Learning from MOM's principles

G. Lecué and M. Lerasle

January 8, 2017

Abstract

We obtain estimation error rates and sharp oracle inequalities for a Birgé's T-estimator using a regularized median of mean principle as based tests. The results hold with exponentially large probability – the same one as in the gaussian framework with independent noise– under only weak moments assumption like a L_4/L_2 assumption and without assuming independence between the noise and the design X. The obtained rates are minimax optimal. The regularization norm we used can be any norm. When it has some sparsity inducing power we recover sparse rates of convergence and sparse oracle inequalities. As in [29], the size of the sub-differential of the regularization norm in a neighborhood of the oracle plays a central role in our analysis.

Moreover, the procedure allows for robust estimation in the sense that a large part of the data may have nothing to do with the oracle we want to reconstruct. The number of such irrelevant data (which can be seen as outliers) may be as large as *(sample size)*×*(rate of convergence)* as long as the quantity of useful data is larger than a proportion of the number of observations.

As a proof of concept, we obtain the "exact" minimax rate of convergence $s\log(ed/s)/N$ for the problem of recovery of a s-sparse vector in $\mathbb{R}^d$ via a median of mean version of the LASSO under a L_{q_0} assumption on the noise for some $q_0 > 2$ and a $C_0 \log(ed)$ moment assumption on the design matrix. As mentionned previously this result holds with exponentially large probability as if the noise and the design were independent and standard gaussian random variables.

1 Introduction

An important problem in learning theory is to estimate a minimizer $f^* \in \operatorname{argmin}_{f\in F} P(Y - f(X))^2$ over a convex class of functions F of the integrated square-loss based on a data set $(X_i, Y_i)_{i=1,\dots,N}$. The Empirical Risk Minimizer (ERM) of [42] and later on, its penalized versions propose to replace the unknown distribution P by the empirical distribution P_N based on the sample $(X_i, Y_i)_{i=1,\dots,N}$, to choose a non-negative function pen : $F \to \mathbb{R}$ and to define

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

This estimator is well understood now and is known to suffer several drawbacks when the data are heavy-tailed or in the presence of "outliers" [15]. 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 being to "smooth" the shape of the square function at infinity to make it less sensitive to large data. For example, the Hüber loss [21] replaces the function $x \to x^2$ by $x \to x^2 I(|x| \le \tau) + [\tau(2|x| - \tau)]I(|x| > \tau)$ that is it interpolates between the square loss that leads to the unbiased (but non robust) empirical mean estimator and the absolute loss that leads to the (more robust but biased) empirical median. Beside the asymptotic results of [21], this estimator has been studied in a non-asymptotic framework, see for example [16, 19] and the references therein. An alternative to the Hüber function has been proposed by Catoni [15] and used in learning frameworks by Audibert and Catoni [2] in least-squares regression and for more general loss functions by Brownlees, Joly and Lugosi [13].

Another line of research for building robust estimators and robust estimator selection procedures has been proposed by Birgé [9], Baraud [4] and Baraud, Birgé and Sart [6] following ideas of Le Cam [26, 25]. It is based on *comparison* between the elements of F. More precisely, their idea is to build tests statistics $T_N(\cdot,\cdot)$ to choose between any two elements in F, these tests are used to build a"confidence region" $\mathcal{B}_{T_N}(f)$ for any $f \in F$ containing all g's that have been preferred to f and to select the estimator having the smallest confidence region given a way to measure its diameter – which is directly related to the statistical performance one wants to prove. Usually, these methods focus on the Hellinger loss and were considered impossible to compute for a long time until [7, 40], for example, managed to compute these estimators in particular cases.

In a related but different approach, Lugosi and Mendelson [32] have recently introduced the notion of "median of means tournaments". Their idea is to use the Median of means principle of [1, 22, 38] to make comparisons between the elements of F. The authors call "champion" any element of F with a confidence region smaller than the one of the oracle. They prove that any champion has a risk controlled by the one of the oracle. In Section 3 we shall show why any possibly penalized empirical loss function can also be seen as a Birgé's estimator and it is clear from its definition that Birgé's estimator built with the median of means tests is a "Champion of a Median of Means tournament".

In this paper, we build on the idea of Lugosi and Mendelson [32] and use (a regularized version of) their Median Of Mean tests (see Section 4.1) in a Birgé's procedure. Its performances are studied with respect to the square-loss. The main advantage of this approach compared to Birgé's original one is that it does not require to work with the Hellinger loss, allowing more classical losses functions in learning that typically fail in the presence of heavy-tailed data. We shall illustrate this idea by focusing on the square-loss. Compared to the Huber loss or Catoni's loss, this approach allows to control easily the risk of our estimators by using classical tools from empirical process theory. Compared to [32], we stress the link between the notions of "champions" and Birgé's estimation procedure by aggregation of tests, and we study the robustness of the approach with respect to "outliers", that will be defined as data whose distribution are not related to P.

More precisely, we build three different types of estimators and prove three different types of results. First, to use Birgé's procedure, we have to compute the "radii" of the sets $\mathcal{B}_{T_N}(f), f \in F$ for various "metrics" associated to some statistical measure of performance such as the regularization norm, the L_P^2-norm or exact oracle inequalities. The radius can naturally be evaluated with respect to the regularization norm since this norm is known. It is chosen by the statistician in advance and it is usually used to promote sparsity or smoothness. If one wants estimation result w.r.t. the L_P^2-norm then ideally the diameter of the sets $\mathcal{B}_{T_N}(f), f \in F$ should be measure w.r.t. the L_P^2-norm. The problem with the L_P^2-norm is that it is unknown in general since the distribution of X is not known in general. So we shall first evaluate empirically the metric structure induced by L_P^2 over F. We will do so using again a median of means principle and then show that the corresponding Birgé's estimator is well located. The main issue here is that we want to infer the metric structure associated with L_P^2 only very weak moment assumption. We cannot use the classical "isomorphic approach" used to compare the empirical metric $L_{P_N}^2$ to the actual metric L_P^2 since this approach requires strong subgaussian properties of the design vector X that we don't have here.

Note that to obtain estimation results w.r.t. both the regularization norm and the L_P^2-norm, we need to slightly extend Birgé's criterion since we need to compute the diameter of the $\mathcal{B}_{T_N}(f), f \in F$ for two criteria simultaneously. Nevertheless, once the mechanism used to measure the diameter of the $\mathcal{B}_{T_N}(f), f \in F$ w.r.t. two metrics (one of which being empirical) one can use this idea a third time to obtain exact oracle inequalities. For such statistical results and as in the case of the estimation of the L_P^2 diameter of the $\mathcal{B}_{T_N}(f), f \in F$, we have to estimate the "metric" structure induced by the excess risk over F using again a median of means principle (still because we do not have strong concentration property that allows for an "isomorphic approach" between the empirical and actual excess risk). We show that the resulting estimator based on the measure of the diameter of the $\mathcal{B}_{T_N}(f), f \in F$ w.r.t. the previous three criteria is well localized w.r.t. both the regularization norm and the L_P^2-norm as well as its excess risk is properly bounded. Note that the third criteria we added to Birgé's procedure is based on the median of means principle to control the correlation between the noise $Y - f^*(X)$ and $(f - f^*)(X)$ for all $f \in F$. All these results are based on (regularized) median of means tests and depend on the number K of blocks of data that should be ultimately chosen using an unknown parameter associated to the oracle f^* like its sparsity in the case where the regularization norm has some sparsity inducing power. To overcome this issue, we use Lepski's method [31] as in [18] to select K adaptively and get rid of this dependence and get a fully data-dependent procedure. As we will see later this adaptation step is the reason why we can get the exact minimax rate in sparse recovery $s\log(ed/s)/N$ (cf. [8]) for a MOM version of the LASSO whereas the classical LASSO estimator achieves the rate $s\log(ed)/N$.

There are four important features in our approach. First, all the results are proved under weak assumptions on the noise, that is only required to have an $L^{2+\epsilon}$ moment, which is almost the minimal condition under which the problem at hand makes sense and the class of functions may only satisfy a weak moment condition as a "L_4/L_2" comparison. Second, the performances of the estimator are not affected by the presence of complete outliers, as long as their number remains comparable to *(number of observations)*×*(rates of convergence)* which is for the problem of sparse-recovery of the order of the sparsity of the oracle. Third, the results are non-asymptotic and do not require the regression function $x \mapsto \mathbb{E}[Y|X = x]$ to belong to the class F. In particular, the noise $Y - f^*(X)$ can be correlated with $f^*(X)$ and/or Y. Finally, even the "informative data" that is the one that are not "complete outliers" are not requested to be i.i.d. $\sim P$, but only to have close first and second moments for all $f \in F$. Nevertheless, the estimators are shown to behave as well as the ERM when i.i.d. $\sim P$ data, $\mathbb{E}[Y|X = \cdot] \in F$ and the noise $\xi = Y - f^*(X)$ is independent from the design and has Gaussian distribution.

An example: sparse-recovery in $\mathbb{R}^d$ via the MOM LASSO. As a proof of concept, these properties are illustrated in the classical example of sparse-recovery in high-dimensional spaces using the ℓ_1-norm as penalization. We only study this example because it has been one of the most studied example in high dimensional statistics [14, 20] even though our approach also applies to other procedures like Slope [11, 41], trace-norm regularization and kernel methods for instance. This example shows the minimax optimality of our rates obtained under weak moment assumption on the noise and the design – which may be dependent – and the robustness to outliers property of the new estimator. Moreover, the parameter λ used to balance between adequacy to the data and regularization is chosen adaptively in our final estimator and does not a priori depend on the sparsity parameter, as it has to be the case for the Lasso estimator if one wants to achieve the exact rate of convergence which is $s\log(ed/s)/N$. Let us now show the advantage of our procedure over the classical regularization methods for the problem of sparse recovery. We first recall the setup of this problem then recall a result from [29] on the LASSO and then state our result for this example.

We assume that X is an isotropic random vector in $\mathbb{R}^d$ (i.e. $\mathbb{E}\left\langle X, t\right\rangle^2 = \|t\|_2^2$ for all $t \in \mathbb{R}^d$) and Y is a real-valued random vector. We define $t^* \in \operatorname{argmin}_{t\in\mathbb{R}^d} \mathbb{E}(Y - \left\langle X, t\right\rangle)^2$. We are given a dataset $(X_i, Y_i)_{i\in[N]}$ of independent (not necessarily identically distributed) random variables which has been contaminated by outliers: that is a set of data $(X_i, Y_i)_{i\in O}$ such that the distribution of (X_i, Y_i) for $i \in O \subset [N]$ has nothing to do with the distribution of (X, Y). We denote by $I = [N]\backslash O$, the set of indices of the *informative data* $(X_i, Y_i)_{i\in I}$ and we assume that for all $i \in I$, (X_i, Y_i) is distributed like (X, Y).

In the high-dimensional statistics setup, one has $N \leq d$ but it is believed that t^* has a small support of size s such that $s < N$. For the reconstruction of such a vector, one may use the ℓ_1-norm $\|\cdot\|_1$ for the regularization norm because it promotes zero coordinates by thresholding and therefore induces sparsity. It follows from Theorem 1.4 in [29] that for $0 < \delta < 1$ if t^* is s-sparse, $N \geq c_0 s\log(ed/s)$,

i) $|I| = N$ and so $|O| = 0$ (there is no outliers in the dataset),

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

iii) for all $t \in \mathbb{R}^d$ and every $p \in \mathbb{N}$, $\left\|\left\langle X, t\right\rangle\right\|_{L_p} \leq L\sqrt{p}\left\|\left\langle X, t\right\rangle\right\|_{L_2}$

then the LASSO estimator

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

is such that with probability at least $1 - \delta$, for every $1 \leq p \leq 2$,

$$\left\| \tilde{t} - t^* \right\|_p \leq c_2(L, \delta) \left\| \xi \right\|_{L_{q_0}} s^{1/p} \sqrt{\frac{\log(ed)}{N}}.$$

Compare with the MOM LASSO $\hat{t}_{LE}$ that we introduce later: we obtain that if t^* is s-sparse, $N \geq c_0 s \log(ed/s)$,

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

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

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

iv) $\operatorname{var}(\xi \left\langle X, t \right\rangle) \leq C_\xi^2 \left\| t \right\|_2^2$ for all $t \in \mathbb{R}^d$,

the MOM LASSO estimator $\hat{t}_{LE}$ is such that with probability at least $1 - c_2 \exp(-c_3 s \log(ed/s))$, for every $1 \leq p \leq 2$,

$$\left\| \hat{t}_{LE} - t^* \right\|_p \leq c_2(L) \left\| \xi \right\|_{L_{q_0}} s^{1/p} \sqrt{\frac{1}{N} \log\left(\frac{ed}{s}\right)}.$$

Note that iv) holds with $C_\xi = \left\| \xi \right\|_{L_4}$ if a L_4/L_2 assumption holds: for all $t \in \mathbb{R}^d$, $\left\| \left\langle X, t \right\rangle \right\|_{L_P^4} \leq c_0 \left\| \left\langle X, t \right\rangle \right\|_{L_P^2}$ – which is a much weaker requirement than condition iii) for the LASSO.

One may pinpoint several differences between the LASSO $\hat{t}$ and the MOM LASSO $\hat{t}_{L_E}$: the estimation rate obtained for the later is the exact minimax rate; the deviation parameter δ for the LASSO is obtained using a Chebyshev's inequality and is like $1/N^{(q_0/2-1)}$ whereas it is exponentially small for the MOM LASSO – note however that when the noise ξ is assumed to be subgaussian (i.e. $\left\| \xi \right\|_{L_p} \leq C\sqrt{p} \left\| \xi \right\|_{L_2}$ for all $p \in \mathbb{N}$), the deviation parameter δ for the LASSO is the same as the one for the MOM LASSO that is $\delta = c_2 \exp(-c_3 s \log(ed/s))$ – finally, the MOM LASSO allows for a number of outliers proportional to the sparsity s whereas such property is either unknown or simply not true at all for the LASSO. As a consequence, the theoretical properties of the MOM LASSO outperforms in several way the one of the LASSO.

From a mathematical point of view, our results are based on a slight extension of the Small Ball Method of [23] to handle non-i.i.d. data. The SBM is also extended to bound both the quadratic and the multiplier parts of the decomposition of the quadratic loss. Otherwise, all other arguments have been made simple, 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 present Birgé's construction of estimators based on tests. We also show why many learning procedures may be seen as Birgé's estimators. The construction of the 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 $[x] = \{1, \ldots, \lfloor x \rfloor\}$ if $x \geq 1$. For any finite set A, let $|A|$ denote its cardinality. For every vector $t = (t_j)_1^d \in \mathbb{R}^d$ and $1 \leq p \leq +\infty$ denote the ℓ_p^d-norm of t by $\left\| t \right\|_p = \left(\sum_{j=1}^d |t_j|^p \right)^{1/p}$ and the associated unit ball by B_p^d and unit sphere by S_p^{d-1}.

2 Setting

Let $\mathcal{X}$ denote a measurable space and let $(X,Y),(X_i,Y_i)_{i\in[N]}$ denote independent 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

$$\|f\|_{L^2_Q}<\infty,\qquad \text{where}\qquad \|f\|_{L^2_Q}=\left(Qf^2\right)^{1/2}\ .$$

Let $F\subset L^2_P$ denote a convex class of functions $f:\mathcal{X}\to\mathbb{R}$. Assume that $Y\in L^2_P$ and let, for any $f\in F$,

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

We want to estimate f^* based on $(X_i,Y_i)_{i\in[N]}$. Let $\|\cdot\|$ denote a norm defined onto a linear subspace E of L^2_P containing F, that will be used as a regularization function.

Example : ℓ_1-regularization of linear functionals

As a proof of concept, we apply our results to ℓ_1^d regularization of linear functionals, which has been one of the most studied regularization method in high-dimensional statistics during the last decade (cf. [20, 14]). For this problem, F is the class of all linear functionals indexed by $\mathbb{R}^d$ and the regularization function is (with a slight abuse of notation) the ℓ_1-norm over $\mathbb{R}^d$. That is

$$F=\{\langle\cdot,t\rangle:t\in\mathbb{R}^d\}\qquad\text{and}\qquad \|\langle\cdot,t\rangle\|=\|t\|_1\ .$$

For this example, t^* denotes the "coefficient" oracle that is $f^*=\langle\cdot,t^*\rangle\in F$, where

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

The aim of this example is to show that the parameters introduced can be computed in particular examples and to present typical results that follow from our analysis in a well studied case. For this example, we focus on the rates of convergence and the confidence bounds. In particular, we do not consider the non-i.i.d. setup even if it directly follows from our analysis. We always assume that $P=P_i$ for all $i\in[N]$ and we shall write L^q for L^q_P to shorten notations.

3 Learning from tests

3.1 General Principle

Our approach is based on pairwise comparisons between elements of F. More precisely, given two functions f and g in F, we would like to prefer f to g if $P[(Y-f(X))^2]\le P[(Y-g(X)^2]$, or, equivalently, if

$$0\le P[(Y-g(X))^2-(Y-f(X))^2]=P[(g(X)-f(X))^2-2(Y-f(X))(g(X)-f(X))]=T_{\mathrm{id}}(f,g)\ .$$

As the distribution P is unknown, such a direct comparison is impossible and our purpose is to design test statistics $T(f,g,(X_i,Y_i)_{i\in[N]})\equiv T_N(f,g)$ that is, real random variables such that $T_N(f,g)+T_N(g,f)=0$, $T_N(f,g)$ being an estimation of $T_{\mathrm{id}}(f,g)$. These statistics are used to *compare* f and g, simply by saying that f *T_N-beats* g iff $T_N(f,g)\ge 0$.

The next step of our construction is to use Birgé's idea (see for example [9] and the references therein), which is also sometimes attributed to Le Cam (see [25] and the references therein). Define

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

or more generally, to compare only the elements of a subset $\mathcal{F} \subset F$, typically a maximal ϵ-net, define

$$\forall f \in \mathcal{F} : \qquad \mathcal{B}_{T_N}(f, \mathcal{F}) = \{g \in \mathcal{F} : T_N(g, f) \geq 0\} \ . \tag{1}$$

Notice that, from the assumption $T_N(f, g) + T_N(g, f) = 0$, one has always $f \in \mathcal{B}_{T_N}(g)$ or $g \in \mathcal{B}_{T_N}(f)$ (both happen if $T_N(f, g) = 0$). Therefore, if one is given a pseudo-metric d and if we set $C_{T_N}(f) = \sup_{g \in \mathcal{B}_{T_N}(g)} d(f, g)$, we always have $d(f, g) \leq C_{T_N}(f) \vee C_{T_N}(g)$, thus, from the definition of $\hat{f}_{T_N}$,

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

To upper bound the risk of $\hat{f}_{T_N}$ it is therefore sufficient to upper bound the radius of $\mathcal{B}_{T_N}(f^*)$. The definition of "radius" depends on the distance d that should be chosen according to the properties we seek for our estimators (see Section 4.4 for details regarding the different norms we use and the corresponding results on the estimators). More precisely, we are interested with at least two natural norms : the regularization norm $\|\cdot\|$ measuring the "sparsity" and the norm $\|.\|_{L^2_P}$ measuring the excess risk. We intend to design an estimator that is performant for both norms and we will slightly extend Birgé's construction to that purpose. Assume first that the norms $\|f\|_{L^2_P}$ for $f \in F$ can be computed (which is the case, for example, if the distribution of the design is known). Since, we have two different norms at hands, it is not clear which one we should use to compute the radius of $\mathcal{B}_{T_N}(f)$. To deal with this issue, we first rewrite Birgé's criterion (for $\|\cdot\|$) by

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

A minimizer of this criterion is well localized but may have a large excess risk. To control also the excess risk, we use a function $r : \mathbb{R}^+ \to \mathbb{R}^+$ (r will be defined in Definition 1 in Section 4.3) that basically maps the optimal rate of convergence in the reference norm $\|\cdot\|$ to the optimal rate in $\|\cdot\|_{L^2_P}$. The extension of Birgé's criterion is then defined by

$$C^{(2)}_{T_N}(f) = \min\{\rho \geq 0 : \forall g \in \mathcal{B}_{T_N}(f),\ \|g - f\| \leq \rho \text{ and } \|f - g\|_{L^2_P} \leq r(\rho)\} \ . \tag{3}$$

In general, the L^2_P-norm between two functions cannot be directly computed, so we have to estimate the L^2_P-distance between elements of $\mathcal{B}_{T_N}(f)$. Moreover, we shall also be interested in proving "exact" oracle inequalities. This involves a third distance, and a third criterion that extends again a bit Birgé's criterion using the same ideas (see Section 4.4 for details).

Birgé's approach has been used to define T-estimators [5, 9, 10], ρ-estimators [3, 6] (although the aggregation of tests is a bit different there) and to build general estimator selection procedures [4, 7]. This procedure is in general computationally expensive. It also extends many common procedures in statistical learning theory under some assumptions on the tests the procedure relies on. To illustrate this purpose, we present now three examples of estimators that could be seen as Birgé's estimators.

3.2 Examples

Example 1 : Empirical (penalized) minimizers. Assume that T_N can be decomposed $T_N(f, g) = \ell_N(g) - \ell_N(f)$ and denote by $\hat{f} = \arg\min_{f \in F} \ell_N(f)$ (provided that such minimizer 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 Birgé's estimator coïncide with $\hat{f}$.

In particular, any possibly penalized empirical risk minimizer $\hat{f} = \arg\min_{f \in F}\{P_N \ell_f + \text{pen}(f)\}$ is a Birgé estimator corresponding to the tests $T_N(f, g) = P_N(\ell_g - \ell_f) + \text{pen}(g) - \text{pen}(f)$, where $\ell_f(x, y) = (y - f(x))^2$ in this paper.

Example 2 : Median Of Means estimators of the mean Another, perhaps less obvious example is the median of means estimator [1, 22, 38] of the expectation PZ of a real valued random variable Z. Recall that this estimator is built using a sample $Z_1, \dots, Z_N$ of i.i.d. copies of Z and a partition $B_1, \dots, B_K$ of $[N]$ into bins of equal sizes N/K. This estimator is then defined as the (empirical) median of the set of empirical means $\left\{P_{B_k}Z = \frac{1}{|B_k|}\sum_{i\in B_k} Z_i : k \in [K]\right\}$ and is denoted by $\mathrm{MOM}_K(Z)$. To see why this estimator can be obtained from Birgé's procedure, we remind that the expectation is the minimizer $PZ = \arg\min_{m\in\mathbb{R}} P(Z-m)^2$. A natural way to define a test comparing two real numbers m and m' is thus to define

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

Using basic properties of the median (recalled in Equations (4) and (5) in Section 4.1), we have

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

Defining $\ell_N(m) = (m - \mathrm{MOM}_K(Z))^2$, the test T_N is decomposed as in the previous example $T_N(m,m') = \ell_N(m') - \ell_N(m)$ and the unique minimizer $\hat{m}$ of $\ell_N(m)$ is $\mathrm{MOM}_K(Z)$.

Example 3 : "Champions" of a Tournament In a recent paper, Lugosi and Mendelson [32] introduced the notion of median of means tournaments. More precisely, they used the median of means tests (see Section 4.1) to compare the elements of F. Notice that these tests, as well as those used by Birgé [9, 10] don't satisfy a relation of the type $T_N(f,g) = \ell_N(g) - \ell_N(f)$ in general, and Birgé's estimator over the all class F cannot be obtained as a minimizer $\hat{f} = \arg\min_{f\in F} \ell_N(f)$. Lugosi and Mendelson [32] used the following (a priori different from Birgé) strategy to overcome this issue. They compute a tractable upper bound on the radius of the oracle and call "champion" any element of F with a radius smaller than this upper bound. They prove that any champion has a radius smaller than the upper bound. It is clear that, by definition the radius of Birgé's estimator is smaller than the one of the oracle and Birgé's estimator is therefore a champion.

4 Construction of our estimators

4.1 Quantile of means processes and Median Of Means tests

In this paper, we used Median Of Means (MOM) tests of [32]. To introduce these tests, useful quantities and basic properties in our analysis, we start with a general presentation of the *quantiles of means* processes. For any $\alpha \in [0,1]$, $\ell \geq 1$ and any $z \in \mathbb{R}^\ell$, denote the set of α-empirical quantiles of z by

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

For any non-empty subset $B \subset [N]$ and any vector $z = (z_1, \dots, z_N)^T \in \mathbb{R}^N$, let $P_B z = \frac{1}{|B|}\sum_{i\in B} z_i$. Let $K \in [N]$ and let $(B_1, \dots, B_K)$ denote an equipartition of $\{1, \dots, N\}$ with bins of size $|B_i| = N/K$. Note that when K does not divide N, one can remove a few data from the dataset. For any real number $\alpha \in [0,1]$ and any vector $z = (z_1, \dots, z_N)^T \in \mathbb{R}^N$, we denote the set of α-quantiles of empirical means by

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

With a slight abuse of notations, we shall repeatedly denote by $Q_{\alpha,K}(z)$ any element in $\mathcal{Q}_{\alpha,K}(z)$ and write $Q_{\alpha,K}(z) = y$ if $y \in \mathcal{Q}_{\alpha,K}(z)$, $Q_{\alpha,K}(z) \geq y$ if $\sup \mathcal{Q}_{\alpha,K}(z) \geq y$, $Q_{\alpha,K}(z) \leq y$ if $\inf \mathcal{Q}_{\alpha,K}(z) \leq y$, and $Q_{\alpha,K}(z)+Q_{\alpha',K}(z')$ any element in the Minkowski sum $\mathcal{Q}_{\alpha,K}(z)+\mathcal{Q}_{\alpha',K}(z')$. Let also $\text{MOM}_K(z)=Q_{1/2,K}(z)$ denote the empirical median of the empirical means on the blocs B_k. Empirical quantiles are not linear processes, but they satisfy nevertheless the following properties

$$\forall \lambda \geq 0, \forall z \in \mathbb{R}^N, \forall \alpha \in [0,1], \qquad Q_{\alpha,K}(\lambda z)=\lambda Q_{\alpha,K}(z)\ , \tag{4}$$

$$\forall z \in \mathbb{R}^N, \forall \alpha \in [0,1], \qquad Q_{\alpha,K}(-z)=-Q_{1-\alpha,K}(z)\ , \tag{5}$$

$$\forall z,z' \in \mathbb{R}^N, \qquad Q_{1/4,K}(z)+Q_{1/4,K}(z') \leq \text{MOM}_K(z+z') \leq Q_{3/4,K}(z)+Q_{3/4,K}(z')\ . \tag{6}$$

To balance the empirical risk and the regularization norm, it is classical to introduce a parameter $\lambda > 0$ called a regularization parameter that needs to be fit in a proper way to achieve an optimal trade-off between adequacy to the data and regularization. Define respectively the (quadratic) loss and regularized (quadratic) loss as the functions from $F \times \mathcal{X} \times \mathbb{R}$ to $\mathbb{R}$ by

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

To compare/test functions f and g in F, the median of means test between f and g of [32] is defined by

$$T_{K,\lambda}(f,g)=\text{MOM}_K(\ell_g^\lambda-\ell_f^\lambda)=\text{MOM}_K(\ell_g-\ell_f)+\lambda(\|g\|-\|f\|)\ . \tag{7}$$

Notice that, from (5), $T_{K,\lambda}(f,g)+T_{K,\lambda}(g,f)=0$, therefore, $T_{K,\lambda}$ is a test statistic as defined in Section 3.

4.2 Main assumptions

One of our motivations in this paper is to show the robustness properties of median of means estimators in statistical learning. We shall therefore denote by $I \cup O$ a partition of $[N]$, where O has cardinality K_o. The data $(X_i,Y_i)_{i\in O}$ are considered as *outliers*, that is the set of data for which no assumptions on P_i is made. The remaining set $(X_i,Y_i)_{i\in I}$ is the set of *informative* data, that is the ones one can use for estimation. And, of course, given the data $(X_i,Y_i)_{i=1}^N$ no one knows in advance which data is informative or not.

Even for the informative data, we do not assume that $(X_i,Y_i)_{i\in I}$ are i.i.d. $\sim P$, but that the moments of order 1 and 2 of all functions $f\in F$ are close to those of P as well as the correlations between f and the noise ξ. Let $\tau \geq 0$.

Assumption 1 ($R(\tau)$)**.** *Property $R(\tau)$ holds when, for any $i \in I$ and any $f,g\in F$, $P_i(f-g)^2 \leq \tau^2 P(f-g)^2$.*

Of course, property $R(1)$ holds in the i.i.d. framework, with $I=[N]$. The second assumption bounds the correlation between the noise $\xi = Y - f^*(X)$ and the design on the class $F-f^*$.

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

$$\sqrt{\text{var}_Q(\xi(f-f^*))}=\sqrt{Q(\xi(f-f^*))^2-(Q\xi(f-f^*))^2}\leq C_\xi\left\|f-f^*\right\|_{L_P^2}\ .$$

If $\|\xi\|_{L_Q^4}<\infty$ and, for every $f\in F$, $\|f-f^*\|_{L_Q^4}\leq c_0\left\|f-f^*\right\|_{L_P^2}$, then, by Cauchy-Schwarz inequality,

$$\sqrt{\text{var}_Q(\xi(f-f^*))}\leq\|\xi(f-f^*)\|_{L_Q^2}\leq\|\xi\|_{L_Q^4}\left\|f-f^*\right\|_{L_Q^4}\leq c_0\left\|\xi\right\|_{L_Q^4}\left\|f-f^*\right\|_{L_P^2}$$

so Assumption 2 holds for $C_\xi = c_0 \max_{Q\in\{P,(P_i)_{i\in I}\}}\|\xi\|_{L_Q^4}$.

The last assumption essentially states that the distribution of X uniformly spreads in the directions of F. It has been introduced in [24, 34] and it's called the small ball property.

Assumption 3 (The small ball property (SBP))**.** *There exist constants $0<\beta<1$ and $0<u\leq 1$, such that, for any $Q\in\{P,(P_i)_{i\in I}\}$ and any $f,h\in F\cup\{0\}$, $Q\left(|f-h|\geq\beta\|f-h\|_{L_P^2}\right)\geq u$.*

Numerous examples satisfy SBP, we refer to [23, 28, 33, 35, 36, 39] for some of them. Let us give here a classical example where it can be checked. Assume that there exists $c_0 > 0$ such that $\|f-h\|_{L^4_Q} \leq c_0(\|f-h\|_{L^2_Q} \wedge \|f-h\|_{L^2_P})$ and $\|f-h\|_{L^2_P} \leq \|f-h\|_{L^4_Q}$ for every $f, h \in F$ and any $Q \in \{P, (P_i)_{i\in I}\}$, then Assumptions 3 holds (and we've seen that Assumption 2 also). Indeed, let β be such that $c_0\beta < 1$, denoting by $p = Q\left(|f-h| \geq \beta\|f-h\|_{L^2_P}\right)$, the Paley-Zygmund argument [17, Proposition 3.3.1]) shows that

$$\|f-h\|^2_{L^4_Q} \leq c_0^2\|f-h\|^2_{L^2_Q} \leq c_0^2\left(Q(f-g)^2 I(|f-h| \leq \beta\|f-h\|_{L^2_P}) + Q(f-h)^2 I(|f-h| \geq \beta\|f-h\|_{L^2_P})\right)$$
$$\leq c_0^2\left(\beta^2\|f-g\|^2_{L^2_P} + \sqrt{p}\|f-h\|^2_{L^4_Q}\right) \leq c_0^2\|f-h\|^2_{L^4_Q}\left(\beta^2+\sqrt{p}\right),$$

therefore, Assumption 3 holds with $u = (1-c_0^2\beta^2)^2/c_0^2$.

4.3 Complexity parameters

The main purpose of this section is to define the function r connecting the norms in our extension (3) of Birgé's criterion. Let us start with some notations : for any $r, \rho \geq 0$ and any $f \in L^2_P$, let

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

Let also $rB_2 = B_2(0,r)$ and $rS_2 = S_2(0,r)$. For any $f \in E$, let

$$B(f,\rho) = \{g \in E : \|f-g\| \leq \rho\}, \qquad S(f,\rho) = \{g \in E : \|g-f\| = \rho\}, \qquad \rho B = B(0,\rho), \qquad \rho S = S(0,\rho) .$$

Definition 1. *Let $(\epsilon_i)_{i\in I}$ be independent Rademacher random variables, independent from $(X_i, Y_i)_{i=1}^N$. For any $\gamma_Q, \gamma_M > 0$ and $\rho > 0$ let*

$$r_Q(\rho,\gamma_Q) = \sup_{f^\star\in F}\inf\left\{r > 0 : \forall J \subset I, |J| \geq \frac{|I|}{2},\ \mathbb{E}\sup_{\substack{f,g\in F\cap B(f^\star,\rho)\\ \|f-g\|_{L^2_P}\leq r}}\left|\sum_{i\in J}\epsilon_i(f-g)(X_i)\right| \leq \gamma_Q|J|r\right\} ,$$

$$r_M(\rho,\gamma_M) = \sup_{f^\star\in F}\inf\left\{r > 0 : \forall J \subset I, |J| \geq \frac{|I|}{2},\ \mathbb{E}\sup_{\substack{f\in F\cap B(f^\star,\rho)\\ \|f-f^\star\|_{L^2_P}\leq r}}\left|\sum_{i\in J}\epsilon_i\xi_i(f-f^\star)(X_i)\right| \leq \gamma_M|J|r^2\right\} ,$$

and let $\rho \to r(\rho, \gamma_Q, \gamma_M)$ be a continuous and increasing function such that for every $\rho > 0$,

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

Explicit computations of the complexity parameters $r_Q(\cdot)$ and $r_M(\cdot)$ may follow from classical results on the expectation of symmetrized empirical processes. For instance, upper bounds can be obtained using Gaussian mean widths : for any $V \subset \mathbb{R}^d$, the Gaussian mean width of V is defined by

$$\ell^*(V) = \mathbb{E}\sup_{v=(v_j)\in V}\sum_{j=1}^d g_j v_j, \qquad \text{where} \qquad (g_1,\ldots,g_d) \sim \mathcal{N}_d(0, I_d) . \tag{8}$$

Now, the complexity parameters $r_Q(\cdot)$ and $r_M(\cdot)$ in the ℓ_1^d-regularization example are computed in [37, Theorem 1.6]. Since the dual norm of the ℓ_1^d-norm is 1-unconditional with respect to the canonical basis of $\mathbb{R}^d$ (cf. [37, Definition 1.4]); [37, Theorem 1.6] applies under the following set of assumptions.

Assumption 4. *There exist constants $q_0 > 2$ and L such that:*

A1 $\xi \in L_{q_0}$,

A2 X is isotropic : $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ for every $t \in \mathbb{R}^d$,

A3 the coordinates of X have $C_0 \log d$ sub-gaussian moments: for every $1 \leq j \leq d$

$$\left\|\langle X, e_j\rangle\right\|_{L_p} \leq L\sqrt{p} \textit{ for every } 1 \leq p \leq C_0 \log d$$

where $(e_1, \dots, e_d)$ is the canonical basis of $\mathbb{R}^d$ and C_0 is some absolute constant.

The unit ball of the regularization norm plays a central role in our analysis. For the ℓ_1^d-regularization example, this ball is the unit ℓ_1^d-ball $B_1^d = \{t \in \mathbb{R}^d : \|t\|_1 \leq 1\}$. Under Assumption 4, [37, Theorem 1.6] shows that, for every $\rho > 0$ the expectation of the symmetrized empirical process satisfies

$$\mathbb{E} \sup_{v \in \rho B_1^d \cap r B_2^d} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i \langle v, X_i\rangle \right| \leq \frac{c_2 \ell^*(\rho B_1^d \cap r B_2^d)}{\sqrt{N}} \tag{9}$$

and for the multiplier process

$$\mathbb{E} \sup_{v \in \rho B_1^d \cap r B_2^d} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i \xi_i \langle v, X_i\rangle \right| \leq \frac{c_2 \|\xi\|_{L_{q_0}} \ell^*(\rho B_1^d \cap r B_2^d)}{\sqrt{N}} \quad . \tag{10}$$

As a consequence, it only remains to bound from above the local Gaussian mean widths $\ell^*(\rho B_1^d \cap r B_2^d)$. This is done in [29, Lemma 5.3] and therefore the computation of the fixed points r_M and r_Q follow (note that we drop off γ_M and γ_Q since they will play the role of constants later)

$$r_M^2(\rho) \lesssim_{L,q_0} \begin{cases} \frac{\|\xi\|_{L_{q_0}}^2 d}{N} & \text{if } \rho^2 N \geq \|\xi\|_{L_{q_0}}^2 d^2 \\ \rho \|\xi\|_{L_{q_0}} \sqrt{\frac{1}{N} \log\left(\frac{e\|\xi\|_{L_{q_0}} d}{\rho\sqrt{N}}\right)} & \text{otherwise} \end{cases}, r_Q^2(\rho) \begin{cases} = 0 & \text{if } N \gtrsim_L d \\ \lesssim_L \frac{\rho^2}{N} \log\left(\frac{c(L)d}{N}\right) & \text{otherwise} \end{cases} .$$

In this example therefore,

$$r^2(\rho) \sim_{L,q_0} \begin{cases} \max\left(\rho \|\xi\|_{L_{q_0}} \sqrt{\frac{1}{N} \log\left(\frac{e\|\xi\|_{L_{q_0}} d}{\rho\sqrt{N}}\right)}, \frac{\|\xi\|_{L_{q_0}}^2 d}{N}\right) & \text{if } N \gtrsim_L d \\ \max\left(\rho \|\xi\|_{L_{q_0}} \sqrt{\frac{1}{N} \log\left(\frac{e\|\xi\|_{L_{q_0}} d}{\rho\sqrt{N}}\right)}, \frac{\rho^2}{N} \log\left(\frac{c(L)d}{N}\right)\right) & \text{otherwise} \end{cases} . \tag{11}$$

4.4 The estimators

As explained in Section 3, our procedure build on Birgé's construction of T-estimators (cf. [9, 10]) and uses the radii of the sets $\mathcal{B}_{K,\lambda}(f) = \{g \in F : T_{K,\lambda}(g, f) \geq 0\}$, where $T_{K,\lambda}$ has been defined in (7), as criteria to deduce our estimators. These radii are evaluated using different norms depending on the results one wants to obtain. More precisely, we shall prove deviation bounds w.r.t. the regularization norm and the L_P^2-norm as well as prediction result by proving sharp oracle inequalities. For each of these results, we design a way to measure the size of the sets $\mathcal{B}_{K,\lambda}(f)$ for $f \in F$.
Let us first start with the regularization norm and define Birgé's criterion $R_{K,\lambda}^{\text{reg}}(f) = \sup_{g \in \mathcal{B}_{K,\lambda}(f)} \{\|g - f\|\}$. Denote the corresponding estimator $\hat{f}_{K,\lambda}^{(1)} \in \arg\min_{f \in F} R_{K,\lambda}^{reg}(f)$. The risk of this estimator will be bounded w.r.t. the regularization norm.

To get estimation results for both the regularization norm and the L_P^2-norm, one needs to add another way to measure the size of the sets $\mathcal{B}_{K,\lambda}(f)$ related to this norm. Given that P is unknown in general, we have to introduce an empirical way to measure the L_P^2 diameter of $B_{K,\lambda}(f)$. The classical empirical $L_{P_N}^2$-metric would work under strong assumptions on the design, like a subgaussian property or any other property insuring some isometry between $L_{P_N}^2$ and L_P^2 metrics above some level (usually given by $r_Q(\cdot)$, cf.

[27]). We want to relax also this assumption and consider therefore another "empirical norm". Instead of taking the empirical mean, we simply take the empirical median, that is, we define the empirical pseudo-norm: for every $f \in F$, $\|f\|_{L_{2,N}} = \mathrm{MOM}_N(f)$ and the radius of $\mathcal{B}_{K,\lambda}(f)$ w.r.t. $L_{2,N}$ is denoted by $R_{K,\lambda}^{L_{2,N}}(f) = \sup_{g\in\mathcal{B}_{K,\lambda}(f)} \left\{\|g-f\|_{L_{2,N}}\right\}$. The second criterion is then given by

$$C^{(2)}(f) = \min\left\{\rho \geq 0 : R_{K,\lambda}^{\mathrm{reg}}(f) \leq \rho,\ R_{K,\lambda}^{L_{2,N}}(f) \leq \eta r(\rho)\right\} \ ,$$

where η is defined in Theorem 1 below and $r(\rho) = r(\rho, \gamma_Q, \gamma_M)$ for the choice of constants γ_Q and γ_M in Theorem 1 below. The estimator associated to this criterion is $\hat{f}_{K,\lambda}^{(2)} \in \arg\min_{f\in F} C^{(2)}(f)$.

Finally, we want to design an estimator satisfying simultaneously a sharp oracle inequality and optimal risk bounds for both the regularization and L_P^2 norms. We introduce a third pseudo-distance :

$$\forall f, g \in F, \qquad d_K(g,f) = \mathrm{MOM}_K[(Y-f)(f-g)] \ ,$$

the radius of $\mathcal{B}_{K,\lambda}(f)$ w.r.t. d_K is denoted by $R_{K,\lambda}^{d_K}(f) = \sup_{g\in\mathcal{B}_{K,\lambda}(f)}\{d_K(g,f)\}$. The third criterion is given by

$$C^{(3)}(f) = \min\left\{\rho \geq 0 : R_{K,\lambda}^{\mathrm{reg}}(f) \leq \rho,\ R_{K,\lambda}^{L_{2,N}}(f) \leq \eta r(\rho), R_{K,\lambda}^{d_K}(f) \leq \gamma r^2(\rho)\right\}$$

where η and γ are specified in Theorem 1 below and the associated estimator is $\hat{f}_{K,\lambda}^{(3)} \in \arg\min_{f\in F} C^{(3)}(f)$.

4.5 The sparsity equation

As shown in (2), the performances of our estimators are bounded by those of the oracle f^*. To control for example $C^{(2)}(f^*)$ and deduce estimation bounds for f^* with respect to $\|\cdot\|$ and $\|\cdot\|_{L_P^2}$, we have to prove that $T_{K,\lambda}(f^*, f) \geq 0$ for any f such that $\|f - f^*\|$ or $\|f - f^*\|_{L_P^2}$ is large. Recall that

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

Given $f \in F$ and a radius ρ such that $\|f - f^*\| = \rho$. When $\|f - f^*\|_{L_P^2}$ is small, the quadratic term $(f-f^*)^2$ in this decomposition may be small as well and therefore of little help if one wants to prove that $T_{K,\lambda}(f^*, f) \geq 0$. Therefore, we need to use the term $\lambda(\|f\| - \|f^*\|)$ coming from penalization to prove that $T_{K,\lambda}(f^*, f) \geq 0$ when $\|f - f^*\|_{L_P^2}$ is small. The *sparsity equation* [29] connects $\|f - f^*\|$ with $\|f\| - \|f^*\|$ for all f closed to f^* in L_P^2. To bound from below $\|f\| - \|f^*\|$, we introduce the subdifferentials of $\|\cdot\|$:

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

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

For any $\rho > 0$, let H_ρ denote the set of functions "closed" to f^* in L_P^2 and at distance ρ from f^* in regularization norm. Let $\Gamma_{f^*}(\rho)$ denote the set of subdifferentials of vectors "closed" to f^* :

$$H_\rho = \{f \in F : \|f - f^*\| = \rho \text{ and } \|f - f^*\|_{L_P^2} \leq r(\rho)\} \text{ and } \Gamma_{f^*}(\rho) = \bigcup_{f\in F:\|f-f^*\|\leq\rho/20} (\partial\|\cdot\|)_f \ .$$

Now, the idea underlying the sparsity equation is that, if there exists a "sparse" f^{**} vector in $\{f \in F : \|f^* - f\| \leq \rho/20\}$, this vector will have a "large subdifferential" meaning that $\|f\| - \|f^{**}\|$ is large for any $f \in H_\rho$ so $\|f\| - \|f^*\|$ can be lower bounded by $\|f\| - \|f^{**}\| - \|f^* - f^{**}\|$ and thus be large as well. More precisely, let us introduce

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

$\Delta(\rho)$ is the uniform lower bound over all $f \in H_\rho$ of what can be interpreted as the maximal lower bound on $\|f\| - \|f^{**}\|$ for $f^{**} \in \Gamma_{f^*}(\rho)$. The maximal value of $\|f\| - \|f^*\|$ when $f \in H_\rho$ is ρ by the triangular inequality. According to our previous analysis, we will have $\|f\| - \|f^*\| \gtrsim \rho$ if there exists $c_0 > 1$ such that, for all $f \in H_\rho$, $\sup_{f^{**}\in\Gamma_{f^*}(\rho)}(\|f\| - \|f^{**}\|) \geq c_0\rho$. As explained, $\Delta(\rho)$ will be used to lower bound $\sup_{f^{**}\in\Gamma_{f^*}(\rho)}(\|f\| - \|f^{**}\|)$ and our goal will follow from the following inequality, introduced in [29] and called the sparsity equation.

Definition 2. *A radius $\rho > 0$ is said to satisfy the* ***sparsity equation*** *when $\Delta(\rho) \geq \frac{4\rho}{5}$.*

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

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

Example (continued): ℓ_1^d-regularization

In this example, the sparsity equation has been solved in [29]. The key idea is that the ℓ_1^d-norm has singularities at sparse vectors. Therefore, if the oracle t^* is close to a sparse vector then the size of $\Gamma_{f^*}(\rho)$ (for $f^*(\cdot) = \langle \cdot, t^*\rangle$) will be large because the subdifferential in this neighbor of f^* is large when ρ is large enough. Let us recall this result.

Lemma 1. *[29, Lemma 4.2] If there exists $v \in \mathbb{R}^d$ such that $v \in t^* + (\rho/20)B_1^d$ and $100|\mathrm{supp}(v)| \leq c(\rho/r(\rho))^2$ 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 $\mathbb{R}^d$.

When $N \gtrsim s\log(ed/s)$, it follows from Lemma 1 and the choice of the function $r(\cdot)$ in (11) that

$$\rho^* \sim_{L,q_0} \|\xi\|_{L_{q_0}}\, s\sqrt{\frac{1}{N}\log\left(\frac{ed}{s}\right)}$$

if t^* is "close" to a s-sparse vector (i.e. there exists a s-sparse vector in $t^* + (\rho^*/20)B_1^d$). With this ρ^*,

$$r^2(\rho^*) \sim \frac{\|\xi\|_{L_{q_0}}\, s}{N}\log\left(\frac{ed}{s}\right).$$

5 Main results

5.1 Performances of the estimators

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

Theorem 1. *Assume that F satisfies Assumption 3 (SBP) with constants $0 < \beta < 1$ and $3/4 < u \leq 1$, that ξ satisfies Assumption 2 and that property $R(\tau)$ holds for some $\tau \geq 1$ (see Assumption 1). Assume that there exists positive absolute constants $c_i \in (0, 1)$, such that*

$$(1 - c_2)(1 - c_6)\,(u - 3c_1) \geq \frac{3}{4}, \quad (1 - c_2)\left(1 - 4\frac{c_5}{c_3^2} - c_7 - \frac{16c_4}{c_3}\right) \geq \frac{3}{4}, \quad \frac{40}{7}c_3 \leq \frac{10}{11}\left(\frac{3}{4}c_6 - 4c_3\right) \quad . \tag{12}$$

Let $\gamma_Q = c_1\beta$, $\gamma_M = c_4\beta^2$, $c = c_5\frac{\beta^4}{C_\xi}$, let, for all $\rho > 0$, $r(\rho) = r(\rho, \gamma_Q, \gamma_M)$ and let $K^* = \lceil cNr^2(\rho^*)\rceil \vee \frac{K_o}{c_2}$. For any $K \geq 1$ define ρ_K as the solution of $r_M(\rho_K) = \sqrt{\frac{K}{cN}}$. Let

$$\frac{40}{7}c_3\beta^2\frac{r^2(\rho_K)}{\rho_K} \leq \lambda \leq \frac{10}{11}\left(\frac{3}{4}c_6 - 4c_3\right)\beta^2\frac{r^2(\rho_K)}{\rho_K} \quad .$$

Assume that for every $i \in I$, every integer $K \in [K^*, N]$ and every $f \in F \cap B(\rho_K, f^*)$,

$$|(P_i - P)\xi(f - f^*)| \leq c_3\beta^2 \max\left(\|f - f^*\|_{L_P^2}^2, r(\rho_K)^2\right) \ . \tag{13}$$

There exists $\Omega_1(K)$ such that $\mathbb{P}(\Omega_1(K)) \geq 1 - 2\exp(-\frac{1}{2}\min(c_1^2 N, c_7^2 K))$, where the estimators $\hat{f}_{K,\lambda}^{(j)}$ for $j = 1, 2, 3$ defined in Section 4.4 with $\eta = \tau/\sqrt{c_1}$ and $\gamma = \frac{3c_3}{\sqrt{c_1}}\tau\beta$ satisfy

1. $\left\|\hat{f}_{K,\lambda}^{(1)} - f^*\right\| \leq \rho_K$
2. $\left\|\hat{f}_{K,\lambda}^{(2)} - f^*\right\| \leq \rho_K$ *and* $\left\|\hat{f}_{K,\lambda}^{(2)} - f^*\right\|_{L_P^2} \leq \frac{\tau}{\beta\sqrt{c_1}} r(\rho_K)$
3. $\left\|\hat{f}_{K,\lambda}^{(3)} - f^*\right\| \leq \rho_K$, $\left\|\hat{f}_{K,\lambda}^{(3)} - f^*\right\|_{L_P^2} \leq \frac{\tau}{\beta\sqrt{c_1}} r(\rho_K)$ *and* $R(\hat{f}_{K,\lambda}^{(3)}) \leq R(f^*) + C_4 r^2(\rho_K)$ *where* $C_4 = (\frac{\tau}{\beta\sqrt{c_1}} + 2c_3\beta^2 + \frac{10c_3}{\sqrt{c_1}}\beta\tau)$.

One cannot hope to get any result if the probability distributions $\{P_i\}_{i\in I}$ are not related to the probability distribution of the "new" data (X, Y), which is P. As mentioned, Assumption 1 is automatically satisfied in the i.i.d. case; this is also the case for assumption (13). In the i.i.d. setup one just may forget about those two assumptions.

Our aim is to go beyond the i.i.d. setup and the standard strong subgaussian assumptions by designing procedures having the same statistical properties as in the i.i.d. setup when the data satisfy the Gaussian regression model (that is for a subgaussian class of functions and a subgaussian noise ξ independent of X). To that end, we relax the i.d. assumptions by saying that, for most data (those with $i \in I$), the $L_{P_i}^2$ moments of the functions in F are close to those for the distribution P, and that the noise is not two weirdly correlated to the design.

More precisely, our aim is to estimate in L_P^2 the oracle f^* which is defined as the best L_P^2 approximation of Y in F from data distributed according to the probability distribution given by the P_i's. Note that the best $L_{P_i}^2$ approximation of Y_i in F may be different from f^*. It seems unavoidable to assume that for most data, these oracles are not too far from each other, which is the role of Assumption (13). Finally, note that Assumption (13) is used only to prove the exact oracle inequality for $\hat{f}_{K,\lambda}^{(3)}$.

Example (continued): ℓ_1^d regularization

Let us compute explicit value of the radius ρ_K and the associated regularization parameter $\lambda \sim r^2(\rho_K)/\rho_K$ in the ℓ_1^d-regularization case. Let $K \in [N]$ and denote by $\sigma = \|\xi\|_{L_{q_0}}$. We have $K = cr^2(\rho_K)N$ when

$$\rho_K \sim_{L,q_0} \frac{K}{\sigma}\sqrt{\frac{1}{N}\log^{-1}\left(\frac{\sigma^2 d}{K}\right)}. \tag{14}$$

Therefore, the regularization parameter should be taken as

$$\lambda \sim \frac{r^2(\rho_K)}{\rho_K} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma d}{\rho_K\sqrt{N}}\right)} \sim_{L,q_0} \sigma\sqrt{\frac{1}{N}\log\left(\frac{e\sigma^2 d}{K}\right)} \ . \tag{15}$$

As usual the regularization parameter depends on the "variance" of the "noise" which is played here by the L_{q_0}-norm of ξ. This parameter is not known in general. In practice, one may use a cross-validation approach to fit the regularization parameter. In theory, one may construct an estimator $\hat{\sigma}$ of σ as in [20, Sections 5.4 and 5.6.2] and simply replace σ by this estimator in the regularization parameter.

5.2 Adaptive choice of K by Lepski's method

The main drawback of Theorem 1 is that we only get the optimal rates of convergence for our estimators when we can choose $K \approx K^*$, which is unknown in general. In this section, we use a Lepski method to select K.

For any integer $K \in [cr(\rho^*)^2 N, N]$, let ρ_K and λ be defined as in Theorem 1 and for every $j = 1, 2, 3$ denote by $\hat{f}_K^{(j)} = \hat{f}_{K,\lambda}^{(j)}$ for this choice of λ. These estimators are the building blocks for the following confidence sets. Let $f \in F$. The empirical "balls" centered in f associated to $L_{2,N}$ is defined by

$$\hat{B}_K^{L_{2,N}}(f) = \left\{g \in F : \|g - f\|_{L_{2,N}} \le \mu r(\rho_K)\right\}$$

and the one associated to d_K is defined by

$$\hat{B}_K^{d_K}(f) = \{g \in F : d_K(g, f) \ge -\nu r^2(\rho_K)\},$$

where μ and ν will be set in Theorem 2. The notion of proximity w.r.t. d_K used here to define $\hat{B}_K^{d_K}(f)$ is somehow the opposite of the one used to criteria $C^{(3)}(\cdot)$ and the associated estimator $\hat{f}_K^{(3)}$. Now, let

$$R_K^{(1)} = B(\hat{f}_K^{(1)}, \rho_K), \quad R_K^{(2)} = B(\hat{f}_K^{(2)}, \rho_K) \cap \hat{B}_K^{L_{2,N}}(\hat{f}_K^{(2)}) \text{ and } R_K^{(3)} = B(\hat{f}_K^{(3)}, \rho_K) \cap \hat{B}_K^{L_{2,N}}(\hat{f}_K^{(3)}) \cap \hat{B}_K^{d_K}(\hat{f}_K^{(3)})$$

and for every $j = 1, 2, 3$, let

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

Finally, define adaptive (to K) estimators via the Lepski's method: for $j = 1, 2, 3$, $\hat{f}_{LE}^{(j)} \in \bigcap_{J=\hat{K}^{(j)}}^{N} R_J^{(j)}$.

Theorem 2. *Grant the assumptions and notations of Theorem 1. There exist an absolute constant c_0 such that Lepski's estimators $\hat{f}_{LE}^{(j)}$ for $j = 1, 2, 3$ defined for $\mu = \frac{\tau}{\sqrt{c_1}}$, $\nu = \frac{3c_3}{\sqrt{c_1}}\beta\tau$ satisfy for every $K \in [K^*, N]$, with probability $1 - c_0 e^{-K/c_0}$,*

1. $\left\|\hat{f}_{LE}^{(1)} - f^*\right\| \le 2\rho_K$
2. $\left\|\hat{f}_{LE}^{(2)} - f^*\right\| \le 2\rho_K$ *and* $\left\|\hat{f}_{LE}^{(2)} - f^*\right\|_{L_P^2} \le \left(1 + \frac{1}{\beta}\right) \frac{\tau}{\sqrt{c_1}} r(2\rho_K)$
3. *if, denoting by $\zeta = \left(1 + \frac{1}{\beta}\right)\frac{\tau}{\sqrt{c_1}}$, for every $i \in I$, every $K \in [K^*, N]$ and every $f, g \in F \cap B(f^*, 2\rho_K) \cap B_2(f^*, \zeta r(2\rho_K))$,*
$$|P_i(Y - g)(f - g) - P(Y - g)(f - g)| \le \alpha_c' r(2\rho_K)^2 \ , \qquad (16)$$
then $\left\|\hat{f}_{LE}^{(3)} - f^*\right\| \le 2\rho_K$, $\left\|\hat{f}_{LE}^{(3)} - f^*\right\|_{L_P^2} \le \zeta r(2\rho_K)$ and $R(\hat{f}_{LE}^{(3)}) \le R(f^*) + C_5 r^2(\rho_K)$, for some $C_5 > 0$ independent from K and N.

In particular, if $K^* \ge K_o/c_2$ and $K = K^*$, with probability $1 - c_0 e^{-cr^2(\rho_{K^*})N/c_0}$,

$$\left\|\hat{f}_{LE}^{(3)} - f^*\right\| \le 2\rho_{K^*}, \left\|\hat{f}_{LE}^{(3)} - f^*\right\|_{L_P^2} \le \zeta r(2\rho_{K^*}) \text{ and } R(\hat{f}_{LE}^{(3)}) \le R(f^*) + C_5 r^2(2\rho_{K^*}).$$

To comment Theorem 2, recall an optimality result from [27]. Assume that all (X_i, Y_i) are distributed according to (X, Y^{f^*}), where $f^* \in F$ and $Y^{f^*} = f^*(X) + \xi$ and ξ is a centered Gaussian variable with variance σ independent of X. Assume that F is L-subgaussian i.e. that, for every $f \in F$: $\|f\|_{L^p} \le L\sqrt{p}\,\|f\|_{L^2}$ for every $p \ge 2$. Then, [27, Theorem A′] proves that, for every $r > 0$, if $\tilde{f}_N$ is a procedure such that, for every $f^* \in F$, with probability at least $1 - c_0 \exp(-\sigma^{-1} r^2 N / c_0)$, $\left\| \tilde{f}_N - f^* \right\|_{L_P^2} \le \zeta_N$, then, necessarily, its rate of convergence ζ_N must satisfy

$$\zeta_N \gtrsim \min\left(r, \operatorname{diam}(F, L_P^2)\right) . \tag{17}$$

In the case where the statistical model $Y^{f^*} = f^*(X) + \xi$ is true, $c \sim 1/C_\xi \sim 1/\sigma$. As a consequence, [27, Theorem A′] with $r = r(\rho_K)$ for some $K \ge K^*$ shows that no procedure can estimate f^* in L_P^2 uniformly over F with confidence at least $1 - c_0 \exp(-K/c_0)$ at better rate than $r(\rho_K)$ (we implicitly assumed that $r(\rho_K) \le \operatorname{diam}(F, L_P^2)$ since $r(\rho_K)$ can obviously be replaced by $r(\rho_K) \wedge \operatorname{diam}(F, L_P^2)$ in all the results).

Moreover, [27, Theorem A] also shows that the ERM over $\rho_K B$, $\hat{f}_N^{ERM} \in \operatorname{argmin}_{f \in \rho_K B} P_N \ell_f$, satisfies, with probability at least $1 - c_0 \exp(-\sigma^{-1} r^2(\rho_K) N / c_0)$, $\left\| \hat{f}_N^{ERM} - f^* \right\|_{L_2} \lesssim r(\rho_K)$ when $\sigma \gtrsim r_Q(\rho_K)$. This proves that, when the noise level is non trivial (that is $\sigma \gtrsim r_Q(\rho_K)$), the ERM is a minimax procedure over $\rho_K B$ in the exponentially high confidence regime.

Therefore, Theorem 2 shows that the procedure $\hat{f}_{LE}$ achieves the same rate of convergence with the same exponentially high confidence as a minimax estimator in the Gaussian regression model (with independent noise) whereas Theorem 2 grants only a very weak stochastic assumption (a simple L_4/L_2 assumption is enough, for instance), no statistical model like the Gaussian regression model with independent noise $Y^{f^*} = f^*(X) + \xi$ is assumed, the data may not be i.i.d. and even contain a few outliers.

Recall that we have seen in Section 3.2 that the $\hat{f}_{K,\lambda}$ are champions of a MOM's tournament in the sense of Lugosi and Mendelson [32]. In this sense, their result is stronger than ours since they prove that *any* champion has a properly controlled risk if K is adequately chosen. On the other hand, the main advantage of our approach is that the *definition* of Birgé's estimator does not require the computation of a theoretical upper bound on the radius of the oracle. Moreover, if the upper bound is pessimistic, as might be the case because they have to consider a supremum over F to get rid of the dependence in f^* of their original bound, then their control of the risk will be pessimistic too. Moreover, 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 rate of convergence.

Example (continued): ℓ_1^d regularization

Finally, let us turn to the main result concerning the ℓ_1^d regularization example. The following result follows from Theorem 2 together with the computation of ρ^* and $r(\cdot)$ for this particular example.

Theorem 3. *We assume that the design vector X and the noise $\xi = Y - \langle X, t^* \rangle$ satisfy the following: there exist $c_1, u > 3/4, \beta, C_\xi$ such that, for every $t \in \mathbb{R}^d$,*

1. *$\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$, $\mathbb{P}[|\langle X, t\rangle| \ge \beta \|t\|_2] \ge u$ and for every $1 \le j \le d$, $\left\|\langle X, e_j\rangle\right\|_{L_p} \le L\sqrt{p}$ for every $1 \le p \le C_0 \log d$ where $(e_1, \dots, e_d)$ is the canonical basis of $\mathbb{R}^d$ and C_0 is an absolute constant,*
2. *$\sigma = \|\xi\|_{L_{q_0}} < \infty$ for some $q_0 > 2$ and $\sqrt{\operatorname{var}\left(\xi \langle X, t\rangle\right)} \le C_\xi \|t\|_2$*
3. *there exists $s \in [N]$ such that $N \ge c_1 s \log(ed/s)$ and for $\rho^* \sim \sigma s \sqrt{\log(ed/s)/N}$, there exists $v \in \mathbb{R}^d$ for which $\|t^* - v\|_1 \le \rho^*/20$ and $|\operatorname{supp}(v)| \le s$.*

Then, there exists a constant c_0 depending only on $c_1, C_0, u > 3/4, \beta, C_\xi$ such that with probability at least $1 - c_0 \exp\left(-\sigma s \log(ed/s)/c_0\right)$, the estimator $\hat{t}_{LE}$ built using Lepski's method on the family of estimator $(\hat{t}_K^{(3)})_{K \in [N]}$ satisfies, for every $1 \le p \le 2$,

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

Proof. It follows from Theorem 2, the computation of $r(\cdot)$ in (11) and ρ_K in (14) that with probability at least $1-c_0\exp(-cr^2(\rho_{K^*})N/c_0)$, $\left\|\hat{t}_{LE}-t^*\right\|_1 \leq \rho_{K^*}$ and $\left\|\hat{t}_{LE}-t^*\right\|_2 \lesssim r(\rho^*)$. Then, the result follows since $\rho_{K^*} \sim \rho^* \sim_{L,q_0} \|\xi\|_{L_{q_0}} s\sqrt{\frac{1}{N}\log\left(\frac{ed}{s}\right)}$ and because of the interpolation inequality $\|v\|_p \leq \|v\|_1^{-1+2/p}\|v\|_2^{2-2/p}$ for all $v \in \mathbb{R}^d$ and $1 \leq p \leq 2$.

■

Theorem 3 is a sub-Gaussian deviation bound obtained under a limited moment assumption and no independence between the noise and the design. The noise ξ only needs to have q_0 moments for some $q_0 > 2$, which is almost the minimal assumption one needs to get a L_2-estimation result. Moreover, the rate of convergence is the minimax one, that is we really get the $\sqrt{\log(ed/s)/N}$ rate of convergence and not only the "classical" $\sqrt{(\log d)/N}$ which is usually found in the literature on the LASSO even under stronger assumption like a statistical model with subgaussian design and independent gaussian noise.

6 Proofs

The proof of Theorem 1 follows from the next three lemmas.

Lemma 2. *Grant the conditions of Theorem 1. There exists an event $\Omega_1(K)$ such that $\mathbb{P}(\Omega_1(K)) \geq 1-2\exp(-\frac{1}{2}\min(c_1^2N, c_7^2K))$, where*

$$\mathcal{B}_{K,\lambda}(f^*) \subset F \cap B(f^*, \rho_K) \cap B_2(f^*, r(\rho_K)) \ ,$$

in particular, on $\Omega_1(K)$, $C^{(1)}_{K,\lambda}(f^*) \leq \rho_K$.

Lemma 3. *Grant the conditions of Theorem 1. Let $\rho > 0$. For any x such that $\frac{|I|}{N}(u-x-c_1) \geq \frac{1}{2}$, then, for every $f, g \in F \cap B(\rho, f^*)$ such that $\|f-g\|_{L_P^2} \geq r_Q(\rho)$, there exists an event $\Omega_Q(\rho, x)$ such that $\mathbb{P}(\Omega_{Q,2}(\rho,x)) \geq 1-e^{-\frac{x^2}{2}N}$ where $\|f-g\|_{L_{2,N}} \geq \beta\|f-g\|_{L_P^2}$.*

For any $A > 1$ and $x > 0$ such that $\frac{|I|}{N}\left(1-\frac{1}{A^2}-x-\frac{2c_1\beta}{A\tau}\right) \geq \frac{1}{2}$, there exists an event $\Omega_{Q,2}(\rho,x)$ such that $\mathbb{P}(\Omega_{Q,2}(\rho,x)) \geq 1-e^{-\frac{x^2}{2}N}$ where, $\forall (f,g) \in F \cap B(f^*, \rho_K) : \|f-g\|_{L_{2,N}} \leq A\tau(\|f-g\|_{L_P^2} \vee r_Q(\rho))$. For example, if $A = \frac{1}{\sqrt{c_1}}$, $x = c_1$ and $\beta \leq 1/8$, with probability larger than $1-2e^{-\frac{c_1^2}{2}N}$, for any $(f,g) \in F \cap B(f^*, \rho_K)$,

$$\|f-g\|_{L_{2,N}} \leq \frac{\tau}{\sqrt{c_1}}(\|f-g\|_{L_P^2} \vee r_Q(\rho_K)) \text{ and, if } \|f-g\|_{L_P^2} \geq r_Q(\rho_K),\ \|f-g\|_{L_{2,N}} \geq \beta\|f-g\|_{L_P^2} \ . \tag{18}$$

In particular, on $\Omega_2(K) = \Omega_Q(\rho_K, c_1) \cap \Omega_{Q,2}(\rho_K, c_1)$,

$$\begin{aligned}\{F \cap B(f^*, \rho_K) \cap B_2(f^*, r(\rho_K))\} &\subset \{B(f^*, \rho_K) \cap \{g \in F : \|g-f^*\|_{2,N} \leq \frac{\tau}{\sqrt{c_1}}r(\rho_K)\}\} \\ &\subset \{F \cap B(f^*, \rho_K) \cap B_2(f^*, \frac{\tau}{\sqrt{c_1}\beta}r(\rho_K))\} \ .\end{aligned}$$

Lemma 4. *Grant the conditions of Theorem 1. On the event $\Omega_1(\rho_K)$ of Lemma 2, for every $f \in F \cap B(f^*, \rho_K) \cap B_2(f^*, \frac{\tau}{\sqrt{c_1}}r(\rho_K))$,*

1. *$d_K(f, f^*) \leq \frac{3c_3}{\sqrt{c_1}}\beta\tau r^2(\rho_K)$*

2. *$Q^{\bar{P}}_{3/4,K}|(Y-f^*)(f^*-f)| \leq \frac{2c_3}{\sqrt{c_1}}\beta\tau r^2(\rho_K)$, where, for any function $G : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$*

$$Q^{\bar{P}}_{3/4,K}|G(X,Y)| \in \mathcal{Q}_\alpha([|(P_{B_k} - \overline{P}_{B_k})G|]_{k\in[K]}) \qquad \text{and} \qquad \overline{P}_{B_k}G = \frac{1}{|B_k|}\sum_{i\in B_k} P_iG \ .$$

In the proof, λ is fixed according to Theorem 1 and the dependence in λ of the estimators is omitted.
Proof of point *1.* of Theorem 1: On $\Omega_1(K)$ defined in Lemma 2, $C^{(1)}(f^*) \le \rho_K$ so $C^{(1)}(\hat{f}_K^{(1)}) \le \rho_K$, thus $\|\hat{f}_K^{(1)} - f^*\| \le C^{(1)}(f^*) \vee C^{(1)}(\hat{f}_K^{(1)}) \le \rho_K$. ■

Proof of point *2.* of Theorem 1: Let $\Omega_1(K)$ and $\Omega_2(K)$ denote the events defined in Lemmas 2 and 3 respectively. On $\Omega_1(K) \cap \Omega_2(K)$,

$$\mathcal{B}_{K,\lambda}(f^*) \subset F \cap B(f^*, \rho_K) \cap B_2(f^*, r(\rho_K)) \subset \{B(f^*, \rho_K) \cap \{g \in F : \|g - f^*\|_{2,N} \le \frac{\tau}{\sqrt{c_1}} r(\rho_K)\}\} \enspace ,$$

therefore $C^{(2)}_{K,\lambda}(f^*) \le \rho_K$ and by definition of $\hat{f}_K^{(2)}$, $C_{K,\lambda}(\hat{f}_K^{(2)}) \le \rho_K$. Thus

$$\hat{f}_K^{(2)} \in \{B(f^*, \rho_K) \cap \{g \in F : \|g - f^*\|_{2,N} \le \frac{\tau}{\sqrt{c_1}} r(\rho_K)\}\} \subset \{F \cap B(f^*, \rho_K) \cap B_2(f^*, \frac{\tau}{\beta\sqrt{c_1}} r(\rho_K))\} \enspace .$$

In other words $\|\hat{f}_K^{(2)} - f^*\| \le \rho_K$ and $\|\hat{f}_K^{(2)} - f^*\|_{L^2_P} \le \frac{\tau}{\beta\sqrt{c_1}} r(\rho_K)$. ■

Proof of point *3.* of Theorem 1: Let $\Omega_1(K)$, $\Omega_2(K)$ denote the events defined respectively in Lemmas 2 and 3. On $\Omega_1(K) \cap \Omega_2(K)$, by point *1.* in Lemma 4, $C^{(3)}(f^*) \le \rho_K$. As seen in the proof of point *2.* of Theorem 1, we have thus $\|\hat{f}_K^{(3)} - f^*\| \le \rho_K \quad$ and $\quad \|\hat{f}_K^{(3)} - f^*\|_{L^2_P} \le \frac{\tau}{\beta\sqrt{c_1}} r(\rho_K)$.

Let us now turn to the proof of the oracle inequality. We have

$$R(\hat{f}_K^{(3)}) - R(f^*) = \left\| \hat{f}_K^{(3)} - f^* \right\|_{L^2_P}^2 + 2P(Y - f^*)(f^* - \hat{f}_K^{(3)}) \le \frac{\tau}{\beta\sqrt{c_1}} r^2(\rho_K) + 2P(Y - f^*)(f^* - \hat{f}_K^{(3)}) \enspace .$$

From Assumption (13), for every $f \in F \cap B(f^*, \rho_K) \cap B_2(f^*, \frac{\tau}{\beta\sqrt{c_1}} r(\rho_K))$, $\forall k \in [K]$ such that $B_k \cap \mathcal{O} = \emptyset$,

$$(P - \bar{P}_{B_k})[\xi(f^* - f)] = \frac{1}{|B_k|} \sum_{i \in B_k} (P - P_i)[\xi(f^* - f)] \le c_3 \beta^2 r^2(\rho_K) \enspace .$$

Thus, for any $K \ge 4K_o$,

$$\begin{aligned}
R(\hat{f}_K^{(3)}) - R(f^*) &\le (\frac{\tau}{\beta\sqrt{c_1}} + 2c_3\beta^2) r^2(\rho_K) + 2Q_{1/4}([\bar{P}_{B_k}[\xi(f^* - \hat{f}_K^{(3)})]]_{k \in [K]}) \\
&\le (\frac{\tau}{\beta\sqrt{c_1}} + 2c_3\beta^2) r^2(\rho_K) + 2\mathrm{MOM}_K[\xi(f^* - \hat{f}_K^{(3)})] \\
&\quad - 2Q_{1/4}([(\bar{P}_{B_k} - P_{B_K})[\xi(f^* - \hat{f}_K^{(3)})]]_{k \in [K]}) \\
&\le (\frac{\tau}{\beta\sqrt{c_1}} + 2c_3\beta^2) r^2(\rho_K) + 2d_K(f^*, \hat{f}_K^{(3)}) + 2Q_{3/4}^{\overline{P}}(|\xi(f^* - f)|) \\
&\le (\frac{\tau}{\beta\sqrt{c_1}} + 2c_3\beta^2 + \frac{10c_3}{\sqrt{c_1}} \beta\tau) r^2(\rho_K) \enspace .
\end{aligned}$$

■

In Section 6.1, we recall the small ball method of Mendelson by presenting a lemma that will be repeatedly used afterwards in the proofs of Lemmas 2, 3 and 4. Section 6.2, we prove an "empirical small ball" result (Corollary 1) for the quadratic process $P_{B_k}(f - f^*)^2$ and a similar result for the multiplier process $P_{B_k}\xi(f - f^*)$ in Section 6.3. The proofs of Lemmas 2, 3 and 4 are given in Section 6.4, using a strategy similar to the proof of [29, Theorem 3.2]. Finally, Theorem 2 is proved in Section 6.6.

6.1 Mendelson's small ball method (SBM)

We will repeatedly use the following lemma.

Lemma 5. *Let $W_1, \dots, W_n$ denote independent random variables with respective distributions $Q_1, \dots, Q_n$, let G be a class of functions, $b \in (0,1)$ and $h \geq 0$. Let $\mathcal{I} \subset [n]$, let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables, independent of the W_i's and $(\gamma_i)_{i\in\mathcal{I}}$ be nonnegative functions such that*

$$\inf_{i\in\mathcal{I}} Q_i(|g| \geq 2\gamma_i(g)) \geq b \qquad \text{and} \qquad \mathbb{E}\sup_{g\in G}\left|\sum_{i\in\mathcal{I}} \epsilon_i \frac{g(W_i)}{\gamma_i(g)}\right| \leq h|\mathcal{I}| \enspace .$$

Then

$$\forall x \geq 0, \qquad \mathbb{P}\left(|\{i \in [n],\ |g(W_i)| \geq \gamma_i(g)\}| \geq |\mathcal{I}|(b - x - 2h)\right) \geq 1 - e^{-\frac{|\mathcal{I}|}{2}x^2} \enspace . \tag{19}$$

If the first condition is replaced by $\inf_{i\in\mathcal{I}} Q_i(2|g| < \gamma_i(g)) \geq b$, then,

$$\forall x \geq 0, \qquad \mathbb{P}\left(|\{i \in [n],\ |g(W_i)| < \gamma_i(g)\}| \geq |\mathcal{I}|(b - x - 4h)\right) \geq 1 - e^{-\frac{|\mathcal{I}|}{2}x^2} \enspace .$$

Proof. Define ϕ for all $t \geq 0$ by

$$\phi(t) = \begin{cases} 0 & \text{if } 0 \leq t \leq 1 \\ t-1 & \text{if } 1 \leq t \leq 2 \\ 1 & \text{otherwise.} \end{cases}$$

Let $F = \{g/\gamma_i(g) : g \in G\}$ and for all $f \in F$, let $Z(f) = \sum_{i=1}^n I(|f(W_i)| \geq 1)$,

$$P_{\mathcal{I}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \delta_{W_i} \text{ and } \overline{P}_{\mathcal{I}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} Q_i.$$

For all $f \in F$, we have

$$\begin{aligned} Z(f) &\geq \sum_{i\in\mathcal{I}} I(|f(W_i)| \geq 1) = |\mathcal{I}|\left(\overline{P}_{\mathcal{I}} I(|f| \geq 2) + P_{\mathcal{I}} I(|f| \geq 1) - \overline{P}_{\mathcal{I}} I(|f| \geq 2)\right) \\ &\geq |\mathcal{I}|\left(b - (P_{\mathcal{I}} - \overline{P}_{\mathcal{I}})\phi(|f|)\right) \geq |\mathcal{I}|\left(b - \sup_{f\in F}(P_{\mathcal{I}} - \overline{P}_{\mathcal{I}})\phi(|f|)\right) \end{aligned}$$

By the bounded difference inequality (see, for instance [12, Theorem 6.2]), for any $x > 0$, with probability larger than $1 - e^{-\frac{|\mathcal{I}|}{2}x^2}$,

$$\sup_{f\in F}(P_{\mathcal{I}} - \overline{P}_{\mathcal{I}})\phi(|f|) \leq \mathbb{E}\sup_{f\in F}(P_{\mathcal{I}} - \overline{P}_{\mathcal{I}})\phi(|f|) + x \enspace .$$

Denote by $P_{\mathcal{I},\epsilon} = |\mathcal{I}|^{-1}\sum_{i\in\mathcal{I}} \epsilon_i\delta_{W_i}$ the symmetrized empirical measure then, by the symmetrization argument,

$$\mathbb{E}\sup_{f\in F}(P_{\mathcal{I}} - \overline{P}_{\mathcal{I}})\phi(|f|) \leq 2\mathbb{E}\sup_{f\in F} P_{\mathcal{I},\epsilon}\phi(|f|) \enspace .$$

Note that ϕ is Lipschitz with Lipschitz constant equal to one and $\phi(0) = 0$. By the contraction principle (see, for example [30, Chapter 4] or [12, Theorem 11.6])

$$\mathbb{E}\sup_{f\in F} P_{\mathcal{I},\epsilon}\phi(|f|) \leq \mathbb{E}\sup_{f\in F} P_{\mathcal{I},\epsilon}|f| = \mathbb{E}\sup_{g\in G}\left|\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \epsilon_i \frac{|g(W_i)|}{\gamma_i(g)}\right| \leq h \enspace .$$

Gathering these results yields (19).

For the last claim, observe that, for all $f \in F$,

$$\sum_{i=1}^n I\left(|f(W_i)| < 1\right) \geq \sum_{i\in\mathcal{I}} I\left(|f(W_i)| < 1\right) \geq |\mathcal{I}| - \sum_{i\in\mathcal{I}} I\left(|f(W_i)| \geq 1\right)$$

and

$$\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} I\left(|f(W_i)|\geq 1\right) = P_{\mathcal{I}} I\left(|f|\geq 1\right) - \overline{P}_{\mathcal{I}}\left(|f|\geq 1/2\right) + \overline{P}_{\mathcal{I}}\left(|f|\geq 1/2\right) \leq 1-b+\left(P_{\mathcal{I}}-\overline{P}_{\mathcal{I}}\right)\psi(|f|)$$

where ψ is defined for all $t\geq 0$ by

$$\psi(t)=\begin{cases} 0 & \text{if } 0\leq t\leq 1/2 \\ 2t-1 & \text{if } 1/2\leq t\leq 1 \\ 1 & \text{otherwise.}\end{cases}$$

Using the same arguments as in the first part of the proof (note however that ψ is 2-Lipschitz), we obtain that for all $x>0$, with probability larger than $1-e^{-\frac{|\mathcal{I}|}{2}x^2}$,

$$\sup_{f\in F}\left(P_{\mathcal{I}}-\overline{P}_{\mathcal{I}}\right)\psi(|f|)\leq 4h+x.$$

This concludes the proof.

■

6.2 Control of the quadratic process via the SBM

The following result is a slight extension of a result in [23] that we reproduce to highlight the small ball method and its robustness properties.

Theorem 4 ([23])**.** *Assume that F satisfies Assumption 3 (small-ball condition) with constants $0<\beta<1$ and $0<u\leq 1$. Let $\rho>0$, then for any $x\geq 0$, on an event $\Omega_Q(\rho,x)$ satisfying $\mathbb{P}(\Omega_Q(\rho,x))\geq 1-e^{-\frac{|I|}{2}x^2}$, for every $f,g\in F\cap B(\rho,f^*)$ such that $\|f-g\|_{L_P^2}\geq r_Q(\rho)$,*

$$\left|\left\{i\in[N]:|(f-g)(X_i)|\geq\beta\left\|f-g\right\|_{L_P^2}\right\}\right|\geq |I|(u-x-\frac{4\gamma_Q}{\beta}) \enspace .$$

In particular, if $\frac{|I|}{N}(u-x-\frac{4\gamma_Q}{\beta})\geq\frac{1}{2}$, then, for every $f,g\in F\cap B(\rho,f^*)$ such that $\|f-g\|_{L_P^2}\geq r_Q(\rho)$

$$\|f-g\|_{L_{2,N}}\geq\beta\left\|f-g\right\|_{L_P^2}$$

Proof. The proof follows from Lemma 5 for $n=N$, $G=\{f-g:f,g\in F\cap B(\rho,f^*),\|f-g\|_{L_P^2}\geq r_Q(\rho)\}$, $W_i=X_i$ for any $i\in[N]$, $\mathcal{I}=I$ and, for any $g\in G$, $\gamma_i(g)=\beta\left\|g\right\|_{L_P^2}/2$.

It follows from the small ball property that Condition 1 of Lemma 5 is satisfied with $b=u$. Therefore, it only remains to bound the expectation in Condition 2 of Lemma 5. We have

$$\mathbb{E}\sup_{g\in G}\left|\sum_{i\in I}\epsilon_i\frac{g(W_i)}{\gamma_i(g)}\right| = \mathbb{E}\sup_{\substack{f,g\in F\cap B(\rho,f^*)\\ \|f-g\|_{L_P^2}\geq r_Q(\rho)}}\left|\sum_{i\in I}\epsilon_i\frac{(f-g)(X_i)}{\gamma_i(f-g)}\right| = \mathbb{E}\sup_{\substack{f,g\in F\cap B(\rho,f^*)\\ \|f-g\|_{L_P^2}=r_Q(\rho)}}\left|\sum_{i\in I}\epsilon_i\frac{(f-g)(X_i)}{\gamma_i(f-g)}\right|$$

$$\leq\frac{2}{\beta r_Q(\rho)}\mathbb{E}\sup_{\substack{f,g\in F\cap B(\rho,f^*)\\ \|f-g\|_{L_P^2}=r_Q(\rho)}}\left|\sum_{i\in I}\epsilon_i(f-g)(X_i)\right|\leq\frac{2\gamma_Q}{\beta}|I| \enspace .$$

Therefore, Condition 2 of Lemma 5 holds with $h=\frac{2\gamma_Q}{\beta}$ and the result follows. ■

Corollary 1. *On the event $\Omega_Q(\rho, x)$ of Theorem 4, for any $f, g \in F \cap B(\rho, f^*)$ such that $\|f-g\|_{L_P^2} \geq r_Q(\rho)$, one has*

$$\forall \delta \in (0,1), \qquad \left|\left\{k \in [K] : P_{B_k}(f-g)^2 \geq \delta\beta^2 (u - x - \frac{4\gamma_Q}{\beta})\frac{|I|}{N} \|f-g\|_{L_P^2}^2\right\}\right| \geq (1-\delta)(u-x-\frac{4\gamma_Q}{\beta})\frac{|I|}{N}K \ .$$

In particular, if $\delta = c_6$, $x = c_1$, $\gamma_Q = c_1\beta$ and $\frac{|I|}{N} \geq 1 - \frac{K_o}{K} \geq 1 - c_2$ as in Theorem 1, then $(1-\delta)(u - x - \frac{4\gamma_Q}{\beta})\frac{|I|}{N} \geq 3/4$, so, on $\Omega_Q(\rho, c_1)$, for any $f, g \in F \cap B(\rho, f^*)$ such that $\|f-g\|_{L_P^2} \geq r_Q(\rho)$, one has

$$Q_{1/4,K}[(f-g)^2] \geq \frac{3}{4}c_6\beta^2 \|f-g\|_{L_P^2}^2 \ .$$

Proof. Denote by $J = \{i \in [N] : |(f-g)(X_i)| \geq \beta \|f-g\|_{L_P^2}\}$. On the event $\Omega_Q(\rho, x)$, $|J| \geq \alpha N$, with $\alpha = (u - x - \frac{4\gamma_Q}{\beta})\frac{|I|}{N}$. If there were less than $(1-\delta)\alpha K$ blocks containing more than $\delta\alpha\frac{N}{K}$ elements in J, there would be at most

$$(1-\delta)\alpha K\frac{N}{K} + (1-(1-\delta)\alpha)K\delta\alpha\frac{N}{K} < \alpha N$$

elements in J. Therefore, there must be at least $(1-\delta)\alpha K$ blocks B_k containing more than $\delta\alpha\frac{N}{K}$ elements in J. For these blocks, one has

$$P_{B_k}(f-g)^2 \geq \frac{1}{|B_k|}\sum_{i \in J \cap B_k}(f-g)^2(X_i) \geq \frac{1}{|B_k|}\delta\alpha\frac{N}{K}\beta^2\|f-g\|_{L_2(P)}^2 = \delta\alpha\beta^2\|f-g\|_{L_2(P)}^2 \ .$$

■

Lemma 6. *Assume that property $R(\tau)$ holds (see Assumption 1) for some $\tau \geq 1$ and let $A > 1$. For any $x \geq 0$ and $\rho > 0$, there exists an event $\Omega_{Q,2}(\rho, x)$ satisfying $\mathbb{P}(\Omega_{Q,2}(\rho, x)) \geq 1 - \exp(-\frac{x^2}{2}|I|)$, where, for every $f, g \in F \cap B(f^*, \rho)$, one has*

$$\left|\left\{i \in [N] : |(f-g)(X_i)| \leq A\tau(\|f-g\|_{L_P^2} \vee r_Q(\rho))\right\}\right| \geq |I|\left(1 - \frac{1}{A^2} - x - \frac{16\gamma_Q}{A\tau}\right) \ .$$

In particular, if $\frac{|I|}{N}\left(1 - \frac{1}{A^2} - x - \frac{4\gamma_Q}{A\tau}\right) \geq \frac{1}{2}$, on $\Omega_{Q,2}(\rho, x)$, $\|f-g\|_{L_{2,N}} \leq A\tau(\|f-g\|_{L_P^2} \vee r_Q(\rho))$.

Proof. We apply Lemma 5 to $n = N$, $G = \{f - g : f, g \in F \cap B(f^*, \rho)\}$, $W_i = X_i$ for $i \in [N]$ and $\mathcal{I} = I$. Define, for any $g \in G$, $\gamma_i(g) = (A\tau/2)(\|g\|_{L_P^2} \vee r_Q(\rho))$. It follows from property $R(\tau)$ that, for every $i \in I$, $\gamma_i(g) \geq (A/2)\|g\|_{L_{P_i}^2}$ and, by Markov's inequality, $\inf_{i \in I} P_i(2|g| < \gamma_i(g)) \geq 1 - \frac{1}{A^2}$. Therefore Condition 1 (bis) of Lemma 5 holds with $b = 1 - A^{-2}$.

From the convexity of F and the definition of $r_Q(\rho)$,

$$\mathbb{E}\sup_{g \in G}\left|\sum_{i \in I}\epsilon_i\frac{g(W_i)}{\gamma_i(g)}\right| \leq \frac{4}{A\tau r_Q(\rho)}\mathbb{E}\left(\sup_{\substack{f,g \in F \cap B(\rho, f^*) \\ \|f-g\|_{L_2(P)} = r_Q(\rho)}}\sum_{i \in I}\epsilon_i(f-g)(X_i)\right) \leq \frac{4\gamma_Q}{A\tau}|I| \ .$$

Therefore, Condition 2. of Lemma 5 holds with $h = \frac{4\gamma_Q}{A\tau}$ and Lemma 6 is proved. ■

6.3 Control of the multiplier process via the SBM

In this section, we provide bounds on the multiplier process via an adaptation of Mendelson's (SBP). The SBP is usually used to control the quadratic process (or any process with nonnegative terms) as in Theorem 4. In this section, we show that SBP can be adapted to control also the multiplier process.

Theorem 5. *Suppose that the noise ξ satisfies Assumption 2. For any $\rho > 0$, $x \geq 0$, $K \geq 2K_o$, there exists an event $\Omega_M(\rho, x)$ such that $\mathbb{P}(\Omega_M(\rho,x)) \geq 1 - e^{-\frac{x^2}{2}(K-K_o)}$ where, $\forall f \in F \cap B(\rho, f^*)$, $\forall A_2 > 0$,*

$$\left|\left\{k \in \mathcal{K} : \left|\frac{2}{|B_k|}\sum_{i\in B_k}\left(\xi_i(f-f^*)(X_i) - P_i(\xi(f-f^*))\right)\right| \leq A_2 \max\left(r_M(\rho)^2, \|f-f^*\|_{L_P^2}^2\right)\right\}\right|$$
$$\geq |\mathcal{K}|\left(1 - \frac{4C_\xi K}{A_2^2 N r_M(\rho)^2} - x - \frac{64\gamma_M}{A_2}\right) =: \alpha K \enspace ,$$

where $\mathcal{K} = \{k \in [K] : B_k \cap \mathcal{O} = \emptyset\}$ denote the set of blocks containing only informative data.

Proof. Let $K \geq 2K_o$. Obviously, by the pigeonhole principle, $|\mathcal{K}| \geq K - K_o$. We apply Lemma 5 to $n = K$, $W_k = ((X_i, Y_i))_{i\in B_k}$ for $k \in [K]$, $F^* = \{f - f^* : f \in F \cap B(\rho, f^*)\}$ and $\mathcal{I} = \mathcal{K}$. For any $f \in F^*$ and $k \in [K]$, define

$$g_f(W_k) = \frac{2}{|B_k|}\sum_{i\in B_k}\left(\xi_i f(X_i) - P_i(\xi f)\right) \text{ and } \gamma_k(g_f) = 2A_2 \max\left(r_M(\rho)^2, \|f\|_{L_P^2}^2\right) \enspace .$$

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

$$\mathbb{P}\left[2\left|g_f(W_k)\right| \geq \gamma_k(g_f)\right] \leq \frac{\mathbb{E}\left[\left(\frac{2}{|B_k|}\sum_{i\in B_k}(\xi_i f(X_i) - P_i(\xi f))\right)^2\right]}{A_2^2 \max(r_M(\rho)^4, \|f\|_{L_P^2}^4)} \leq \frac{4\sum_{i\in B_k}\mathrm{var}_{P_i}(\xi f)}{|B_k|^2 A_2^2 \max(r_M(\rho)^4, \|f\|_{L_P^2}^4)}$$
$$\leq \frac{4C_\xi \|f\|_{L_P^2}^2}{|B_k| A_2^2 \max(r_M(\rho)^4, \|f\|_{L_P^2}^4)} \leq \frac{4C_\xi}{A_2^2}\frac{K}{N r_M(\rho)^2} \enspace .$$

Hence, Condition 1 (bis) of Lemma 5 applies with $b = 1 - \frac{4C_\xi}{A_2^2}\frac{K}{N r_M(\rho)^2}$.

Let $J = \cup_{k\in\mathcal{K}} B_k$, $|J| = |\mathcal{K}|\frac{N}{K} = \frac{K-K_o}{K}N \geq \frac{N}{2}$. By definition,

$$\mathbb{E}\left(\sup_{f\in F^*}\sum_{k\in\mathcal{K}}\epsilon_k \frac{g_f(W_k)}{\gamma_k(g_f)}\right) \leq \frac{4}{A_2}\mathbb{E}\sup_{f\in F^*}\left|\sum_{k\in\mathcal{K}}\frac{\frac{\epsilon_k}{|B_k|}\sum_{i\in B_k}(\xi_i f(X_i) - P_i(\xi f))}{\max\left(r_M(\rho)^2, \|f\|_{L_P^2}^2\right)}\right|$$
$$\leq \frac{4}{A_2 r_M(\rho)^2}\mathbb{E}\sup_{f\in F^*\cap r_M(\rho)B_2}\left|\sum_{k\in\mathcal{K}}\frac{\epsilon_k}{|B_k|}\sum_{i\in B_k}(\xi_i f(X_i) - P_i(\xi f))\right|$$
$$+ \frac{4}{A_2}\mathbb{E}\sup_{f\in F^* \text{ s.t.} \|f\|_{L_P^2} \geq r_M(\rho)}\left|\sum_{k\in\mathcal{K}}\frac{\epsilon_k}{|B_k|}\sum_{i\in B_k}\frac{(\xi_i f(X_i) - P_i(\xi f))}{\|f\|_{L_2(P)}^2}\right| \enspace .$$

Therefore, by convexity of F and the symmetrization argument,

$$\mathbb{E}\left(\sup_{f\in F^*}\sum_{k\in\mathcal{K}}\epsilon_k \frac{g_f(W_k)}{\gamma_k(g_f)}\right) \leq \frac{8K}{A_2 N r_M(\rho)^2}\mathbb{E}\sup_{f\in F^*\cap r_M(\rho)B_2}\left|\sum_{i\in J}\epsilon_{[i]}(\xi_i f(X_i) - P_i \xi f)\right|$$
$$\leq \frac{16K}{A_2 N r_M(\rho)^2}\mathbb{E}\sup_{f\in F^*\cap r_M(\rho)B_2}\left|\sum_{i\in J}\epsilon_i \xi_i f(X_i)\right| \leq \frac{16K\gamma_M |J|}{A_2 N} = \frac{16\gamma_M}{A_2}|\mathcal{K}| \enspace .$$

where, in the first inequality of last line, we used a classical symmetrization argument on the family of independent and centered random variables $(\xi_i f(X_i) - P_i \xi f)_{i=1}^N$. The result now follows from Lemma 5. ■

6.4 Proof of Lemma 2

Reduction to bounds on the quantiles of mean processes. We want to prove that, with large probability, for any f such that $\|f-f^*\|_{L^2_P}\geq r(\rho_K)$ or $\|f-f^*\|\geq\rho_K$,

$$\mathrm{MOM}_K\left[(f-f^*)^2-2\xi(f-f^*)\right]+\lambda(\|f\|-\|f^*\|)\geq 0\ . \tag{20}$$

Using properties (4), (5) and (6), one can lower bound

$$\mathrm{MOM}_K\left[(f-f^*)^2-2\xi(f-f^*)\right]\geq Q_{1/4,K}[(f-f^*)^2]-2Q_{3/4,K}[\xi(f-f^*)]\ . \tag{21}$$

Now, since $\mathcal{K}\subset K$, $Q_{3/4,K}[\xi(f-f^*)]\leq Q_{3/4,\mathcal{K}}[\xi(f-f^*)]$, where $\mathcal{K}=\{k:B_k\cap\mathcal{O}=\emptyset\}$, and, $\forall z\in\mathbb{R}^N$,

$$Q_{3/4,\mathcal{K}}[z]=\inf\{x\in\mathbb{R}:\sum_{k\in\mathcal{K}}I(P_{B_k}z\leq x)\geq\frac{3}{4}K\}\ .$$

Then, by (13), for any $i\in I$, $P_i\xi(f-f^*)\leq c_3\beta^2(\|f-f^*\|^2_{L^2_P}\vee r(\rho_K)^2)$, so

$$\begin{aligned}Q_{3/4,\mathcal{K}}[\xi(f-f^*)]&\leq Q^{\overline{P}}_{3/4,\mathcal{K}}[\xi(f-f^*)]+c_3\beta^2(\|f-f^*\|_{L^2_P}\vee r(\rho_K))\\&\leq Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|]+c_3\beta^2(\|f-f^*\|^2_{L^2_P}\vee r(\rho_K)^2)\ .\end{aligned} \tag{22}$$

It follows from these bounds that (20) holds if

$$Q_{1/4,K}[(f-f^*)^2]-2Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|]+\lambda(\|f\|-\|f^*\|)\geq c_3\beta^2(\|f-f^*\|^2_{L^2_P}\vee r(\rho_K)^2)\ . \tag{23}$$

Conclusion of the proof when the excess risk is large and the regularization distance is small Assume first that $\|f-f^*\|_{L^2_P}\geq r(\rho_K)$ and $\|f-f^*\|\leq\rho_K$, then, by the triangular inequality, $\|f\|-\|f^*\|\geq-\|f-f^*\|\geq-\rho_K$, therefore, (23) holds if

$$Q_{1/4,K}[(f-f^*)^2]-2Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|]\geq\lambda\rho_K+c_3\beta^2\|f-f^*\|^2_{L^2_P}\ . \tag{24}$$

By Corollary 1, if (12) holds, then, on the event $\Omega_Q(\rho_K,c_1)$, since $f\in F\cap B(f^*,\rho_K)$ and $\|f-g\|_{L^2_P}\geq r_Q(\rho_K)$, one has

$$Q_{1/4,K}[(f-g)^2]\geq\frac{3}{4}c_6\beta^2\left\|f-g\right\|^2_{L^2_P}\ .$$

By Theorem 5, for $A_2=c_3\beta^2$ as in Theorem 1, on $\Omega_M(\rho_K,c_7)$,

$$Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|]\leq c_3\beta^2\max\left(r_M(\rho_K)^2,\|f-f^*\|^2_{L^2_P}\right)\leq c_3\beta^2\left\|f-f^*\right\|^2_{L^2_P}\ .$$

Therefore, (24) holds on $\Omega_Q(\rho_K,c_1)\cap\Omega_M(\rho_K,c_7)$ thanks to the upper bound assumption on λ. ■

The homogeneity argument when $\|f-f^*\|$ is large Assume from now on that $\|f-f^*\|\geq\rho_K$. One has for every $f^{**}\in f^*+(\rho_K/20)B$ and every $z^*\in(\partial\left\|\cdot\right\|)_{f^{**}}$,

$$\|f\|-\|f^*\|\geq\|f\|-\|f^{**}\|-\|f^{**}-f^*\|\geq z^*(f-f^{**})-\frac{\rho_K}{20}=z^*(f-f^*)-z^*(f^{**}-f^*)-\frac{\rho_K}{20}\geq z^*(f-f^*)-\frac{\rho_K}{10}\ ,$$

where the last inequality follows from $z^*(f^{**}-f^*)\leq\|f^{**}-f^*\|$. As this holds for any $z^*\in\Gamma_{f^*}(\rho_K)$, (21) holds if

$$Q_{1/4,K}[(f-f^*)^2]-2Q_{3/4,K}[\xi(f-f^*)]+\lambda\sup_{z^*\in\Gamma_{f^*}(\rho_K)}z^*(f-f^*)\geq\lambda\frac{\rho_K}{10}\ . \tag{25}$$

Assume that (25) holds for any $g \in S(f^*, \rho_K)$. Let $g = f^* + \rho_K \frac{f-f^*}{\|f-f^*\|}$ so $g \in S(f^*, \rho_K)$, therefore,

$$Q_{1/4,K}[(g-f^*)^2] - 2Q_{3/4,K}[\xi(g-f^*)] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) \geq \lambda \frac{\rho_K}{10} \ .$$

Hence, if $\kappa = \frac{\|f-f^*\|}{\rho_K} \geq 1$,

$$\begin{aligned} Q_{1/4,K}[(f-f^*)^2] - 2Q_{3/4,K}[\xi(f-f^*)] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \\ &= \kappa^2 Q_{1/4,K}[(g-f^*)^2] - 2\kappa Q_{3/4,K}[\xi(g-f^*)] + \kappa\lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) \\ &\geq \kappa \left(Q_{1/4,K}[(g-f^*)^2] - 2Q_{3/4,K}[\xi(g-f^*)] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) \right) \geq \lambda \frac{\rho_K}{10} \ . \end{aligned}$$

Therefore, (21) holds for any f such that $\|f-f^*\| \geq \rho_K$ if it holds for any $f \in S(f^*, \rho_K)$. Using the bound (22) shows finally that (21) holds for any f such that $\|f-f^*\| \geq \rho_K$ if, for any $f \in S(f^*, \rho_K)$,

$$Q_{1/4,K}[(f-f^*)^2] - 2Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|] + \lambda \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq \lambda \frac{\rho_K}{10} + 2\alpha_c(\|f-f^*\|^2_{L^2_P} \vee r(\rho_K)^2) \ . \quad (26)$$

Conclusion of the proof when $\|f-f^*\|$ is large and $\|f-f^*\|_{L^2_P}$ is small Suppose now that $f \in H_{\rho_K}$, i.e. that $\|f-f^*\| = \rho_K$ and $\|f-f^*\|_{L^P_2} \leq r(\rho_K)$. Then, by definition $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq \Delta(\rho_K)$ and, since $\rho_K \geq \rho^*$, ρ_K satisfies the sparsity equation and thus, $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq \frac{4}{5}\rho_K$. Moreover, $Q_{1/4,K}[(f-f^*)^2] \geq 0$ and by Theorem 5, for $A_2 = c_3\beta^2$, with the parameters set in Theorem 1, on $\Omega_M(\rho_K, c_7)$,

$$Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|] \leq c_3\beta^2 \max\left(r_M(\rho_K)^2, \|f-f^*\|^2_{L^2_P}\right) \leq c_3\beta^2 r(\rho_K)^2 \ .$$

Therefore, (26) holds on $\Omega_M(\rho_K, c_7)$ by the lower bound condition on λ. ■

Conclusion of the proof when both $\|f-f^*\|$ and $\|f-f^*\|_{L^2_P}$ are large Assume finally that $\|f-f^*\| = \rho_K$ and $\|f-f^*\|_{L^2_P} \geq r(\rho_K)$. Then we always have $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq -\|f-f^*\| = -\rho_K$. Moreover, by Corollary 1, by condition (12), on the event $\Omega_Q(\rho_K, c_1)$, since $f \in F \cap B(\rho_K, f^*)$ and $\|f-g\|_{L^2_P} \geq r_Q(\rho_K)$, one has

$$Q_{1/4,K}[(f-g)^2] \geq \frac{3}{4} c_2\beta^2 \|f-g\|^2_{L^2_P}$$

and, by Theorem 5, for $A_2 = c_3\beta^2$, with the parameters set in Theorem 1, on $\Omega_M(\rho_K, c_7)$,

$$Q^{\overline{P}}_{3/4,\mathcal{K}}[|\xi(f-f^*)|] \leq c_3\beta^2 \max\left(r_M(\rho_K)^2, \|f-f^*\|^2_{L^2_P}\right) \leq c_3\beta^2 \|f-f^*\|^2_{L^2_P} \ .$$

Therefore, (26) holds on $\Omega_M(\rho_K, c_7) \cap \Omega_Q(\rho_K, c_1)$ from our upper bound assumption on λ. ■

6.5 Proofs of Lemma 3 and Lemma 4

Proof of Lemma 3: The first part of the Lemma is stated in the conclusion of Theorem 4. The second part is stated in the conclusion of Lemma 6. ■

Proof of Lemma 4: It follows from (22) that, for the parameters set in Theorem 1, on $\Omega_M(\rho_K, c_7)$,

$$Q^{\overline{P}}_{3/4,\mathcal{K}}|\xi(f-f^*)| \leq 2c_3\beta^2 \max\left(r^2_M(\rho_K), \|f-f^*\|^2_{L_2(P)}\right) \ ,$$

Therefore, on $\Omega_M(\rho_K, c_7)$, from (13),

$$d_K(f, f^*) = \mathrm{MOM}_K[\xi(f^* - f)] \leq \max_{k \in \mathcal{K}} \overline{P}_{B_k} \xi(f - f^*) + Q^{\overline{P}}_{3/4,K} |\xi(f - f^*)|$$
$$\leq P\xi(f - f^*) + 3c_3 \beta^2 \max\left(r_M^2(\rho_K), \|f - f^*\|^2_{L_2(P)} \right) \ .$$

By convexity of F and the definition of f^*, $P\xi(f - f^*) \leq 0$, which concludes the proof of the first point. The second point follows from (22).

■

6.6 Proof of Theorem 2

Proof of point 1. of Theorem 2: It follows from Theorem 1 that for every integer $J \in [K^*, N]$, the event $\{f^* \in R_J^{(1)} = B(\rho_J, \hat{f}_J^{(1)})\}$ has probability larger than $1 - 2\exp(-C_7 J)$, where $C_7 = \frac{1}{2}(c_1^2 \wedge c_7^2)$. For every $K \in [K^*, N]$, the event $\{f^* \in \cap_{J=K}^N R_J^{(1)}\}$ has probability

$$\mathbb{P}\left[f^* \in \cap_{J=K}^N R_J^{(1)} \right] \geq 1 - \sum_{J=K}^{N} 2\exp(-C_7 J) \geq 1 - c_8 \exp(-C_7 K), \qquad c_8 = 2e^{C_7}/(1 - e^{-C_7}) \ .$$

On this event, $\hat{K}^{(1)} \leq K$ and so $\hat{f}_{LE}^{(1)} \in R_K^{(1)}$ implying that $\left\| \hat{f}_{LE}^{(1)} - f^* \right\| \leq 2\rho_K$. ■

The proof of points 2 and 3 of Theorem 2 rely on the next result which follows from a straigforward modification (just replace ρ_K by $2\rho_K$) of the proofs of Lemmas 3 and 4.

Lemma 7. *Grant the conditions of Theorem 1. There exists an absolute constant c_9 and an event $\tilde{\Omega}(K)$ such that $\mathbb{P}\left(\tilde{\Omega}(K)\right) \geq 1 - 2\exp\left(-c_9 K\right)$ where, for any $(f, g) \in F \cap B(f^*, 2\rho_K)$,*

$$\|f - g\|_{L_{2,N}} \leq \frac{\tau}{\sqrt{c_1}} (\|f - g\|_{L_P^2} \vee r_Q(2\rho_K)) \quad \textit{and, if } \|f - g\|_{L_P^2} \geq r_Q(2\rho_K), \quad \|f - g\|_{L_{2,N}} \geq \beta \|f - g\|_{L_P^2} \ . \tag{27}$$

On the same event, there exists η, independent from K and N such that

$$\forall f, g \in F \cap B(f^*, 2\rho_K) \cap B_2(f^*, \zeta r(2\rho_K)), \qquad Q^{\bar{P}}_{3/4,K} |(Y - f)(g - f)| \leq \eta r^2(2\rho_K) \ .$$

Proof of point 2. of Theorem 2: It follows from Theorem 1 that for every integer $J \in [cr^2(\rho^*)N, N]$, there exists an event $\bar{\Omega}(J)$ with probability larger than $1 - 2\exp(-C_7 J)$, such that on $\bar{\Omega}(J)$, we have both

$$\left\| \hat{f}_J^{(2)} - f^* \right\| \leq \rho_J \text{ and } \left\| \hat{f}_J^{(2)} - f^* \right\|_{L_P^2} \leq \frac{\tau}{\beta \sqrt{c_1}} r(\rho_J). \tag{28}$$

Now, consider the event $\Omega_2(J)$ (resp. $\tilde{\Omega}(J)$) as defined in Lemma 3 (resp. Lemma 7) and denote $\Omega(J) := \bar{\Omega}(J) \cap \tilde{\Omega}(J) \cap \Omega_2(J)$. On $\Omega(J)$, we have from (28) and Lemma 3 that

$$\|\hat{f}_J^{(2)} - f^*\|_{L_{2,N}} \leq \frac{\tau}{\sqrt{c_1}} \left(\left\| \hat{f}_J^{(2)} - f^* \right\|_{L_2(P)} \vee r_Q(\rho_J) \right) \leq \frac{\tau}{\sqrt{c_1}} r(\rho_J) \ . \tag{29}$$

This implies that $f^* \in R_J^{(2)}$.
Let now $K \in [K^*, N]$. On $\cap_{J=K}^N \Omega(J)$, which has probability $1 - c_0 e^{-J/c_0}$, we have $f^* \in \cap_{J=K}^N R_J^{(2)}$ therefore $\hat{K}^{(2)} \leq K$ and so $\hat{f}_{LE}^{(2)} \in R_K^{(2)}$. This means that

$$\left\| \hat{f}_{LE}^{(2)} - \hat{f}_J^{(2)} \right\| \leq \rho_J \text{ and } \left\| \hat{f}_{LE}^{(2)} - \hat{f}_J^{(2)} \right\|_{L_{2,N}} \leq \frac{\tau}{\sqrt{c_1}} r(\rho_J).$$

In particular, both $\hat{f}_{LE}^{(2)}$ and $\hat{f}_J^{(2)}$ belong to $B(2\rho_J, f^*)$ and one has either $\left\|\hat{f}_{LE}^{(2)} - \hat{f}_J^{(2)}\right\|_{L_P^2} \le r(2\rho_J)$ or, by Lemma 7,

$$\left\|\hat{f}_{LE}^{(2)} - \hat{f}_J^{(2)}\right\|_{L_P^2} \le \frac{1}{\beta}\|\hat{f}_{LE}^{(2)} - \hat{f}_J^{(2)}\|_{L_{2,N}} \le \frac{\tau}{\beta\sqrt{c_1}} r(2\rho_J) \enspace .$$

Together with (28), this gives

$$\left\|\hat{f}_{LE}^{(2)} - f^*\right\|_{L_P^2} \le \left\|\hat{f}_{LE}^{(2)} - \hat{f}_J^{(2)}\right\|_{L_P^2} + \left\|\hat{f}_J^{(2)} - f^*\right\|_{L_P^2} \le \left(1 + \frac{1}{\beta}\right) \frac{\tau}{\sqrt{c_1}} r(2\rho_J) \enspace .$$

■

Proof of point 3. of Theorem 2: It follows from Theorem 1 that for every integer $J \in [K^*, N]$, there exists an event $\bar{\Omega}(J)$ with probability larger than $1 - 2\exp(-C_7 J)$, such that on $\bar{\Omega}(J)$, we have

$$\left\|\hat{f}_J^{(3)} - f^*\right\| \le \rho_J, \left\|\hat{f}_J^{(3)} - f^*\right\|_{L_P^2} \le \frac{\tau}{\beta\sqrt{c_1}} r(\rho_J) \text{ and } R(\hat{f}_J^{(3)}) - R(f^*) \le C_4 r^2(\rho_J) \enspace . \tag{30}$$

Now, consider the event $\Omega_2(J)$ (resp. $\tilde{\Omega}(J)$) as defined in Lemma 3 (resp. Lemma 7) and denote $\Omega(J) := \bar{\Omega}(J) \cap \tilde{\Omega}(J) \cap \Omega_2(J)$. On $\Omega(J)$, by Lemma 4, $d_K(\hat{f}_K^{(3)}, f^*) \ge -\frac{3c_3}{\sqrt{c_1}}\tau\beta r^2(\rho_K)$, thus, $f^* \in R_J^{(3)}$. Then let $K \in [K^*, N]$. On $\cap_{J=K}^N \Omega(J)$, we have $f^* \in \cap_{J=K}^N R_J^{(3)}$, therefore $\hat{K}^{(3)} \le K$ and so $\hat{f}_{LE}^{(3)} \in R_K^{(3)}$. This implies by points 1. and 2. that

$$\left\|\hat{f}_{LE}^{(3)} - f^*\right\| \le 2\rho_K, \quad \left\|\hat{f}_{LE}^{(3)} - f^*\right\|_{L_P^2} \le \left(1 + \frac{1}{\beta}\right) \frac{\tau}{\sqrt{c_1}} r(2\rho_K) \enspace , \tag{31}$$

and, by definition that $d_K(\hat{f}_{LE}^{(3)}, \hat{f}_K^{(3)}) \ge -\nu r^2(\rho_K)$. By (30),

$$\begin{aligned} R(\hat{f}_{LE}^{(3)}) - R(f^*) &= \mathbb{E}[(Y - \hat{f}_K^{(3)}(X) + \hat{f}_K^{(3)}(X) - \hat{f}_{LE}^{(3)}(X))^2] - R(f^*) \\ &= \left\|\hat{f}_{LE}^{(3)} - \hat{f}_K^{(3)}\right\|_{L_P^2}^2 + 2P(Y - \hat{f}_K^{(3)})(\hat{f}_K^{(3)} - \hat{f}_{LE}^{(3)}) + R(\hat{f}_K^{(3)}) - R(f^*) \\ &\le 2P(Y - \hat{f}_K^{(3)})(\hat{f}_K^{(3)} - \hat{f}_{LE}^{(3)}) + (\frac{\tau}{\sqrt{c_1}\beta} + C_4) r^2(\rho_K) \enspace . \end{aligned}$$

then, the third point in Theorem 2 hold on $\cap_{J=K}^N \Omega(J)$ if there exists at least one block B_k such that

$$\begin{aligned} &P(Y - \hat{f}_K^{(3)})(\hat{f}_K^{(3)} - \hat{f}_{LE}^{(3)}) = (P - \bar{P}_{B_k})(Y - \hat{f}_K^{(3)})(\hat{f}_K^{(3)} - \hat{f}_{LE}^{(3)}) \\ &+ (\bar{P}_{B_k} - P_{B_k})(Y - \hat{f}_K^{(3)})(\hat{f}_K^{(3)} - \hat{f}_{LE}^{(3)}) - P_{B_k}(Y - \hat{f}_K^{(3)})(\hat{f}_{LE}^{(3)} - \hat{f}_K^{(3)}) \lesssim r^2(2\rho_K). \end{aligned} \tag{32}$$

Since $d_K(\hat{f}_{LE}^{(3)}, \hat{f}_K^{(3)}) \ge -\nu r^2(\rho_K)$, the last term in (32) is properly bounded on $K/2$ blocks B_k at least. From (31) and assumption (16), the first term of (32) is properly bounded for every $k \in \mathcal{K}$, that is, on at least $3K/4$ blocks B_k under the conditions of Theorem 1. Finally, the second term in (32) is properly bounded on $3K/4$ blocks by the last item of Lemma 7, which applies thanks to (31). As a consequence, there exists at least one block k where (32) holds.

■

References

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

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

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

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

[5] Yannick Baraud and Lucien Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, 143(1-2):239–284, 2009.

[6] Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection : ρ-estimation. *To appear in Invent. Math.*

[7] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1092–1119, 2014.

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

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

[10] Lucien Birgé. Robust tests for model selection. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 47–64. Inst. Math. Statist., Beachwood, OH, 2013.

[11] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015.

[12] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013. ISBN 978-0-19-953525-5.

[13] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *Ann. Statist.*, 43(6):2507–2536, 2015.

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

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

[16] Michaël Chichignoud and Johannes Lederer. A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression. *Bernoulli*, 20(3):1560–1599, 2014.

[17] Víctor H. de la Peña and Evarist Giné. *Decoupling.* Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

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

[19] J. Fan, Q. Li, and Y. Wang. Estimation of high dimensional mean regression in absence of symmetry and light-tail assumptions. *To appear in Journal of the Royal Statistical Society, Serie B.*

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

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

[22] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

[23] Vladimir Koltchinskii and Shahar Mendelson. Bounding the Smallest Singular Value of a Random Matrix Without Concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008, 2015.

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

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

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

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

[28] Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. Technical report, CNRS, Ecole Polytechnique and Technion, 2014. To appear in Journal of the European Mathematical Society.

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

[30] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2001.

[31] O. V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4):645–659, 1991.

[32] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Preprint available on ArXive:1608.00757.*

[33] Shahar Mendelson. Learning without concentration for general loss function. Technical report, Technion, I.I.T., 2013. arXiv:1410.3192.

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

[35] Shahar Mendelson. A remark on the diameter of random sections of convex bodies. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 395–404. Springer, Cham, 2014.

[36] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

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

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

[39] Mark Rudelson and Roman Vershynin. Small ball probabilities for linear images of high dimensional distributions. Technical report, University of Michigan, 2014. International Mathematics Research Notices, to appear. [arXiv:1402.4492].

[40] Mathieu Sart. Estimation of the transition density of a Markov chain. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1028–1068, 2014.

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

[42] W. N. Wapnik and A. J. Tscherwonenkis. *Theorie der Zeichenerkennung*, volume 28 of *Elektronisches Rechnen und Regeln, Sonderband [Electronic Computing and Control, Special Issue].* Akademie-Verlag, Berlin, 1979. Translated from the Russian by Klaus-Günter Stöckel and Barbara Schneider, Translation edited by Siegfried Unger and Klaus Fritzsch.