρ is large and $\|f - f^*\|_{L_P^2}$ is small, $T_{K,\lambda}(f^*, f) > 0$ thanks to the regularization term $\lambda(\|f\| - \|f^*\|)$ in (16) because the quadratic term $(f - f^*)^2$ is likely to be small. We will therefore derive a lower bound on the regularization term when the subdifferential of $\|\cdot\|$ is "large" in the following sense.

First, we recall that the subdifferential of $\|\cdot\|$ in $f \in F$ is the set

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

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 is associated with a notion of "sparsity" if one agrees to say that a non-zero function f^{**} is *sparse* w.r.t. the norm $\|\cdot\|$ when the subdifferential of this norm at f^{**} is a "large subset" of the dual sphere (i.e. the sphere of $(E^*, \|\cdot\|^*)$). 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 are formalized in the sparsity equation. More precisely, let

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

$\Delta(\rho)$ is a uniform lower bound on $\|f\| - \|f^{**}\|$ if $f^{**} \in B(f^*, \rho/20)$. Thus, $\|f\| - \|f^*\| \gtrsim \rho$, if $\sup_{f^{**} \in \Gamma_{f^*}(\rho)}(\|f\| - \|f^{**}\|) \gtrsim \rho$ or if the following *sparsity equation* of Lecué and Mendelson (2016a) holds.

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

$$\rho^* = \inf\left(\rho > 0 : \Delta(\rho) \geq \frac{4\rho}{5}\right) . \tag{17}$$

The sparsity equation in ℓ_1^d-regularization. The equation has been solved in this example in (Lecué and Mendelson, 2016a, Lemma 4.2), recall this result.

Lemma 1. *If there exists $v \in \mathbb{R}^d$ such that $v \in t^* + (\rho/20)B_1^d$ and $|\text{supp}(v)| \leq c\rho^2/r^2(\rho)$ then*

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

where S_1^{d-1} is the unit sphere of the ℓ_1^d-norm and B_2^d is the unit Euclidean ball in $\mathbb{R}^d$.

If $N \gtrsim s\log(ed/s)$ and if there exists a s-sparse vector in $t^* + (\rho/20)B_1^d$, Lemma 1 and the choice of $r(\cdot)$ in (15) imply that for $\sigma = \|\zeta\|_{L^{q_0}}$,

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

then ρ^* satisfies the sparsity equation and $r^2(\rho^*)$ is the rate of convergence of the LASSO (cf. Lecué and Mendelson (2016a)).

5. Main results

5.1. Performances of the estimators

Theorem 3 gathers estimation error bounds satisfied by the estimators $\hat{f}^{(j)}_{K,\lambda}$ for $j = 1, 2$ defined in Section 4.4.

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

$$\gamma_Q = \min\left(\frac{1}{661\theta_0}, \frac{1}{17640\theta_r}\right), \gamma_M = \frac{\epsilon}{168} \text{ and } \epsilon = \frac{3}{331\theta_0^2}.$$

Let ρ^* be defined in (17) and let K^* denote the smallest integer such that

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

For all $K \geq 1$, let ρ_K be a solution of $r^2(\rho_K) = [16\theta_m^2/(\epsilon^2\alpha)]\sqrt{K/N}$. Assume that for every $i \in \mathcal{I}$, $K \in [K^*, N]$ and $f \in F \cap B(f^*, \rho_K)$,

$$2(P_i - P)\zeta(f - f^*) \leq \epsilon \max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho_K, \gamma_M), \|f - f^*\|_{L_P^2}^2\right) \quad . \tag{18}$$

For all $K \in [K^*, N/(84\theta_r^2\theta_0^2)]$, on an event $\Omega_1(K)$ such that $\mathbb{P}(\Omega_1(K)) \geq 1 - 4\exp(-K/1008)$, the estimators $\hat{f}^{(j)}_{K,\lambda}$ for $j = 1, 2$ defined in Section 4.4 satisfy

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

and

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

when the regularization parameter satisfy

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

To the best of our knowledge, Theorem 3 provides the first statistical performance of an estimator operating in such a "nasty" environment: the dataset may be corrupted by complete outliers, the informative data may be heavy-tailed and their distribution P_i for $i \in \mathcal{I}$ is only asked to have a L^2 and L^1 geometry over $F - f^*$ equivalent to that of P. The most surprising thing is that the rate we obtain for $K = K^*$ in Theorem 3, i.e. $r(\rho_{K^*})$ when the number of outliers K_o is less than $Nr^2(\rho^*)$ is the minimax rate we would have gotten in a very good i.i.d. subgaussian framework with independent noise. This means that the quality of a dataset does not have to be as good as it is classically assumed in the literature to make estimation possible: all we need is that a large fraction of the data should be independent (even though we believe that some "weak dependence" could also be introduced) and distributed according to distributions inducing L^1 and L^2 geometries equivalent to the L_P^2 one.

In Theorem 3, K can be as small as the infimum between the number of outliers and N times the minimax rate of convergence. Henceforth, if the optimal rate is known, as in Lugosi and Mendelson (2017), Theorem 3 shows that Le Cam's champion of the median of means tournament with $K = K^*$ reaches the same performances as any champion in this paper. Theorem 3 is thus an extension of Lugosi and Mendelson (2017) to a non-i.d. corrupted setting for Le Cam's champion. Moreover, our control improves theirs if the upper bound on the radius of f^* used in Lugosi and Mendelson (2017) is pessimistic (cf. Example 3.2 in Section 3.2).

Assumption 1 is automatically satisfied in the i.i.d. case and so is Assumption (18). Theorem 3 goes beyond this i.i.d. setup, relaxing the i.d. assumptions into proximity assumptions between $L_{P_i}^2$ and L_P^2 geometries, for informative data.

Risk bounds in ℓ_1^d regularization. Let us now compute explicit values of ρ_K and $\lambda \sim r^2(\rho_K)/\rho_K$ in the ℓ_1^d-regularization case. Let $K \in [N]$ and $\sigma = \|\zeta\|_{L^{q_0}}$. The equation $K = cr(\rho_K)^2 N$ is solved by

$$\rho_K \sim_{L,q_0} \frac{K}{\sigma}\sqrt{\frac{1}{N}\log^{-1}\left(\frac{\sigma^2 d}{K}\right)} \tag{19}$$

for the $r(\cdot)$ function defined in (15). Therefore,

$$\lambda \sim \frac{r^2(\rho_K)}{\rho_K} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho_K\sqrt{N}}\right)} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma^2 d}{K}\right)}\ . \tag{20}$$

The regularization parameter depends on the "level of noise" σ, the L^{q_0}-norm of ζ. This parameter is unknown in practice. Nevertheless, it can be estimated and replaced by this estimator in the regularization parameter as in (Giraud, 2015, Sections 5.4 and 5.6.2).

5.2. Adaptive choice of K by Lepski's method

The main drawback of Theorem 3 is that optimal rates are only achieved when $K \approx K^*$. Since K^* is unknown, it cannot be used in general. This issue is tackled in this section by Lepski's method.

Let $K_1 = K^*$ and $K_2 = N/(84\theta_0^2\theta_r^2)$ be defined as in Theorem 3. For any integer $K \in [K_1, K_2]$, let ρ_K and λ be defined as in Theorem 3 and for $j = 1, 2$ denote by $\hat{f}_K^{(j)} = \hat{f}_{K,\lambda}^{(j)}$ for this choice of λ. These estimators are the building blocks of the following confidence sets. For all $f \in F$, let

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

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

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

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 4. *Grant assumptions and notations of Theorem 3. 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 $1 - c_1 \exp(-c_2 K)$,*

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

and

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

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

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

then,

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

Recall an optimality result from Lecué and Mendelson (2013). 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

$$\zeta_N \gtrsim \min\left(r, \mathrm{diam}(F, L_P^2) \right) . \tag{21}$$

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 $r(\rho_K) \leq \mathrm{diam}(F, L_P^2)$ since $r(\rho_K)$ can obviously be replaced by $r(\rho_K) \wedge \mathrm{diam}(F, L_P^2)$ in all results). Moreover, this rate is minimax since (Lecué and Mendelson, 2013, Theorem A) also shows that the ERM over $\rho_K B$, $\hat{f}_N^{ERM} \in \mathrm{argmin}_{f \in \rho_K B} P_N \ell_f$, 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 4 shows that $\hat{f}_{LE}$ achieves the same rate of convergence with the same exponentially high confidence as a minimax estimator does in the Gaussian regression model (with independent noise). These rates are achieved here under very weak stochastic assumptions allowing the presence of outliers, without assuming that the regression function lies in F or that the data are i.i.d.. Compared to Lugosi and Mendelson (2017), using a Lepski method, we don't have to *choose* the integer K in advance, we let the data decide the best choice and automatically get an estimator with the correct minimax rate of convergence. Moreover, the regularization parameter is chosen adaptively, which yields to exact minimax rates and, since this minimax rate is not required to build the estimators, these are naturally adaptive.

Adaptive results in ℓ_1^d regularization. The following result follows from Theorem 4 together with the computation of ρ^*, r_Q, r_M and r from the previous sections. This is a slight extension of Theorem 2 to the case where the oracle t^* is not exactly sparse but close to a sparse vector.

Theorem 5. *Assume that X is isotropic 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 $\|t^* - v\|_1 \leq \sigma s \sqrt{\log\left(ed/s\right)/N}/20$ and $|\mathrm{supp}(v)| \leq s$.*
- *i') $|\mathcal{I}| \geq N/2$ and $|\mathcal{O}| \leq c_1 s \log(ed/s)$,*
- *ii) $\zeta = Y - \left\langle X, t^* \right\rangle \in L_{q_0}$ for some $q_0 > 2$*
- *iii') for every $1 \leq p \leq C_0 \log(ed)$, $\left\|\left\langle X, e_j \right\rangle\right\|_{L_p} \leq L\sqrt{p}\left\|\left\langle X, e_j\right\rangle\right\|_{L_2}$ where $(e_j)_{j\in[d]}$ is the canonical basis of $\mathbb{R}^d$ and C_0 is some absolute constant,*
- *iv') there exists θ_0 such that $\left\|\left\langle X, t\right\rangle\right\|_{L^1} \leq \theta_0 \left\|\left\langle X, t\right\rangle\right\|_{L^2}$, for all $t \in \mathbb{R}^d$,*
- *v) there exists θ_m such that $\mathrm{var}(\zeta\left\langle X, t\right\rangle) \leq \theta_m^2 \|t\|_2^2$, for all $t \in \mathbb{R}^d$.*

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

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

In particular, Theorem 5 shows that, for our estimator contrary to the one in Lugosi and Mendelson (2017), the sparsity parameter s does not have to be known in advance in the LASSO case.

Proof. It follows from Theorem 4, the computation of $r(\rho_K)$ from (15) and ρ_K in (19) that with probability at least $1 - c_0 \exp(-cr(\rho_K)^2 N/\overline{C})$, $\left\|\hat{t}_{LE} - t^*\right\|_1 \leq \rho_{K^*}$ 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 $1 \leq p \leq 2$. ■

6. Proofs

In all the proof section, we denote by $\mathbb{P}$ the distribution of $(X_1, \dots, X_N)$ 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 $f \in L^2_P$ and $r \geq 0$, let

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

We consider the set of indices of blocks B_k containing only informative data:

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

6.1. Lower Bound on the quadratic process

Lemma 2. *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 $\gamma\,(1 - \alpha - x - 32\theta_0\gamma_Q) \geq 1 - \eta$. Let $K \in [K_o/(1-\gamma), N\alpha/(2\theta_0\theta_r)^2]$.*

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 $\|f - f^*\|_{L^2_P} \geq r_Q(\rho, \gamma_Q)$ then

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

Proof. For all $f \in F - \{f^*\}$, let $n_f = (f - f^*)/\|f - f^*\|_{L^2_P}$. 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 $K \leq [\alpha/(2\theta_0\theta_r)]^2 N$ then $|B_k| = N/K \geq [\alpha/(2\theta_0\theta_r)]^2$ and so we have

$$\mathbb{P}\left(P_{B_k}|n_f| \geq \frac{1}{2\theta_0}\right) \geq 1 - \alpha \ . \tag{22}$$

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 (22) 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 $f \in F$, we have

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

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

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

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. (Ledoux and Talagrand, 1991, Theorem 4.12) or (Boucheron et al., 2013, Theorem 11.6)),

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

Applying again the symmetrization and contraction principles, we get,

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

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 $J = \cup_{k \in \mathcal{K}} B_k$,

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

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

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

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 process

Lemma 3. *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 $\gamma\left(1-\alpha-x-8\gamma_M/\epsilon\right) \geq 1-\eta$. Let $K \in [K_o/(1-\gamma), N]$. 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*

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

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

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

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

$$\begin{aligned}
\mathbb{P}\left[2\left|W_k(f)\right| \geq \gamma_k(f)\right] &\leq \frac{4\mathbb{E}\left[\left(2(P_{B_k}-\overline{P}_{B_k})(\zeta(f-f^*))\right)^2\right]}{\frac{16\theta_m^2}{\alpha}\|f-f^*\|_{L_P^2}^2\frac{K}{N}} \\
&\leq \frac{\alpha\sum_{i\in B_k}\mathrm{var}_{P_i}(\zeta(f-f^*))}{|B_k|^2\theta_m^2\|f-f^*\|_{L_P^2}^2\frac{K}{N}} \leq \frac{\alpha\theta_m^2\|f-f^*\|_{L_P^2}^2}{|B_k|\theta_m^2\|f-f^*\|_{L_P^2}^2\frac{K}{N}} = \alpha \enspace . \qquad (23)
\end{aligned}$$

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

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

where in the last but one inequality we used that F is convex and the same argument as in the proof of Lemma 2. 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,

$$\begin{aligned}
\mathbb{E}\sup_{f\in B(f^*,\rho)}\sum_{k\in\mathcal{K}}\epsilon_k\frac{W_k(f)}{\gamma_k(f)} &\leq \frac{4K}{\epsilon r_M^2 N}\mathbb{E}\sup_{f\in B(f^*,\rho)\cap B_2(f^*,r_M)}\left|\sum_{i\in J}\epsilon_i\zeta_i(f-f^*)(X_i)\right| \\
&\leq \frac{4K}{\epsilon N}\gamma_M|\mathcal{K}|\frac{N}{K} = \frac{4\gamma_M}{\epsilon}|\mathcal{K}| \enspace . \qquad (24)
\end{aligned}$$

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

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

where we used (23) in the last inequality. The bounded difference inequality ensures that, for all $x > 0$, there exists an event $\Omega(x)$ satisfying $\mathbb{P}(\Omega(x)) \geq 1 - \exp(-x^2|\mathcal{K}|/2)$ on which

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

Furthermore, it follows from the symmetrization argument that

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

and, from the contraction principle and (24), that

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

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

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

■

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

Besides the controls of the quadratic and multiplier MOM processes presented in Lemmas 2 and 3 respectively, the estimation error bounds for the MOM estimators rely on the following isometry property of the MOM processus $f \in F \to \mathrm{MOM}_K\left[|f-f^*|\right]$.

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

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

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

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

$$\frac{1}{4\theta_0}\left\|f-f^*\right\|_{L_P^2} \leq MOM_K\left[|f-f^*|\right] \leq \theta_r\left(1+\frac{4}{\alpha}\right)\left\|f-f^*\right\|_{L_P^2}. \tag{25}$$

Proof. It follows from Lemma 2 that on the event $\Omega_Q(K,\rho)$ for all $f \in B(f^*,\rho)$, if $\left\|f-f^*\right\|_{L_P^2} \geq r_Q(\rho,\gamma_Q)$ then $Q_{\eta,K}|f-f^*| \geq (1/(4\theta_0)\left\|f-f^*\right\|_{L_P^2}$. This yields the "lower bound" result in (25).

For the upper bound of the isomorphic result, we essentially repeat the proof of Lemma 3. Let us just highlight the main differences. We will use the same notation as in the proof of Lemma 3 except that for all $f \in F$, we define

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

It follows from Chebyshev's inequality and Assumption 1 that

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

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

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

and then using a symmetrization argument, we obtain that

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

Finally, using the same argument as in the proof of Lemma 3, for all $x > 0$ there exists an event $\Omega(x)$ such that $\mathbb{P}(\Omega(x)) \ge 1 - \exp(-x^2 |\mathcal{K}|/2)$, on which for all $f \in B(f^*, \rho)$,

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

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, $P_{B_k}|f - f^*| \le \overline{P}_{B_k}|f - f^*| + \gamma_k(f)$. Now, the result follows from Assumption 1 since $\overline{P}_{B_k}|f - f^*| \le \theta_r \left\| f - f^* \right\|_{L^2_P}$.

■

6.4. *Conclusion to the proof of Theorem 3*

The proof relies on the following proposition.

Proposition 2. *Grant conditions of Theorem 3. 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 $\mathbb{P}(\Omega_0(K)) \ge 1 - 2\exp(-K/1008)$ and on $\Omega_0(K)$ for all $f \in F$ if $\|f - f^*\|_{L^2_P} \ge r(\rho_K)$ or $\|f - f^*\| \ge \rho_K$ then

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

Proof. Using (8), (9) and (11) together with the quadratic / multiplier decomposition of the excess quadratic loss yields that for all $f \in F$,

$$\begin{aligned}\mathrm{MOM}_K\left[\ell_f - \ell_{f^*}\right] &= \mathrm{MOM}_K\left[(f - f^*)^2 - 2\zeta(f - f^*)\right] \\ &\ge Q_{1/4,K}[(f - f^*)^2] - 2Q_{3/4,K}[\zeta(f - f^*)] \ . \end{aligned} \tag{26}$$

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

$$\eta=\frac{1}{4}, \gamma=\frac{7}{8}, \alpha=\frac{1}{21}, x=\frac{1}{21}, \gamma_Q=\frac{1}{661\theta_0}, \gamma_M=\frac{\epsilon}{168} \text{ and } \epsilon\leq\frac{1}{64\theta_0^2}. \tag{27}$$

For this choice of constants, Lemma 2 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

$$Q_{1/4,K}[(f-f^*)^2]\geq\frac{1}{(4\theta_0)^2}\left\|f-f^*\right\|_{L_P^2}^2 \quad . \tag{28}$$

Moreover, for the choice of parameters as in (27), we also have $\gamma(1-\alpha-x-8\gamma_M/\epsilon)\geq 1-\eta$, hence Lemma 3 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 Assumpion (18), it follows that on the event $\Omega_M(K,\rho_K)$, for all $f\in B(f^*,\rho_K)$,

$$2Q_{3/4,K}[\zeta(f-f^*)]\leq 2\epsilon\max\left(\frac{16\theta_m^2}{\epsilon^2\alpha}\frac{K}{N}, r_M^2(\rho_K,\gamma_M), \|f-f^*\|_{L_P^2}^2\right) \quad . \tag{29}$$

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

$$\mathrm{MOM}_K\left[(f-f^*)^2-2\zeta(f-f^*)\right]\geq -2\epsilon r^2(\rho_K) \quad . \tag{30}$$

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 (26), (28) and (29), the choice of ϵ in (27) and the definition of ρ_K that

$$\mathrm{MOM}_K\left[(f-f^*)^2-2\zeta(f-f^*)\right]\geq\left(\frac{1}{(4\theta_0)^2}-2\epsilon\right)\|f-f^*\|_{L_P^2}^2\geq\frac{\|f-f^*\|_{L_P^2}^2}{32\theta_0^2}. \tag{31}$$

Now, if $\|f-f^*\|_{L_P^2}\leq r^2(\rho_K)$ then it follows from (26), (29) 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 (30) 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 (31), 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 5. *For all $f \in F$ such that $\|f-f^*\| \geq \rho_K$*

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

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

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

■

Lemma 6. *Assume that, for all $f \in F \cap S(f^*, \rho_K)$,*

$$MOM_K\left[(f-f^*)^2 - 2\zeta(f-f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) > \lambda\frac{\rho_K}{10} \quad . \tag{32}$$

Then (32) 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 (32) that for $\kappa = \|f-f^*\|/\rho_K \geq 1$, one has

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

■

Let $f \in F$ be such that $\|f - f^*\| \geq \rho_K$. By Lemma 5,

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

Therefore, it will follow from Lemma 6 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

$$\mathrm{MOM}_K\left[(g-f^*)^2 - 2\zeta(g-f^*)\right] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) > \lambda\frac{\rho_K}{10} \ . \tag{33}$$

Let us now prove that (33) holds. Let $g \in F$ be such that $\|g - f^*\| = \rho_K$. First assume that $\|g - f^*\|_{L^2_P} \leq r(\rho_K)$ 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 (30), when $\lambda > 20\epsilon r^2(\rho_K)/(7\rho_K)$, one has

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

Finally assume that $\|g - f^*\|_{L^2_P} \geq r(\rho_K)$. Since $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq -\|f - f^*\| = -\rho_K$, it follows from (31) that

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

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

End of the proof of Theorem 3. On the event $\Omega_0(K)$ of Proposition 2, $\mathcal{B}_{K,\lambda}(f^*)$ is included in the ball $B(f^*, \rho_K)$, therefore, by definition of $\hat{f}^{(1)}_{K,\lambda}$ (cf. (6)),

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

Again, by Proposition 2, 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 4, for all $f \in \mathcal{B}_{K,\lambda}(f^*)$,

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

where $\alpha = 1/21$ according to (27). Therefore, $C^{(2)}_{K,\lambda}(f^*) \leq \rho_K$, which implies that $\left\| \hat{f}^{(2)}_{K,\lambda} - f^* \right\| \leq \rho_K$ (cf. (6)) and that $C^{(2)}_{K,\lambda}(\hat{f}^{(2)}_{K,\lambda}) \leq \rho_K$ and therefore, by Lemma 4, on $\Omega_0(K) \cap \Omega_{iso}(K)$, either $\left\| \hat{f}^{(2)}_{K,\lambda} - f^* \right\|_{L^2_P} \leq r_Q(\rho_K, \gamma_K)$ and so $\left\| \hat{f}^{(2)}_{K,\lambda} - f^* \right\|_{L^2_P} \leq 340\theta_0\theta_r r(\rho_K)$ or $\left\| \hat{f}^{(2)}_{K,\lambda} - f^* \right\|_{L^2_P} \geq r_Q(\rho_K, \gamma_K)$ and so

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

6.5. Conclusion to the proof of Theorem 4

First, it follows from Theorem 3 that for all $K \in [K_1, K_2]$, with probability at least $1 - c_0 \exp(-c_1 K)$, for both $j = 1, 2$, $f^* \in \cap_{J=K}^{K_2} R^{(j)}_K$, so $\widehat{K}^{(j)} \leq K$, which implies that both f^* and $\widehat{f}^{(j)}_{\mathrm{LE}}$ belong to $B(\widehat{f}^{(j)}_{K,\lambda}, \rho_K)$, therefore, $\left\| f^* - \widehat{f}^{(j)}_{\mathrm{LE}} \right\| \leq 2\rho_K$.

Second, for the L^2_P-estimation error bound of $\widehat{f}^{(2)}_{\mathrm{LE}}$, denote by $r_J = 340\theta_r\theta_0 r(\rho_J)$ the bound on the L^2_P risk of the estimator $\hat{f}^{(2)}_J$ obtained in Theorem 3. Let $K \in [K_1, K_2]$. It follows from Lemma 4 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 $f \in B(f^*, 2\rho_J)$,

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

Let Ω be the event defined as the following intersection:

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

It follows from Theorem 3 that $\mathbb{P}(\Omega) \geq 1 - c_3 \exp(-c_4 K)$. Moreover, on Ω, for all $J \geq K$,

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

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

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

In particular,

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

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

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

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,

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

■

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., 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.

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.

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.