Estimation properties of regularization methods under the small ball property

Guillaume Lecué[1,3]   Shahar Mendelson[2,4,5]

August 22, 2014

Abstract

Regularization procedures are popular methods in Statistics and Learning Theory to enhance low-dimensional structures or smoothness properties of estimators. In this work, we study the estimation property of regularization procedures of the form

$$\hat{f} \in \underset{f \in F}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \big(Y_i - f(X_i)\big)^2 + \lambda \left\| f \right\| \right)$$

where the regularization function $\|\cdot\|$ satisfies rather weak property allowing for various type of functions like ℓ_p- (quasi) norms and S_p-(quasi) norms, for all $0 < p \leq \infty$, atomic norms, max-norm, RKHS norms for which we provide a detailled study. It appears that rather weak moment properties are enough to obtain these estimation results when X satisfies a small ball property.

1 Introduction

Let $\mathcal{X}$ be a space endowed with a probability measure μ. Let $X_1, \ldots, X_N$ be N iid input points in $\mathcal{X}$ distributed according to μ. To each point X_i a real-valued output Y_i is associated such that $(X_i, Y_i)_{i=1}^N$ are N iid random variables taking values in $\mathcal{X} \times \mathbb{R}$. Given a new input X in $\mathcal{X}$, we want to predict an associated output. To that end, we use the set of data $\mathcal{D} =$

[1] CNRS, CMAP, Ecole Polytechnique, 91120 Palaiseau, France.

[2] Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.

[3] Email: guillaume.lecue@cmap.polytechnique.fr

[4] Email: shahar@tx.technion.ac.il

[5] Supported by the Mathematical Sciences Institute – The Australian National University and by ISF grant 900/10.

$\{(X_i, Y_i) : i = 1, \cdots, N\}$ to construct prediction rules $\hat{f}(\cdot) = \hat{f}(\mathcal{D}, \cdot)$ such that $\hat{f}(X)$ will be a good guess of the output Y when (X, Y) is distributed like the (X_i, Y_i)'s.

In the learning theory framework, we do not assume any statistical model behind the data generating process but we are given a class of real-valued functions F defined on $\mathcal{X}$. We assume that F **is convex and closed in** $L_2(\mu)$ and denote by f^* a best approximation of Y in $F \subset L_2(\mu) = L_2$:

$$f^* \in \operatorname*{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2. \tag{1.1}$$

Our aim is to construct a prediction rule $\hat{f}$ which is close to f^* in L_2: with large probability,

$$\left\| \hat{f} - f^* \right\|_{L_2}^2 = \mathbb{E}\big(f^*(X) - \hat{f}(X)\big)^2 \leq \alpha_N(F)^2 \tag{1.2}$$

where $\alpha_N(F)^2$ is called the rate of convergence and we want to construct prediction rules $\hat{f}$ such that $\alpha_N(F)^2$ is as small as possible. Note that even though f^* exists, it may not be unique.

Since, we do not assume any underlying statistical model, the price to pay to insure that f^* is a reasonable prediction of Y is that one has to consider models F of large size and therefore the rate $\alpha_N(F)^2$ will be large as well whatever the procedure $\hat{f}$ is. To avoid this issue, we usually try to find some a priori property satisfied by f^*. Such a property is usually caracterized by a function $\|\cdot\|$ such that $\|f^*\|$ should be small– even though this is no assumption. Note that the function $\|\cdot\|$ does not have to be a norm; properties on $\|\cdot\|$ required by our analysis are given below:

Assumption 1.1 *The function $\|\cdot\|$ is non-negative, convex and satisfies the following properties:*

(N1) There exists $\eta_1 \geq 1$ such that for any $f, g \in F$,

$$\|f - g\|, \|f + g\| \leq \eta_1\big(\|f\| + \|g\|\big).$$

(N2) For any $f, g \in F$, $\lambda \mapsto \|\lambda(f - g)\|$ is a continuous function from $[0, 1]$ onto $[0, \|f - g\|]$. Moreover, for any $\lambda \in [0, 1]$, $\|\lambda(f - g)\| \leq \lambda \|f - g\|$ and $\|\lambda f\| \leq \lambda \|f\|$.

Classical examples of function $\|\cdot\|$ are related to the smoothness of f^* or to some sparsity structure via some norms or quasi-norms (note that any quasi-norm satisfies Assumption 1.1) or mixtures of norms. We will explore several examples of functions $\|\cdot\|$ in the following sections.

A classical approach is to consider **regularized empirical risk minimization procedures** (RERM). A regularized empirical risk minimization procedure is defined by

$$\hat{f} \in \underset{f \in F}{\operatorname{argmin}} \left(P_N \ell_f + \lambda \|f\| \right) \tag{1.3}$$

where P_N is the empirical probability measure, ℓ_f is the loss function associated with f and λ is the regularization parameter. Here we consider only the square loss $\ell_f(x,y) = (y - f(x))^2$ for all $x \in \mathcal{X}$ and $y \in \mathbb{R}$, in particular, $P_N \ell_f = N^{-1} \sum_{i=1}^N (Y_i - f(X_i))^2$.

There are a large number of both theoretical and practical study on the performance of estimators like (1.3) for some particular examples of regularization function. One typical example of such a procedure is the Lasso (cf. [41]). It is obtained when F is a class of linear functional of the form $\langle \cdot, t \rangle$ for $t \in \mathbb{R}^d$ and the regularization function is the ℓ_1^d-norm. Estimation, de-noising, prediction and support recovery results have been obtained for the Lasso during the last decades (cf. [41], [2], we refer the reader to the books [4] and [19] for more references). Few statistical results for the Lasso have been obtained in the random design scenario (cf. [1], [26] and chapter 8.2 in [19]). The vast majority of the results have been obtained in the linear model with sub-gaussian noise and for a fixed design satisfying some weak form of RIP. One typical example of such a property is the *Restricted Eigenvalue Condition* (REC) from [2]. To define it, let us introduce the following notation: for $t \in \mathbb{R}^d$ and a set $S_0 \subset \{1, \dots, d\}$ of cardinality $|S_0| \leq s$, let S_1 be the subset of indexes of the m largest coordinates of $(|x_i|)_{i=1}^n$ that are outside S_0. Let $t_{S_{01}}$ be the restriction of t to the set $S_{01} = S_0 \cup S_1$.

Definition 1.1 *Let Γ be an $N \times d$ matrix. For a constant $c_0 \geq 1$ and an integer $1 \leq s \leq m \leq d$ for which $m + s \leq d$, let the* ***restricted eigenvalue constant*** *be*

$$\kappa(s, m, c_0) = \min \left(\frac{\|\Gamma t\|_2}{\|t_{S_{01}}\|_2} : S_0 \subset \{1, \dots, n\}, |S_0| \leq s, \left\|t_{S_0^c}\right\|_1 \leq c_0 \|t_{S_0}\|_1 \right).$$

The matrix Γ satisfies the ***Restricted Eigenvalue Condition (REC) of order*** *s* ***with a constant*** *c if $\kappa(s, s, 3) \geq c$.*

Several statistical properties of the Lasso have been obtained in [2] under (REC). It appears that this condition is of the same flavor as the one we consider below in $(\mathbf{Q}(\rho))$ in the learning theory setup.

There are also results obtained for general regularization methods as we aim to obtain here. In [34], the authors identify theoretical principles that underlies the analysis of several regularization methods for diverse loss functions and regularization functions. They obtain estimation results under two properties: 1) the regularizing function $\|\cdot\|$ is a norm satisfying the so-called *decomposability property*; 2) a control on the interaction between the regularization function and the loss function called the *restricted strong convexity*. Under those two properties, convergence rates for $\hat{f}$ are obtained. In the linear model with sub-gaussian noise and a fixed design satisfying the (REC) condition, it is proved in [25] that these two conditions are satisfied with large probability for the square loss and several examples of regularization methods when the target vector belongs to some classes of signals.

In [50], the authors obtain results for fairly general concave regularization methods in the linear model with sub-gaussian noise and fixed design. They obtained estimation, de-noising and support selection properties for the regularization method $\hat{f}$ under a condition on the design matrix called the *restricted invertibility factor* and some structures on the regularization functions. Similar results are also obtained for approximate local minimizers in [50].

At some point, all these works had to face the problem of calibration of the regularization parameter λ. The strategy we use to construct a regularization parameter λ is as follows. For any $\rho \geq 0$, we define the sub-model

$$F_\rho = \{f \in F : \|f\| \leq \rho\}. \tag{1.4}$$

We also define the excess loss of $f \in F$ with respect to f^* by:

$$\mathcal{L}_f = \ell_f - \ell_{f^*}. \tag{1.5}$$

We consider the quadratic/linear decomposition of the excess loss:

$$\begin{aligned}\mathcal{L}_f(x,y) &= \big(\ell_f - \ell_{f^*}\big)(x,y) = \big(y - f(x)\big)^2 - \big(y - f^*(x)\big)^2 \\ &= \big(f(x) - f^*(x)\big)^2 + 2\big(f^*(x) - f(x)\big)\big(y - f^*(x)\big).\end{aligned} \tag{1.6}$$

Following this decomposition, an important role will be played by two empirical processes: if we denote by P the actual measure – and recall that P_N is the empirical measure over the data then $P_N\mathcal{L}_f = P_N(f-f^*)^2 - 2P_N(f^*-f)(f^*-Y)$. The first empirical process we consider is the **quadratic empirical process** $\big(P_N(f-f^*)^2 : f \in F\big)$ where

$$P_N(f-f^*)^2 = \frac{1}{N}\sum_{i=1}^N \big(f(X_i) - f^*(X_i)\big)^2. \tag{1.7}$$

Under a weak assumption (cf. the small ball property in Definition 2.1 below), this process has the following property.

Definition 1.2 *Let $\rho \geq 0$. We say that the quadratic process satisfies an* ***empirical small ball property of level*** *$s_Q(\rho) \geq 0$ when for all $f \in F_\rho$ satisfying $\|f - f^*\|_{L_2} \geq s_Q(\rho)$, we have*

$$P_N(f - f^*)^2 \geq \kappa_0 \|f - f^*\|_{L_2}^2 \qquad (\mathbf{Q}(\rho))$$

where κ_0 is an absolute constant.

One way to see this condition is as a generalization of the (REC) to the learning theory framework. It is interesting to note that the restricted strong convexity from [34] turns also to be a type of (REC) when dealing with the square loss. This condition (on the quadratic empirical process) seems to be a key property to obtain statistical property of regularized procedures. The main result from Section 2 below shows that this condition is almost always true (with large probability) as long as X satisfies the so-called small ball property which is a pretty weak requirement (for instance a vector of iid Cauchy variables satisfies this condition even though Cauchy variables don't even have a mean). An output of our results is to show that the key property ($\mathbf{Q}(\rho)$) that appeared in many works, in a form or another, is satisfied by fairly general design vectors X. What costs more is the control of the linear process due to the noise $Y - f^*(X)$.

The other empirical process that plays a central role in our analysis is the **linear empirical process** $\big(P_N(f^* - f)(f^* - Y) : f \in F\big)$ where

$$P_N(f^* - f)(f^* - Y) = \frac{1}{N}\sum_{i=1}^{N} \big(f^*(X_i) - f(X_i)\big)\big(f^*(X_i) - Y_i\big). \qquad (1.8)$$

When a statistical model is assumed to hold then $Y - f^*(X)$ is equal to the noise. In the more general learning theory setup, the quantity $Y - f^*(X)$ can still be considered as a noise. We will therefore call it like that. With this terminology, the empirical process (1.8) is an empirical measure of the correlation between the noise and the class (centered around f^*). In particular when there is no noise (i.e. $Y = f^*(X)$), this process equals zero and does not enter the analysis. As in Definition 1.2 concerning the quadratic empirical process, we will be interested in some particular property of the linear empirical process (1.8).

Before introducing the property of this process that plays a central role in our analysis, we introduce some notations. We recall that L_2 is the Hilbert space endowed with the norm $\|f\|_{L_2} = \big(\mathbb{E}f(X)^2\big)^{1/2}$. We denote by S_{L_2} (resp. $\mathcal{D}$) the unit sphere (resp. ball) of L_2; in particular, if $f^* \in L_2$ and $s > 0$, $f^* + s\mathcal{D} = \{f \in L_2 : \|f - f^*\|_{L_2} \leq s\}$.

Definition 1.3 *For all $s \geq 0$ and $\rho \geq 0$, define*

$$\phi_N(\rho, s) = \sup_{f \in F_\rho \cap (f^* + s\mathcal{D})} \frac{1}{N} \sum_{i=1}^{N} (f^* - f)(X_i)(f^*(X_i) - Y_i).$$

*Let $\rho \geq 0$. We say that the linear process satisfies a **noise/class interaction of level** $s_L(\rho) \geq 0$ when, for some absolute constant κ_1,*

$$\phi_N(\rho, s_L(\rho)) \leq \kappa_1 s_L(\rho)^2, \qquad (\mathbf{L}(\rho))$$

when $s_L(\rho) > 0$ and $\phi_N(\rho, s) \leq 0$ for all $s \geq 0$ when $s_L(\rho) = 0$.

The two conditions $(\mathbf{L}(\rho))$ and $(\mathbf{Q}(\rho))$ characterize the order of magnitude of the linear and quadratic process under two types of conditions: some deviation conditions on the "noise" $Y - f^*(X)$ and the class F for property $(\mathbf{L}(\rho))$ and a small ball property for $(\mathbf{Q}(\rho))$. Those properties are studied in Section 2 and 3. The two functions $s_L(\cdot)$ and $s_Q(\cdot)$ characterize the rate of convergence of the Empirical risk minimization procedure over all sub-models F_ρ (cf. [22]).

Our first result is to show that under properties $(\mathbf{L}(\rho^*))$ and $(\mathbf{Q}(\rho^*))$ for $\rho^* = \eta_1 C_0 \|f^*\|$, the RERM $\hat{f}$ satisfies some Model Selection properties in the sense that it belongs to the "correct" model F_{ρ^*} and some estimation properties of f^* in L_2. It should be noted that those properties concern only the "true" model F_{ρ^*} and not the entire space F which should be much bigger than F_{ρ^*} when f^* is such that $\|f^*\|$ is small (which is the motivation behind the study of (1.3)). Before that we introduce conditions on the regularization parameter λ under which (1.3) will be proved to have the expected Model Selection and Estimation properties.

Assumption 1.2 $(\rho, \mathbf{R}^*, \mathbf{c_0}, \mathbf{C_0})$ *Let η_1, κ_0 and κ_1 be the constants appearing in Assumption 1.1 and Definitions 1.2 and 1.3. Let $s(\cdot)$ be such that*

$$s(\rho) \geq \max\big(s_Q(\rho), s_L(\rho)\big). \qquad (1.9)$$

There exists $\rho > 0$, $R^* > 0$, $c_0 > 0$, and $C_0 > 2\eta_1 + 1$ such that:

i) $2\lambda\rho \leq \kappa_0 c_0^2 s^2\big(\eta_1^2 C_0 \rho\big)$

ii)

$$\Big(\frac{C_0-2\eta_1-1}{2\eta_1}\Big)\lambda \geq \frac{\phi_N\big(\eta_1 C_0\rho, c_0 s(\eta_1^2 C_0\rho)\big)}{\rho}.$$

iii) If $\|f^*\|=0$ *then*

$$\frac{\lambda}{2\eta_1} \geq \sup_{0<r\leq \eta_1 C_0 R^*} \frac{\phi_N\big(\eta_1 r, c_0 s(\eta_1^2 C_0 R^*)\big)}{r}$$

For a choice of regularization function satisfying Assumption 1.2, we obtain the following model selection and estimation result for the RERM $\hat{f}$.

Theorem A: *Let* $\eta_1\geq 1$*,* $C_0>2\eta_1+1$ *and* $R^*>0$*. Let* $\|\cdot\|$ *satisfying Assumption 1.1. Assume properties (*$\mathbf{L}(\eta_1^2 C_0\|f^*\|)$*) and (*$\mathbf{Q}(\eta_1^2 C_0\|f^*\|)$*) hold and properties (*$\mathbf{L}(\eta_1^2 C_0 R^*)$*) and (*$\mathbf{Q}(\eta_1^2 C_0 R^*)$*) hold as well for some* $\kappa_1\geq 0$ *and* $\kappa_0>0$*. Let* λ *satisfy Assumption 1.2 with parameters (*$\|\mathbf{f}^*\|, \mathbf{R}^*, \mathbf{c_0}, \mathbf{C_0}$*) for some* $c_0\geq \max(4\kappa_1/\kappa_0, 1)$*. Then, the RERM* $\hat{f}$ *is such that:*

A) $\left\|\hat{f}-f^*\right\|_{L_2}\leq c_0 s\big(\eta_1^2 C_0\|f^*\|\big)$,

B) $\left\|\hat{f}\right\|\leq \eta_1 C_0\|f^*\|$.

Estimation results follow from Theorem A when one is able to choose λ and $s(\cdot)$ such that Assumption 1.2 is satisfied. One way to choose those parameters is such that

$$\lambda \sim \sup_{\rho>0}\frac{\mathbb{E}\phi_N(\rho,\infty)}{\rho} \tag{1.10}$$

so that points *ii)* and *iii)* in Assumption 1.2 hold with large probability (under appropriate stochastic assumptions introduced in the two following sections) and then take

$$s^2(\rho)=\max\Big(s_L^2(\rho), s_Q^2(\rho), \frac{2\lambda\rho}{\kappa_0 c_0^2\eta_1^2 C_0}\Big)$$

so that point *i)* of Assumption 1.2 is satisfied. This choice will be made in the main applications of Theorem A: Theorems 4.2, 4.4 and 4.3 below.

Theorem A is a deterministic result in the same spirit as the results in [45, 44] or [25, 34]. All the stochastic/complexity part of our analysis is contained in the two properties ($\mathbf{L}(\rho^*)$) and ($\mathbf{Q}(\rho^*)$) and in upper bounds on the linear empirical processes ϕ_N. Obtaining results on these two properties and ϕ_N is the aim of Section 2 and 3. In particular, Theorem A has nothing to do with the fact that the data $(X_i, Y_i)_{i=1}^N$ introduced at the beginning are i.i.d.; Theorem A applies for instance when the data are dependent. Note also that Theorem A can be generalized to other loss functions than the quadratic one thanks to Taylor expansions of the loss. Finally, note that another advantage of a deterministic result like Theorem A is that the regularization parameter λ can be expressed in function of the data. Three directions that will not be studied here.

Note that Theorem A also apply when $\|f^*\| = 0$. In this case, $\hat{f}$ satisfies $\left\|\hat{f} - f\right\|_{L_2} \leq c_0 s(0)$ and $\left\|\hat{f}\right\| = 0$. In particular, when the regularizing function $\|\cdot\|$ is a norm, $\hat{f} = f = 0$; that is exact reconstruction even though there is some noise. This can be surprising at a first glance but this is one consequence of the regularization function that forces the estimator $\hat{f}$ towards zero which is the correct target when $f = 0$.

Note that we don't use the definition of f^* from (1.1) as an oracle. In fact, Theorem A applies to any function $f^* \in F$ for which the two properties ($\mathbf{L}(\rho^*)$) and ($\mathbf{Q}(\rho^*)$) hold. This may be useful when one wants to obtain oracle inequalities instead of estimation results.

The paper is organized as follows. In order to apply Theorem A, we have to construct functions $s_L(\cdot)$ and $s_Q(\cdot)$ for which ($\mathbf{L}(\rho)$) and ($\mathbf{Q}(\rho)$) hold with large probability for a given $\rho > 0$. We also need to upper bound $\phi_N(\rho, s)$ (cf. definition 1.3) in order to choose λ such that Assumption 1.2 holds. This is the purpose of the two following sections. This is where the stochastic (small ball and moment assumptions) and complexity arguments (Gaussian mean width and expected suprema) enter the analysis. In Section 4, we prove in Theorem 4.2 that the only computation of the Gaussian mean width of the unit ball associated with $\|\cdot\|$ allows for a calibration of the regularization parameter and for the identification of the rate of estimation. We study several examples of application of this result in Section 4 to 7 just by computing some Gaussian mean widths. This result does not require any statistical model and is true if the noise is in L_q, for some $q > 2$. In the case where a statistical model is true, a similar result is stated in Section 4.2 under weak moment assumptions on the coordinates of X as long as it satisfies a small ball property. A property that will play an important role in our analysis which introduced in the next section

Notation. $(e_1, \ldots, e_d)$ denotes the canonical basis of $\mathbb{R}^d$; for every $p > 0$, ℓ_p^d is the space $\mathbb{R}^d$ endowed with the (quasi)-norm $\|t\|_{\ell_p^d} = \left(\sum_{j=1}^d |t_j|^p\right)^{1/p}$.

The unit balls and spheres of the ℓ_p^d spaces are denoted by $B_p^d = \{t \in \mathbb{R}^d : \|t\|_{\ell_p^d} \leq 1\}$ and $S_p^{d-1} = \{t \in \mathbb{R}^d : \|t\|_{\ell_p^d} = 1\}$.

The set L_2 denotes the Hilbert space $L_2(\mathcal{X}, \mu)$, its norm is denoted by $\|f\|_{L_2} = \big(\mathbb{E}f^2(X)\big)^{1/2}$, its unit ball is denoted by $\mathcal{D} = \{f \in L_2 : \big(\mathbb{E}f^2(X)\big)^{1/2} \leq 1\}$ and its unit sphere is denoted by S_{L_2}. If $H \subset L_2$ and $f^* \in H$ then $H - H = \{h - g : h, g \in H\}$ and $H - f^* = \{h - f^* : h \in H\}$. When there will be no ambiguity, we will also use the notation $\|f(X)\|_{L_2}$ for $\big(\mathbb{E}f^2(X)\big)^{1/2}$.

2 Property ($\mathbf{Q}(\rho)$) under the small ball assumption

In this section, we study condition ($\mathbf{Q}(\rho)$) under the following small ball assumption introduced in [29]:

Definition 2.1 *Let H be a class of functions and define*

$$Q_H(u) = \inf_{h \in H} P\big(|h| \geq u \, \|h\|_{L_2}\big).$$

*We say that H satisfies the **small ball assumption** when there exists positive constants u_0 and β_0 such that $Q_H(u_0) \geq \beta_0$.*

Several examples of classes satisfying the small ball property are given for linear functionals: $h(X) = \big\langle X, \beta \big\rangle$ in [29, 31, 23]. The following result is a key one for the understanding of the role played by the small ball property in condition ($\mathbf{Q}(\rho)$).

Lemma 2.2 *Let H be a class of functions. Assume that there exists some $u_0 > 0$ such that $Q_H(u_0) > 0$ and*

$$\mathbb{E} \sup_{h \in H} \Big| \frac{1}{N} \sum_{i=1}^N \varepsilon_i I\big(|h(X_i)| \geq u_0 \, \|h\|_{L_2}\big) \Big| \leq \frac{Q_H(u_0)}{8} \tag{2.1}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are iid Rademacher variables. Then with probability larger than $1 - \exp(-N Q_H^2(u_0)/8)$, for every $h \in H$, $P_N h^2 \geq \kappa_0 \, \|h\|_{L_2}^2$ for $\kappa_0 = u_0^2 Q_H(u_0)/2$.

Proof. Let

$$G(X_1, \ldots, X_N) = \sup_{h \in H} \big|(P - P_N) I(|h(\cdot)| \geq u_0 \, \|h\|_{L_2})\big|.$$

It follows from the bounded difference inequality (cf. Theorem 6.2 in [3]) that with probability larger than $1-\exp(-x)$,

$$G(X_1,\dots,X_N) \le \mathbb{E}G(X_1,\dots,X_N) + \sqrt{\frac{x}{2N}}.$$

It follows from a symmetrization argument (cf. [24]) and (2.1) that $\mathbb{E}G(X_1,\dots,X_N) \le Q_H(u_0)/4$. So, if one takes $x = NQ_H(u_0)^2/8$ then, with probability larger than $1-\exp(-NQ_H(u_0)^2/8)$, $G(X_1,\dots,X_N) \le Q_H(u_0)/2$. Moreover, by definition of $Q_H(\cdot)$, any $h \in H$ is such that $P(|h(X)| \ge u_0 \left\|h\right\|_{L_2}) \ge Q_H(u_0)$ so, with the same probability estimate,

$$P_N I\big(|h(\cdot)| \ge u_0 \left\|h\right\|_{L_2}\big) \ge Q_H(u_0)/2$$

and therefore, for any $h \in H$,

$$\frac{1}{N}\sum_{i=1}^N h(X_i)^2 \ge u_0^2 \left\|h\right\|_{L_2}^2 P_N I\big(|h(\cdot)| \ge u_0 \left\|h\right\|_{L_2}\big) \ge \frac{u_0^2 \left\|h\right\|_{L_2}^2 Q_H(u_0)}{2}.$$

■

Theorem 2.3 *Let $\left\|\cdot\right\|$ be some function defined on L_2 satisfying Assumption 1.1 for some $\eta_1 \ge 1$. Let $\rho^* > 0$ and $f^* \in F_{\rho^*}$. Assume that the following hold:*

- *there exists u_0 and β_0 such that, $Q_{F_{\eta_1\rho^*}-f^*}(u_0) \ge \beta_0$,*
- *there exists $s_Q > 0$ such that*

$$\mathbb{E} \sup_{h \in (F_{\eta_1\rho^*}-f^*)\cap s_Q S_{L_2}} \Big|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\Big| \le \frac{\beta_0 u_0}{8} s_Q. \tag{2.2}$$

*Then, with probability at least $1-\exp(-\beta_0^2 N/8)$, for any $f \in F_{\rho^*}$, if $\left\|f-f^*\right\|_{L_2} \ge s_Q$ then $P_N(f-f^*)^2 \ge (u_0^2\beta_0/2)\left\|f-f^*\right\|_{L_2}^2$. In other words, (**Q**$(\rho^*)$) holds for $s_Q(\rho^*) = s_Q$ and $\kappa_0 = u_0\beta_0^2/2$ with probability larger than $1-\exp(-\beta_0^2 N/8)$.*

Proof. Denote $H = F_{\eta_1\rho^*} - f^*$. It follows from the contraction principle (cf. Chapter 4 in [24]) that

$$\begin{aligned}&\mathbb{E} \sup_{h \in H\cap s_Q S_{L_2}} \Big|\frac{1}{N}\sum_{i=1}^N \varepsilon_i I\big(|h(X_i)| \ge u_0 \left\|h\right\|_{L_2}\big)\Big| \\ &\le \frac{1}{u_0 s_Q}\mathbb{E} \sup_{h \in H\cap s_Q S_{L_2}} \Big|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\Big| \le \frac{\beta_0}{8}.\end{aligned}$$

It follows from Lemma 2.2 that with probability at least $1-\exp(-\beta_0^2 N/8)$, for any $h \in H \cap s_Q S_{L_2}$, $P_N h^2 \geq \kappa_0 \left\|h\right\|_{L_2}^2$ for $\kappa_0 = u_0^2\beta_0/2$. Therefore, with the same probability estimate, if one takes $f \in F_{\rho^*}$ such that $\left\|f-f^*\right\|_{L_2} \geq s_Q$ then by convexity of F and Assumption 1.1, $g = (s_Q/\left\|f-f^*\right\|_{L_2})(f-f^*) \in H \cap s_Q S_{L_2}$, so $P_N g^2 \geq \kappa_0 \left\|g\right\|_{L_2}^2$ and $P_N(f-f^*)^2 \geq \kappa_0 \left\|f-f^*\right\|_{L_2}^2$. ■

3 Properties ($\mathbf{L}(\rho)$) under moments and sub-gaussian assumption

In this section, we provide tools to check assumption ($\mathbf{L}(\rho)$) and to control the linear process ϕ_N. In what follows $\left\|\cdot\right\|$ is some function defined on L_2 satisfying Assumption 1.1 for some $\eta_1 \geq 1$.

In order to to prove that ($\mathbf{L}(\rho)$) holds with large probability, it is enough to obtain an upper bound on the quantity $\phi_N(\rho, s)$ for a given $s \geq 0$ that holds with large probability.

We first obtain such a bound under moment assumption on the design X and the noise $\zeta = Y - f^*(X)$ when X and ζ are independent. We use tools from chapter 2.9 in [46] from which we recall the notation:

$$\left\|\zeta\right\|_{2,1} = \int_0^\infty \sqrt{P(|\zeta| > x)}dx.$$

When $\zeta \in L_q$ for some $q > 2$ then $\left\|\zeta\right\|_{2,1}$ is finite. The set of all random variables ζ such that $\left\|\zeta\right\|_{2,1}$ is finite is denoted by $L_{2,1}$. We now work under the assumption that $\left\|\zeta\right\|_{2,1}$ is finite. Then, a direct application of Lemma 2.9.1 in [46] shows that the following holds.

Proposition 3.1 *Let $\zeta = Y - f^*(X)$ and assume that ζ is independent of X and $\mathbb{E}[\zeta] = 0$. Let $\rho^* \geq \left\|f^*\right\|$ and $s \geq 0$. Let $0 < \delta < 1$. With probability larger than $1-\delta$,*

$$\phi_N(\rho^*, s) \leq \frac{c_1 \left\|\zeta\right\|_{2,1}}{\delta\sqrt{N}} \max_{1 \leq k \leq N} \mathbb{E} \sup_{h \in F_{\rho^*-f^*} \cap s\mathcal{D}} \left|\frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i h(X_i)\right|$$

for $c_1 = 2\sqrt{2}$ and $\varepsilon_1, \ldots, \varepsilon_N$ iid Rademacher variables independent of the (Y_i, X_i)'s. In particular, if s_L is such that

$$\left\|\zeta\right\|_{2,1} \max_{1 \leq k \leq N} \mathbb{E} \sup_{h \in F_{\rho^*-f^*} \cap s_L\mathcal{D}} \left|\frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i h(X_i)\right| \leq \sqrt{N} s_L^2 \tag{3.1}$$

then ($\boldsymbol{L}(\rho^*)$) is satisfied with probability larger than $1-\delta$ for $\kappa_1 = c_1/\delta$ and $s_L(\rho^*) = s_L$. Also, with the same probability, point ii) of Assumption 1.2 is satisfied when λ is such that

$$\lambda\rho^* \geq \frac{2c_1\eta_1 \left\|\zeta\right\|_{2,1}}{(C_0 - 2\eta_1 - 1)\delta\sqrt{N}} \max_{1\leq k\leq N} \mathbb{E} \sup_{h\in F_{\rho^*-f^*}\cap s\mathcal{D}} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} \varepsilon_i h(X_i)\Big|. \qquad (3.2)$$

Proof. It follows from Markov inequality that with probability larger than $1-\delta$, $\phi_N(\rho, s) \leq \delta^{-1}\mathbb{E}|\phi_N(\rho, s)|$. Then, by Lemma 2.9.1 in [46],

$$\mathbb{E}|\phi_N(\rho, s)| \leq \frac{c_1 \left\|\zeta\right\|_{2,1}}{\sqrt{N}} \max_{1\leq k\leq N} \mathbb{E} \sup_{f\in F:\|f\|\leq\rho, \|f-f^*\|_{L_2}\leq s} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} \varepsilon_i (f-f^*)(X_i)\Big|.$$

■

It follows from Theorem 2.3 and the previous result that the only quantity that remains to be studied to calibrate the regularization parameter λ and get statistical properties for $\hat{f}$ thanks to Theorem A is

$$\max_{1\leq k\leq N} \mathbb{E} \sup_{h\in F_{\eta_1\rho^*-f^*}\cap s\mathcal{D}} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} \varepsilon_i h(X_i)\Big|. \qquad (3.3)$$

In what follows we obtain bounds on this quantity under moment assumptions on the coordinates of X.

The last result holds under the assumption that the noise ζ is independent of X. A typical case of application will be when a statistical model holds like $Y = f^*(X) + \zeta$ where ζ is a noise independent of X. This type of assumption may not be true in the general learning theory setup – where no statistical model is assumed. For this setup, we can still obtain some result if the design X is subgaussian by using a result on multiplier processes from [30] in place of lemma 2.9.1 from [46] used in the proof of Proposition 3.1. Before stating this result, we recall some definition.

Definition 3.2 *Let F be a class of functions in L_2 and $L > 0$. We say that F is a **L-subgaussian class with respect to** X when for all $f \in F$ and every $u \geq 1$,*

$$P\big[|f(X)| \geq Lu \left\|f\right\|_{L_2}\big] \leq 2\exp(-u^2).$$

*We define the **Gaussian mean width of** F by $\ell^*(F) = \mathbb{E}\sup_{f,g\in F} G_{f-g}$ where $(G_{f-g} : f, g \in F)$ is the canonical Gaussian process associated with the set $F - F = \{f - g : f, g \in F\}$.*

Note that the Gaussian mean width is the natural complexity parameter associated with a L-subgaussian class of functions. Assuming other deviation properties for F than the sub-gaussian one will result in other complexity parameters. A direction that will not be pursued here.

We are now in position to recall a result on multiplier processes that will be used to prove property $(\mathbf{L}(\rho))$ and to check point *ii)* and *iii)* in Assumption 1.2.

Theorem 3.3 (cf. Theorem 3.10 in [30]) *There are two absolute constants c_1 and c_2 such that the following holds. Let $X_1, \ldots, X_N$ be iid copies of some vector X and let H be a L-sub-gaussian class of functions with respect to X. Let $\zeta_1, \ldots, \zeta_N$ be iid copies of some real-valued random variable ζ such that $\zeta \in L_q$ for some $q > 2$. Then for all $u \geq 1$, with probability larger than $1 - (c_1/u)^q$,*

$$\sup_{h\in H}\Big|\frac{1}{N}\sum_{i=1}^N \zeta_i h(X_i) - \mathbb{E}\zeta h(X)\Big| \leq c_2(u\log(eu))\left\|\zeta\right\|_{L_q}\frac{\ell^*(H)}{\sqrt{N}}.$$

Note that in Theorem 3.3, the X_i's and ζ_i's do not have to be independent – this may be the case when the "noise" $\zeta := Y - f^*(X)$ depends on X, which is in general the case in learning theory. The strategy that we use in the applications of Theorem A below to prove $(\mathbf{L}(\rho))$ and check Assumption 1.2 is based on the following result.

Theorem 3.4 *Let F be a convex class, $\rho > 0$ and $u \geq 1$. Assume that F is a L-subgaussian class with respect to X for some $L \geq 1$ and that $\zeta = Y - f^*(X) \in L_q$ for some $q > 2$. Then with probability at least $1 - 2(c_1/u)^q$,*

1. *$(\boldsymbol{L}(\rho))$ holds for $s_L(\rho) = s_L$ and $\kappa_1 = \kappa_1(u) = c_2 u\log(eu)$ when $s_L \geq 0$ is such that*
$$\left\|\zeta\right\|_{L_q}\ell^*\big(F_\rho \cap (f^* + s_L\mathcal{D})\big) \leq \sqrt{N}s_L^2.$$
2. *point ii) of Assumption 1.2 is satisfied when*
$$\lambda\rho \geq \frac{2\eta_1\kappa_1(u)\left\|\zeta\right\|_{L_q}}{C_0 - 2\eta_1 - 1}\frac{\ell^*\big(F_{\eta_1 C_0\rho}\cap(f^* + c_0 s(\eta_1^2 C_0\rho)\mathcal{D})\big)}{\sqrt{N}}.$$

Proof. Since F is convex, by definition of f^*, one has for every $f \in F$: $\mathbb{E}(f^* - f)(X)(f^*(X) - Y) \leq 0$. Therefore, for any $\rho, s \geq 0$, $\phi_N(\rho, s)$ is smaller than

$$\sup_{f\in F_\rho\cap(f^*+s\mathcal{D})}\frac{1}{N}\sum_{i=1}^N (f^* - f)(X_i)(f^*(X_i) - Y_i) - \mathbb{E}(f^* - f)(X)(f^*(X) - Y).$$

The last quantity can be bounded for any $(\rho, s) \in \{(\rho, s_L), (\eta_1 C_0 \rho, c_0 s^2(\eta_1^2 C_0 \rho))\}$ using Theorem 3.3 and the result follows. ■

4 Learning linear functional in Hilbert spaces

In this section, we assume that the data are N iid couples $(Y_i, X_i)_{i=1}^N$ distributed like (Y, X) where Y is a real-valued output and X is a random vector in a Hilbert space $\mathcal{H}$. The inner product in $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle$ and its associated norm is $\|\cdot\|_2$. We also denote by $B_2 = \{t \in H : \|t\|_2 \leq 1\}$ the unit ball of H. Classical examples studied later are: ℓ_2^d, the space $\mathbb{R}^d$ endowed with the Euclidean norm; $\mathbb{R}^{m \times T}$ the space of $m \times T$ matrices endowed with the Frobenius norm and Reproducing Kernel Hilbert Spaces. We denote by $\Sigma = \mathbb{E} X X^\top$ the covariance operator of X and by $\mathcal{E}$ the associated ellipsoid: $\mathcal{E} = \{t \in \mathcal{H} : \mathbb{E}\langle X, t \rangle^2 \leq 1\}$.

We consider $T \subset \mathcal{H}$ a closed and convex set and denote

$$t^* \in \operatorname*{argmin}_{t \in T} \mathbb{E}(Y - \langle X, t \rangle)^2$$

so that $\langle X, t^* \rangle$ is the best linear approximation of Y in L_2 restricted to vectors in T. We want to obtain estimation results on t^* using Theorem A. Note that when $T = \mathcal{H}$, any $t \in \mathcal{H}$ is such that $\mathbb{E}(Y - \langle X, t \rangle)^2 = \mathbb{E}\langle X, t - t^* \rangle^2 + \mathbb{E}(Y - \langle X, t^* \rangle)^2$, so that predicting the best output associated with X by linear forms like $\langle X, t \rangle$ is equivalent to estimate $\langle X, t^* \rangle$ in L_2: prediction of Y and estimation of $\langle X, t^* \rangle$ are therefore equivalent task.

Let $\|\cdot\|$ be a function on $\mathcal{H}$ satisfying Assumption 1.1. We want to estimate t^* knowing that $t^* \in T$ w.r.t. the semi-norm $(\mathbb{E}\langle X, \cdot \rangle^2)^{1/2}$ (or equivalently to estimate $f^*(\cdot) = \langle \cdot, t^* \rangle$ in $L_2(\mu)$) and "believing" that $\|t\|$ is small. To that end, we consider the regularization procedure

$$\hat{t} \in \operatorname*{argmin}_{t \in T} \Big(\frac{1}{N} \sum_{i=1}^N \big(Y_i - \langle X_i, t \rangle\big)^2 + \lambda \|t\| \Big) \tag{4.1}$$

for some regularization parameter λ chosen such that Assumption 1.2 should hold with large probability.

In order to obtain prediction results for $\hat{t}$ by applying Theorem A, we need to check properties $(\mathbf{Q}(\rho))$ and $(\mathbf{L}(\rho))$ and to obtain bounds on ϕ_N.

Property $(\mathbf{Q}(\rho))$ follows from Section 2 if we assume that the design X satisfies the small ball property: there exists u_0 and β_0 such that for all $t \in \mathcal{H}$,

$$P\big[|\langle X, t \rangle| \geq u_0 \|\langle X, t \rangle\|_{L_2}\big] \geq \beta_0. \tag{4.2}$$

The study of ($\mathbf{L}(\rho)$) follows from some bounds on ϕ_N which requires some moments on the design and the noise. We study in the next sections two different types of such assumptions as in Section 3.

4.1 Results for a sub-Gaussian design and noise in L_q, $q > 2$

In this section ,we assume that the design X is sub-gaussian and the noise $\zeta = Y - \langle X, t^* \rangle$ is in L_q for some $q > 2$ – but we do not assume that ζ is independent of X, in particular, we do not assume any statistical model.

We recall that the random vector X is a **L-sub-Gaussian vector** for some $L > 0$, in $\mathcal{H}$ when for every $t \in \mathcal{H}$ and $u \geq 1$,

$$P\big[|\langle X, t \rangle| \geq Lu \|\langle X, t \rangle\|_{L_2}\big] \leq 2\exp(-u^2). \tag{4.3}$$

Under the sub-gaussian assumption (4.3), the small ball property (4.2) is satisfied for $u_0 = 1/2$ and $\beta_0 = \big(3/(4(2\sqrt{2}eL)^2)\big)^2$. Indeed, it follows from (4.3) that, for any $t \in \mathcal{H}$, $\big\|\langle X, t \rangle\big\|_{L_4} \leq 2\sqrt{2}eL \, \big\|\langle X, t \rangle\big\|_{L_2}$ (cf., for instance, Theorem 1.1.5 in [8]). Then (4.2) follows from the Paley-Zygmund inequality (cf. Corollary 3.3.2 in [11]). In particular, we can apply Theorem 2.3 to prove that condition ($\mathbf{Q}(\rho)$) is satisfied when (4.3) holds as long as there exists some level s_Q such that (2.2) holds.

As in the previous sections, a key role will be played by the nested family of sub-models $(T_\rho)_{\rho \geq 0}$ where $T_\rho = \{t \in T : \|t\| \leq \rho\}$. Moreover, under the sub-gaussian assumption on the design, the complexity parameters of the problem (appearing in both the regularization parameter and the rates of convergence) are driven by the local and global Gaussian mean widths of the sub-models T_ρ for all $\rho \geq 0$. This quantity was introduced in Definition 3.2 for function classes. In the case where T is a subset of a Hilbert space $\mathcal{H}$, a simple construction of the Gaussian mean width is given as follows:

$$\ell^*(T) = \mathbb{E} \sup_{u, v \in T} \langle G, u - v \rangle \tag{4.4}$$

where G is a standard (centered) Gaussian vector of $\mathcal{H}$.

We can recast the problem of learning linear functional in a Hilbert space in the general setup considered in Section 1, by considering the class $F = \{\langle \cdot, t \rangle : t \in T\}$ and the linear function $f^*(\cdot) = \langle \cdot, t^* \rangle \in F$. It follows from Theorem A together with Theorem 3.4, for condition ($\mathbf{L}(\rho)$) and Theorem 2.3, for condition ($\mathbf{Q}(\rho)$) that the following result holds for the problem of learning linear functional in a Hilbert space for a subgaussian design and a noise in L_q, $q > 2$.

Theorem 4.1 *There are absolute constants c_1 and c_2 such that the following holds. Let $s_L(\cdot)$ and $s_Q(\cdot)$ be two non-decreasing functions such that for every $\rho \geq 0$:*

- $\star$ $\sigma_q \ell^*\big(T_{\eta_1^4\rho} \cap s_L(\rho)\mathcal{E}\big) \leq \sqrt{N} s_L(\rho)^2$, *where* $\sigma_q = \big\|Y - \big\langle X, t^*\big\rangle\big\|_{L_q}$,
- $\star\star$ $\ell^*\big(T_{\eta_1^4\rho} \cap s_Q(\rho)\mathcal{E}\big) \leq \frac{c_1\beta_0 u_0}{L}\sqrt{N} s_Q(\rho)$

where $u_0 = 1/2$ and $\beta_0 = \big(3/(4(2\sqrt{e}L))^2\big)^2$ have been introduced in (4.2). Let $s(\rho) \geq \max\big(s_L(\rho), s_Q(\rho)\big)$. Let $u \geq 1$, $\kappa_1(u) = c_2(u\log(eu))$, $\kappa_0 = u_0\beta_0^2/2$ and some $c_0 \geq \max\big(4\kappa_1(u)/\kappa_0, 1\big)$ and $C_0 > 2\eta_1 + 1$. Let $\lambda > 0$ be such that:

1. $2\lambda\rho \leq \kappa_0 c_0^2 s^2\big(\eta_1^2 C_0\rho\big)$,
2. $$\lambda\rho \geq \frac{2\eta_1\kappa_1(u)\sigma_q}{C_0 - 2\eta_1 - 1}\frac{\ell^*\big(F_{\eta_1 C_0\rho} \cap (f^* + c_0 s(\eta_1^2 C_0\rho)\mathcal{D})\big)}{\sqrt{N}}.$$

Let $R^* > 0$ and consider the event Ω^* on which, when $\|t^*\| = 0$,

$$\frac{\lambda}{2\eta_1} \geq \sup_{0<r\leq\eta_1 C_0 R^*} \frac{\phi_N\big(\eta_1 r, c_0 s(\eta_1^2 C_0 R^*)\big)}{r} \tag{4.5}$$

Then, for this choice of regularization parameter λ, the regularization procedure $\hat{t}$ defined in (4.1) is such that with probability larger than $1-2\exp(-N\beta_0^2/8)-5(c_2/u)^q - P\big[(\Omega^*)^c\big]$,

$$\big\|\big\langle X, \hat{t} - t^*\big\rangle\big\|_{L_2} \leq c_0 s(\eta_1^2 C_0 \|t^*\|) \text{ and } \big\|\hat{t}\big\| \leq \eta_1 C_0 \|t^*\|.$$

Proof. By assumption, $\|\cdot\|$ satisfies Assumption 1.1. Moreover, thanks to Theorem 3.4, λ satisfies Assumption 1.2 with probability larger than $1-(c_1/u)^q - P\big[(\Omega^*)^c\big]$. Hence, in order to apply Theorem A, it only remains to check conditions ($\mathbf{L}(\rho^*)$) and ($\mathbf{Q}(\rho^*)$) for both $\rho^* \in \{\eta_1^2 C_0 \|f^*\|, \eta_1^2 C_0 R^*\}$. We first start by proving that ($\mathbf{Q}(\rho^*)$) holds thanks to Theorem 2.3.

Under the sub-gaussian assumption (4.3), a generic chaining argument (cf. Chapter 1 in [40]) shows that

$$\mathbb{E}\sup_{t\in(T_{\eta_1\rho^*}-t^*)\cap s_Q(\rho^*)\mathcal{E}}\Big|\frac{1}{N}\sum_{i=1}^N \varepsilon_i\big\langle X_i, t\big\rangle\Big| \leq \frac{c_5 L}{\sqrt{N}}\ell^*\big((T_{\eta_1\rho^*} - t^*) \cap s_Q(\rho^*)\mathcal{E}\big)$$

for some absolute constant c_5. Therefore, by definition of $s_Q(\cdot)$ in $\star\star$ and since $T_{\eta_1\rho^*} - T_{\eta_1\rho^*} \subset T_{\eta_1^2\rho^*}$, we have, for an appropriate choice of constant c_1 in $\star\star$,

$$\mathbb{E}\sup_{t\in(T_{\eta_1\rho^*}-t^*)\cap s_Q(\rho^*)\mathcal{E}}\Big|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\big\langle X_i,t\big\rangle\Big|\le\frac{\beta_0u_0}{8}s_Q(\rho^*).$$

It follows from the small ball property (4.2) and Theorem 2.3 that $(\mathbf{Q}(\rho^*))$ holds with probability larger than $1-\exp(-\beta_0^2N/8)$ for $\kappa_0 = u_0\beta_0^2/2$.

Then, it follows from Theorem 3.4 that $(\mathbf{L}(\rho^*))$ holds with probability larger than $1-(c_1/u)^q$ for the function $s_L(\cdot)$ defined in $\star$ and $\kappa_1 = c_2(u\log(eu))$ where c_2 is the constant appearing in Theorem 3.3 since $T_{\rho^*} - f^* \subset T_{\eta_1\rho^*}$. Finally, the result follows from Theorem A. ■

In order to obtain rates of convergence and model selection properties for $\hat{t}$ for the problem of learning linear functional with a noise $\zeta = Y - f^*(X)$ in L_q, $q>2$ and a sub-Gaussian design X, one may apply Theorem 4.1. For that matter, it is enough to compute the local Gaussian mean widths $\ell^*(T_\rho\cap s\mathcal{E})$ for any $\rho\ge 0$ and $s\ge 0$ and to find a probability estimate for the event Ω^* in (4.5). Rates of convergence and regularization functions follow from these two unique quantities.

Controlling these two quantities when $\|\cdot\|$ is sub-linear and $T=\mathcal{H}$ easily follows from Theorem 4.1. We now state this result when $\|\cdot\|$ is such that for every $x,y\in\mathcal{H}$ and $\lambda\ge 0$,

$$\|x\| = \|-x\|\,,\qquad \|x+y\|\le\eta_1\big(\,\|x\|+\|y\|\,\big)\text{ and }\|\lambda x\|\le\lambda\,\|x\|\,. \tag{4.6}$$

Note that since $T=\mathcal{H}$, the next result also provides a prediction result.

Theorem 4.2 *Let $q>2$, $u\ge 1$, $c_0, C_0>3, c_1,\beta_0$ and u_0 be the constants introduced in Theorem 4.1. Let $\|\cdot\|$ satisfying (4.6) and denote by $B_{\|\cdot\|}=\{t\in\mathcal{H}:\|t\|\le 1\}$ its unit ball for some $L>0$. Assume that X is L-subgaussian. Set $\sigma_q = \big\|Y-\big\langle X,t^*\big\rangle\big\|_{L_q}$ and $\kappa_1(u)=c_1u\log(eu)$. Consider the RERM*

$$\hat{t}\in\underset{t\in\mathcal{H}}{\operatorname{argmin}}\Big(\frac{1}{N}\sum_{i=1}^{N}(Y_i-\big\langle X_i,t\big\rangle)^2+2\eta_1^3\kappa_1(u)\sigma_q\,\|t\|\,\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}\Big).$$

Then, for c_0 large enough, with probability at least $1-2\exp(-c_0N)-5(c_1/u)^q$,

$$\big\|\big\langle X,\hat{t}-t^*\big\rangle\big\|_{L_2}\le c_0s(\eta_1^2C_0\,\|t^*\|)\ \textit{and}\ \big\|\hat{t}\big\|\le\eta_1C_0\,\|t^*\|$$

where $s^2(\rho) = 2\eta_1^4\kappa_1(u)\sigma_q\rho\ell^*(B_{\|\cdot\|})/\sqrt{N}$ *when* $N \gtrsim \ell^*(\mathcal{E})^2$ *and when* $N \lesssim \ell^*(\mathcal{E})^2$,

$$s^2(\rho) = \max\Big(2\eta_1^4\kappa_1(u)\sigma_q\rho\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}, \frac{\eta_1^8 L^2}{c_1\beta_0^2u_0^2}\rho^2\frac{\ell^*(B_{\|\cdot\|})^2}{N}\Big). \tag{4.7}$$

Proof. Theorem 4.2 will follow from Theorem 4.1 after controlling the probability estimate of the event introduced in (4.5) and checking that conditions of Theorem 4.1 are satisfied by $s(\cdot)$ defined in (4.7) and

$$\lambda = 2\eta_1^3\kappa_1(u)\sigma_q\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}. \tag{4.8}$$

First note that $T_{\eta_1^4\rho} \subset \eta_1^4\rho B_{\|\cdot\|}$ and so $\ell^*(\eta_1^4\rho \cap s\mathcal{E}) \leq \eta_1^4\rho\ell^*(B_{\|\cdot\|})$ for any $s \geq 0$. Therefore, if one takes

$$s_L^2(\rho) = \frac{\eta_1^4\rho\sigma_q\ell^*(B_{\|\cdot\|})}{\sqrt{N}} \text{ and } s_Q^2(\rho) = \begin{cases} 0 & \text{if } (c_1\beta_0u_0)^2N \geq L\ell^*(\mathcal{E})^2 \\ \frac{L^2\eta_1^8\rho^2\ell^*(B_{\|\cdot\|})^2}{c_1^2\beta_0^2u_0^2N} & \text{otherwise.} \end{cases}$$

both conditions $\star$ and $\star\star$ of Theorem 4.1 are satisfied.

Now, we turn to controlling the probability measure of the event Ω^*. Let $R^* > 0$ and denote $s = c_0s(\eta_1^2C_0R^*)$. We have to control the probability measure of the event Ω^* introduced in (4.5) when $\|t^*\| = 0$. Using the sub-linearity of $\|\cdot\|$ from (4.6), we get

$$\begin{aligned}
&\sup_{0<r\leq\eta_1C_0R^*}\frac{\phi_N(\eta_1r, s)}{r} \\
&= \sup\Big(P_N\big\langle\cdot, \frac{t^*-t}{r}\big\rangle(\big\langle\cdot, t^*\big\rangle - Y) : 0 < r \leq \eta_1C_0R^*, \|t\| \leq \eta_1r, \mathbb{E}\big\langle X, t-t^*\big\rangle^2 \leq s^2\Big) \\
&\leq \sup_{0<r\leq\eta_1C_0R^*}\ \sup_{t:\|t\|\leq\eta_1^2r,} P_N\big\langle\cdot, \frac{t}{r}\big\rangle(-\zeta) \leq \eta_1^2\sup_{\|t\|=1}\frac{1}{N}\sum_{i=1}^N(-\zeta_i)\big\langle X_i, t\big\rangle.
\end{aligned} \tag{4.9}$$

Then, using Theorem 3.3 and the same argument as in the proof of Theorem 3.4, we obtain that, with probability at least $1 - (c_1/u)^q$,

$$\sup_{0<r\leq\eta_1C_0R^*}\frac{\phi_N(\eta_1r, s)}{r} \leq \eta_1^2\kappa_1(u)\sigma_q\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}$$

for $\kappa_1(u) = c_2u\log(eu)$. Therefore, if one chooses λ such that point *1)* and *2)* in Theorem 4.1 holds and

$$\lambda \geq 2\eta_1^3\kappa_1(u)\sigma_q\ell^*(B_{\|\cdot\|})/\sqrt{N}$$

are satisfied then we can apply Theorem 4.1. It appears that for the choices of λ in (4.8) and $s(\cdot)$ in (4.7), these conditions are satisfied. ■

We will provide several examples of applications of Theorem 4.2 in what follows. A typical example is given now for the Lasso that is when $\mathcal{H} = \mathbb{R}^d$ and the regularization function is the ℓ_1^d-norm. Note that the following result holds under no statistical model and requires only that the noise $\zeta = Y - \left\langle X, t^* \right\rangle$ is in L_q for some $q > 2$ and the design X is sub-Gaussian.

Theorem 4.3 *Assume that X is a L-sub-gaussian random vector in $\mathbb{R}^d$. Assume that the noise $\zeta = Y - \left\langle X, t^* \right\rangle$ is in L_q for some $q > 2$ and denote $\sigma_q = \left\| Y - \left\langle X, t^* \right\rangle \right\|_{L_q}$. Let $u > 1$ and $\kappa_1(u) = c_1 u \log(eu)$. Then, the Lasso*

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \Big(\frac{1}{N} \sum_{i=1}^N \big(Y_i - \left\langle X_i, t \right\rangle \big)^2 + 2\kappa_1(u) \sigma_q \left\| t \right\|_1 \sqrt{\frac{\log(ed)}{N}} \Big)$$

is such that with probability at least $1 - 2\exp(-\beta_0^2 N) - 5(c_1/u)^q$,

$$\left\| \left\langle X, \hat{t} - t^* \right\rangle \right\|_{L_2} \leq c_0 s(C_0 \left\| t^* \right\|_1) \ \textit{and} \ \left\| \hat{t} \right\|_1 \leq C_0 \left\| t^* \right\|_1$$

where $s^2(\rho) \sim \kappa_1(u) \sigma_q \rho \sqrt{\log(ed)/N}$ *when* $N \gtrsim \ell^*(\mathcal{E})^2$ *and when* $N \lesssim \ell^*(\mathcal{E})^2$,

$$s^2(\rho) \sim \max \Big(\kappa_1(u) \sigma_q \rho \sqrt{\frac{\log(ed)}{N}}, \rho^2 \frac{\log(ed)}{N} \Big).$$

In the next section, we show that a similar result holds when a weaker moment assumption on the design X is satisfied but under the assumption of a Statistical model. Comparing Theorem 4.3 with the other classical estimation results for the Lasso (cf. [2], Chapter 8.2 in [19] or Chapter 6.2 in [4] among others), it appears that even under weak moment assumptions on the noise and no Statistical model, we still consider the same regularization parameter: λ is of the order of $\sigma\sqrt{\log(ed)/N}$ where σ measures the "variance" of the noise in different scenarii. About the optimality of the result in Theorem 4.3, it appears that the rate of convergence $s(C_0 \left\| t^* \right\|)$ is, up to some log factor, the minimax rate of convergence in $C_0 \left\| t^* \right\|_1 B_1^d$ as proved in Section 5.1 of [22].

4.2 Results under moments assumption on the design and independent noise in $L_{2,1}$ for the Lasso

In this section, we assume that the X_i's take their values in $\mathbb{R}^d$ and that a statistical model holds:

$$Y = \left\langle X, t^* \right\rangle + \zeta \tag{4.10}$$

where ζ is a mean zero noise independent of X and $t^* \in T \subset \mathbb{R}^d$. We also consider a regularization function $\|\cdot\|$ defined on $\mathbb{R}^d$ satisfying Assumption 1.1.

Results similar to Theorem 4.1 and Theorem 4.2 can be obtained in this setup where the (local and global) Gaussian mean width is replaced by the complexity parameter (3.3) and the noise level $\sigma_q = \|\zeta\|_{L_q}$ is replaced by $\|\zeta\|_{2,1}$. We do not reproduce here these two results for the sake of shortness. Instead, we show how to get a result in the special case of the Lasso under the small ball property and moment assumptions in model (4.10) via Theorem A.

Theorem 4.4 *Assume that X is a random vector in $\mathbb{R}^d$ satisfying the small property* (4.2) *in $\mathcal{H} = \mathbb{R}^d$ for some β_0 and u_0. Assume that the coordinates of $X = (x_1, \ldots, x_d)$ have $c_0 \log(ed)$ sub-gaussian moments for some $c_0 > 1$: for every $1 \leq j \leq d$, $\|x_j\|_{L_p} \leq \kappa\sqrt{p}$ for every $1 \leq p \leq c_0 \log(ed)$. Let $0 < \delta < 1$. Then, in the statistical model* (4.10) *with an independent noise ζ such that $\|\zeta\|_{2,1}$ is finite, the Lasso*

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \Big(\frac{1}{N} \sum_{i=1}^N \big(Y_i - \langle X_i, t\rangle\big)^2 + \frac{c_1 \|\zeta\|_{2,1}}{\delta} \|t\|_1 \sqrt{\frac{\log(ed)}{N}} \max_{1 \leq j \leq d} \|x_j\|_{L_2} \Big)$$

is such that with probability at least $1 - 4\delta - 2\exp(-\beta_0^2 N)$,

$$\big\|\langle X, \hat{t} - t^*\rangle\big\|_{L_2} \leq c_0 s(C_0 \|t^*\|_1) \text{ and } \big\|\hat{t}\big\|_1 \leq C_0 \|t^*\|_1$$

where $s^2(\rho) \sim \|\zeta\|_{2,1} \rho \sqrt{\log(ed)/N} \max_{1 \leq j \leq d} \|x_j\|_{L_2}$ when $N \gtrsim \mathbb{E}\|X\|_{\ell_2^d}^2$ and when $N \lesssim \mathbb{E}\|X\|_{\ell_2^d}^2$,

$$s^2(\rho) = \max\Big(\frac{c_1 \|\zeta\|_{2,1}}{\delta} \rho \sqrt{\frac{\log(ed)}{N}} \max_{1 \leq j \leq d} \|x_j\|_{L_2}, \rho^2 \frac{\log(ed)}{N} \Big).$$

Again comparing Theorem 4.4 with the classical Lasso procedure studied in many works, it appears that even under weak moment assumptions on the design and the noise, we still consider a same regularization parameter of the order of $\sigma\sqrt{\log(ed)/N}$.

Proof. First note that $\|\cdot\|_1$ is a norm so it satisfies Assumption 1.1 for $\eta_1 = 1$. Then, to apply Theorem A, we need to check properties ($\mathbf{L}(\rho)$) and ($\mathbf{Q}(\rho)$) for $\rho \in \{C_0 \|t^*\|_1, C_0 R^*\}$ for some $C_0 > 3$ and $R^* > 0$ and to choose λ and some function $s(\cdot)$ so that Assumption 1.2 is satisfied. Thanks to Theorem 2.3 and Proposition 3.1 this will follow from a bound on the quantity in (3.3) that we control now.

Let $1 \le k \le N$, $\rho^* \ge \|t^*\|_1$ and $s > 0$. We have

$$(\star) = \mathbb{E} \sup_{h \in (F_{\rho^*} - f^*) \cap s\mathcal{D}} \Big| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i h(X_i) \Big| = \mathbb{E} \sup_{t \in (\rho^* B_1^d - t^*) \cap s\mathcal{E}} \Big| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i \big\langle X_i, t \big\rangle \Big|$$

$$\le \min \left(2\rho^* \mathbb{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i X_i \right\|_{\ell_\infty^d}, s\mathbb{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i X_i \right\|_{\ell_2^d} \right).$$

Therefore, to obtain a bound on $(\star)$, we just have to control the two last expectations in the minimum. For the second expectation, we have:

$$\mathbb{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i X_i \right\|_{\ell_2^d} \le \left(\mathbb{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i X_i \right\|_{\ell_2^d}^2 \right)^{1/2} \le \big(\mathbb{E} \|X\|_{\ell_2^d}^2 \big)^{1/2}.$$

For the first one, we prove two intermediate results.

Proposition 4.5 *Let z be a mean zero variance one random variable and let $z_1, \ldots, z_N$ be N iid copies of z. Assume that there exists $\kappa > 0$ and $p_0 \ge 2$ such that $\|z\|_{L_p} \le \kappa\sqrt{p}$ for every $2 \le p \le p_0$. Then, for some absolute constant c_0, for every $1 \le k \le N$ and every $1 \le p \le p_0$,*

$$\left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k z_i \right\|_{L_p} \le c_0 \kappa \sqrt{p}.$$

Proof. Let $p \le p_0$. It follows from Latała's inequality (cf. [21]) that

$$\left\| \sum_{i=1}^k z_i \right\|_{L_p} \sim \sup \Big(\frac{p}{s} \Big(\frac{k}{p} \Big)^{1/s} \|z\|_{L_s} : \max(2, p/k) \le s \le p \Big). \tag{4.11}$$

Let $H(s) = (p/s)(k/p)^{1/s} \kappa \sqrt{s}$. Since H is decreasing, H attains its maximum on the interval $\max(2, p/k) \le s \le p$ at $\max(2, p/k)$ for which it is less than $c_0 \kappa \sqrt{pk}$ whatever is $p \le p_0$ and $1 \le k \le N$. The result then follows from (4.11) and the moment assumption. ■

Lemma 4.6 *Let $X = (x_1, \ldots, x_d)$ be a random vector in $\mathbb{R}^d$ such that for every $1 \le j \le d$ and every $1 \le p \le c_0 \log(ed)$, $\|x_j\|_{L_p} \le \kappa\sqrt{p}$ for some absolute constant κ and $c_0 > 1$. Let $X_1, \ldots, X_N$ be iid copies of X. Then, for every $1 \le k \le N$,*

$$\mathbb{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i X_i \right\|_{\ell_\infty^d} \le c_1 \kappa \sqrt{\log(ed)} \max_{1 \le j \le d} \|x_j\|_{L_2}.$$

Proof. We write $X_i = (x_{ij})_{j=1}^d$ for every $1 \le i \le N$ and $V_j = k^{-1/2}\sum_{i=1}^k x_{ij}$ for every $1 \le j \le d$. We have

$$\mathbb{E}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_i X_i\right\|_{\ell_\infty^d} = \mathbb{E}\max_{1\le j\le d}|V_j|.$$

For every $1 \le j \le d$, $(\varepsilon_i x_{ij})_{i=1}^N$ is a family of iid mean zero random variables distributed like εx_j and such that $\|\varepsilon x_j\|_{L_p} = \|x_j\|_{L_p} \le \kappa\sqrt{p}$ for any $1 \le p \le c_0\log(ed)$. It follows from Proposition 4.5 that $\|V_j\|_{L_p} \le c_1\kappa\sqrt{p}\,\|x_j\|_{L_2}$ for every $1 \le p \le c_0\log(ed)$ where c_1 is some absolute constant. In particular, it follows from Markov inequality that for every $u > 0$ and for every $1 \le p \le c_0\log(ed)$,

$$\begin{aligned} P\big[\max_{1\le j\le d}|V_j| \ge u\big] &\le \sum_{j=1}^d P[|V_j| \ge u] \le \sum_{j=1}^d \Big(\frac{\|V_j\|_{L_p}}{u}\Big)^p \\ &\le d\Big(\frac{c_1\kappa\sqrt{p}\max_{1\le j\le d}\|x_j\|_{L_2}}{u}\Big)^p. \end{aligned}$$

Hence, for $p = c_0\log(ed)$ and $u = tc_1\kappa\sqrt{p}\max_{1\le j\le d}\|x_j\|_{L_2}$ for some $t > 0$,

$$P\Big[\max_{1\le j\le d}|V_j| \ge c_1 t\kappa\sqrt{p}\max_{1\le j\le d}\|x_j\|_{L_2}\Big] \le \Big(\frac{e}{t}\Big)^{c_0\log(ed)}.$$

Now the result follows by integrating the last inequality. ■

Therefore, we obtain

$$(\star) \le \min\Big(c_1\rho^*\kappa\sqrt{\log(ed)}\max_{1\le j\le d}\|x_j\|_{L_2}, s\big(\mathbb{E}\|X\|_{\ell_2^d}^2\big)^{1/2}\Big). \tag{4.12}$$

Conditions required to apply Theorem A follow from this result. Indeed, let us first check condition $(\mathbf{Q}(\rho))$.

If $\beta_0 u_0\sqrt{N} \ge 8\big(\mathbb{E}\|X\|_{\ell_2^d}\big)^{1/2}$ then (2.2) is satisfied for every $s_Q \ge 0$ so one can take $s_Q(\rho^*) = 0$. If not then (2.2) is satisfied for

$$s_Q^2(\rho) = \frac{c_3}{\beta_0 u_0}\rho^2\max_{1\le j\le d}\|x_j\|_{L_2}\frac{\log(ed)}{N}. \tag{4.13}$$

It also follows from (4.12) that (3.1) is satisfied for $s_L = s_L(\rho^*)$ when

$$s_L^2(\rho) = 2\rho\|\zeta\|_{2,1}\sqrt{\frac{\log(ed)}{N}}\max_{1\le j\le d}\|x_j\|_{L_2} \tag{4.14}$$

and (3.2) holds as well when

$$\lambda \geq \frac{c_3 \left\|\zeta\right\|_{2,1}}{\delta(C_0-3)}\sqrt{\frac{\log(ed)}{N}} \max_{1\leq j\leq d}\left\|x_j\right\|_{L_2}. \tag{4.15}$$

Therefore, it follows from Theorem 2.3 and Proposition 3.1 that for $\rho \in \{C_0 \left\|t^*\right\|_1, C_0R^*\}$, $\kappa_0 = u_0\beta_0^2/2$ and $\kappa_1 = \kappa_1(\delta) = c_1/\delta$, $(\mathbf{Q}(\rho))$, $(\mathbf{L}(\rho))$ and point *ii)* of Assumption 1.2 are satisfied with probability at least $1 - 2\exp(-\beta_0^2 N/8) - 2\delta$.

Now, we turn to point *iii)* of Assumption 1.2. Following the same argument as in (4.9), we obtain

$$\sup_{0<r\leq C_0R^*}\frac{\phi_N(r, c_0 s(C_0R^*))}{r} \leq \sup_{\left\|t\right\|_1=1}\Big|\frac{1}{N}\sum_{i=1}^N \zeta_i\langle X_i, t\rangle\Big|.$$

Following the same argument as in the proof of Proposition 3.1 (that is Markov inequality and Lemma 2.9.1 from [46]), we obtain that with probability larger than $1-\delta$,

$$\sup_{0<r\leq C_0R^*}\frac{\phi_N(r, c_0 s(C_0R^*))}{r} \leq c_4\left\|\zeta\right\|_{2,1}\kappa\sqrt{\frac{\log(ed)}{N}}\max_{1\leq j\leq d}\left\|x_j\right\|_{L_2}.$$

Therefore, if we choose λ such that

$$\lambda \geq \frac{2c_4\left\|\zeta\right\|_{2,1}\kappa}{\delta}\sqrt{\frac{\log(ed)}{N}}\max_{1\leq j\leq d}\left\|x_j\right\|_{L_2} \tag{4.16}$$

then point *iii)* of Assumption 1.2 holds with probability at least $1-\delta$. Hence, there exists c_5 an absolute constant large enough so that (4.15) and (4.16) are satisfied for

$$\lambda = \frac{c_5\left\|\zeta\right\|_{2,1}\kappa}{(C_0-3)\delta}\sqrt{\frac{\log(ed)}{N}}\max_{1\leq j\leq d}\left\|x_j\right\|_{L_2}. \tag{4.17}$$

Finally, point *i)* of Assumption 1.2 holds when we choose the function $s(\cdot)$ such that

$$\begin{aligned} s^2(\rho) &= \max\left(s_L^2(\rho), s_Q^2(\rho), \frac{2\lambda\rho}{C_0\kappa_0c_0^2}\right)\\ &= \begin{cases} \frac{2\lambda\rho}{C_0\kappa_0c_0^2} & \text{if } N \geq \left(\frac{8}{u_0\beta_0}\right)^2\mathbb{E}\left\|X\right\|_{\ell_2^d}\\ \max\left(\frac{\rho^2}{\beta_0u_0}\frac{\log(ed)}{N}\max_{1\leq j\leq d}\left\|x_j\right\|_{L_2}, \frac{2\lambda\rho}{C_0\kappa_0c_0^2}\right) & \text{otherwise.}\end{cases}\end{aligned}$$

■

5 Regularization methods in $\mathbb{R}^d$

In this section, we consider the learning theory setup of Section 4 (in particular, we do not assume that a statistical model holds) where the input variables X_i's belong to $\mathbb{R}^d$ and are L-sub-gaussian for some $L > 0$ and the noise $\zeta = Y - \left\langle X, t^* \right\rangle$ is in L_q for some $q > 2$. In this framework and for a regularization function $\|\cdot\|$ satisfying (4.6), it follows from Theorem 4.2 that the regularization method

$$\hat{t} \in \operatorname*{argmin}_{t \in \mathbb{R}^d} \Big(\frac{1}{N} \sum_{i=1}^N (Y_i - \left\langle X_i, t \right\rangle)^2 + 2\eta_1^3 \kappa_1(u) \sigma_q \|t\| \frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}} \Big) \tag{5.1}$$

is such that, for some large enough constant c_0 and c_1, with probability at least $1 - 2\exp(-c_0 N) - 5(c_1/u)^q$,

$$\left\| \left\langle X, \hat{t} - t^* \right\rangle \right\|_{L_2} \leq c_0 s(\eta_1^2 C_0 \|t^*\|) \text{ and } \left\| \hat{t} \right\| \leq \eta_1 C_0 \|t^*\|$$

where $s^2(\rho) \sim \kappa_1(u) \sigma_q \rho \ell^*(B_{\|\cdot\|})/\sqrt{N}$ when $N \gtrsim \ell^*(\mathcal{E})^2$ and when $N \lesssim \ell^*(\mathcal{E})^2$,

$$s^2(\rho) = \max\Big(\kappa_1(u) \sigma_q \rho \frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}, \rho^2 \frac{\ell^*(B_{\|\cdot\|})^2}{N} \Big).$$

As a consequence, any time the Gaussian mean width of the unit ball $B_{\|\cdot\|}$ is known, one can derive an estimation result for the regularization method (5.1) associated with $\|\cdot\|$ thanks to Theorem 4.2. In the next section, we apply this result to some classical examples.

5.1 ℓ_p-regularization for $0 < p \leq \infty$

In this section, we study rates of convergence and model selection properties of $\hat{t}$ for the regularization functions $\|t\| = \|t\|_p$ for $p > 0$. Note that Assumption 1.1 for $\eta_1 = 1$ is satisfied because $\|\cdot\|_p$ is a norm when $p \geq 1$ and for $\eta_1 = 2^{1/p}$ for $0 < p < 1$ because $\|\cdot\|_p$ is a p-norm for $0 < p < 1$ (cf. page 2 in [13]).

When $p < 1$, ℓ_p-regularization functions, even though being non-convex, received a particular attention for the fixed design model (cf. [36, 37, 47] and also [12] in the sequence space model) and the random design model in [48]. Denote by $(e_1, \ldots, e_d)$ the canonical basis of $\mathbb{R}^d$. We have $\{\pm e_1, \ldots, \pm e_d\} \subset B_p^d$ so $\ell^*(B_p^d) \sim \sqrt{\log(ed)}$. We therefore recover the same rates of estimation for the ℓ_p-regularization methods than for the Lasso in Theorem 4.3.

The same is true for $1 \leq p \leq 1+(\log(ed))^{-1}$ since there exists an absolute constant c_0 such that $B_1^d \subset B_p^d \subset c_0 B_1^d$, so $\ell^*(B_p^d) \sim \ell^*(B_1^d) \sim \sqrt{\log(ed)}$.

When $1 + (\log(ed))^{-1} \leq p$, define q such that $p^{-1} + q^{-1} = 1$ then by duality $\ell^*(B_p^d) \sim \sqrt{q} d^{1/q}$.

5.2 weak-ℓ_p-regularization for $0 < p \leq 1$

We recall the definition

$$\|t\|_{p\infty} = \max_{1 \leq j \leq d} j^{1/p} t_j^* \text{ and } B_{p\infty}^d = \{t \in \mathbb{R}^d : t_j^* \leq j^{-1/p} \text{ for every } 1 \leq j \leq d\}$$

where $t_1^* \geq t_2^* \geq \ldots \geq t_d^*$ is the non-increasing rearrangement of the absolute values of the coordinates of t. Those quasi-norms have been used in sparse signal recovery for instance in [14].

Proposition 5.1 (cf. Theorem B in [16]) *Set* $0 < p \leq 1$.

$$\ell^*(B_{p\infty}) \lesssim \begin{cases} \frac{\sqrt{\log(ed)}}{p-1} & \text{if } 0 < p < 1 \\ \big(\log(ed)\big)^{3/2} & \text{if } p = 1. \end{cases}$$

5.3 Micchelli, Morales and Pontil's regularization functions

General structured sparsity norms have been introduced in [27] in the following way: let Θ be a nonempty convex cone in $(0,\infty)^d$, define for all $t \in \mathbb{R}^d$

$$\Omega(t|\Theta) = \inf_{\theta \in \Theta} \frac{1}{2} \sum_{j=1}^d \Big(\frac{t_j^2}{\theta_j} + \theta_j \Big). \tag{5.2}$$

It is shown in [27] that $\Omega(\cdot|\Theta)$ is a norm on $\mathbb{R}^d$. Given this particular form of the norm, [27] suggested an alternating minimization algorithm for constructing the regularization function (5.1) with regularization function $\Omega(\cdot|\Theta)$.

Several classical regularization functions can written like $\Omega(\cdot|\Theta)$ for an appropriate choice of cone Θ. For instance, the ℓ_1^d-norm is obtained for $\Theta = (0,\infty)^d$. The group Lasso procedure from [49] is also a special case: if $(G_1, \cdots, G_T)$ is a partition of $\{1, \ldots, d\}$ and

$$\Theta = \{\theta \in (0,\infty)^d \text{ constant within groups } G_t\}$$

then

$$\Omega(t|\Lambda) = \sum_{t=1}^T \sqrt{|G_t|} \, \|t_{|G_t}\|_2$$

where $|G_t|$ is the size of G_t and $t_{|G_t}$ is the restriction of t to G_t for all t's. Many other examples can be found in [27] and [28].

Global Rademacher complexities have been studied in [28] for the norms (5.2) from which generalization bounds for constrained empirical risk minimization procedure for bounded loss can be derived. In the following result we compute the global Gaussian mean width of the unit ball associated with the norm $\Omega(\cdot|\Theta)$ so that estimation and model selection results for the regularization method associated with $\Omega(\cdot|\Theta)$ may follow from Theorem 4.2.

Proposition 5.2 *Let Θ be a nonempty convex cone of $(0,\infty)^d$. Denote by $\mathrm{Ext}(\Theta \cap S_1^{d-1})$ the set of extreme points of the closure of Θ intersected with the unit sphere S_1^{d-1}. The Gaussian mean width of the unit ball $B_{\Omega(\cdot|\Theta)}$ associated with the norm $\Omega(\cdot|\Theta)$ is such that*

$$\ell^*(B_{\Omega(\cdot|\Theta)}) \leq 2 + M\sqrt{2\log\big(M|\mathrm{Ext}(\Theta \cap S_1^{d-1})|\big)}.$$

where $M = \max_{a\in\mathcal{E}} \|a\|_\infty^{1/2}$.

Proof. The argument is adapted from the one in [28]. First note that the bound is void when $\mathrm{Ext}(\Theta \cap S_1^{d-1})$ is infinite. We assume now that this set is finite and denote $\mathcal{E} = \mathrm{Ext}(\Theta \cap S_1^{d-1})$. It follows from [27] that the dual norm of $\Omega(\cdot|\Theta)$ is

$$\Omega^*(t|\Theta) = \max_{a\in\mathcal{E}} \Big(\sum_{j=1}^d a_j t_j^2\Big)^{1/2}. \tag{5.3}$$

Therefore, if $G = (g_1, \ldots, g_d)$ denotes a Standard Gaussian vector of $\mathbb{R}^d$ then $\ell^*(B_{\Omega(\cdot|\Theta)}) = \mathbb{E}\Omega^*(G|\Theta)$.

For every $a \in \mathcal{E}$, we have $\mathbb{E}\Big(\sum_{j=1}^d a_j g_j^2\Big)^{1/2} \leq \|a\|_1^{1/2} = 1$. Therefore, it follows that for every $\delta > 0$,

$$\begin{aligned}
&\mathbb{E}\Omega^*(G|\Theta) = \int_0^\infty P[\Omega^*(G|\Theta) \geq v]dv \\
&\leq 1 + \delta + \int_{1+\delta}^\infty P\Big[\max_{a\in\mathcal{E}}\Big(\sum_{j=1}^d a_j g_j^2\Big)^{1/2} \geq v\Big]dv \\
&\leq 1 + \delta + \sum_{a\in\mathcal{E}} \int_{1+\delta}^\infty P\Big[\sum_{j=1}^d a_j g_j^2 \geq v^2\Big]dv \\
&\leq 1 + \delta + \sum_{a\in\mathcal{E}} \int_{\delta}^\infty P\Big[\Big(\sum_{j=1}^d a_j g_j^2\Big)^{1/2} \geq \mathbb{E}\Big(\sum_{j=1}^d a_j g_j^2\Big)^{1/2} + v\Big]dv
\end{aligned}$$

It follows from Borell inequality (cf. Chapter 3.1 in [24]) that for every $a \in \mathcal{E}$

$$P\Big[\Big(\sum_{j=1}^{d} a_j g_j^2\Big)^{1/2} \geq \mathbb{E}\Big(\sum_{j=1}^{d} a_j g_j^2\Big)^{1/2} + v\Big] \leq \exp\big(-v^2/(2\sigma_a^2)\big)$$

where $\sigma_a = \sup_{v \in B_2^d}\big(\sum_{j=1}^d a_j v_j^2\big)^{1/2} = \|a\|_\infty^{1/2}$. Hence, for $M = \max_{a\in\mathcal{E}} \|a\|_\infty^{1/2}$, we obtain

$$\mathbb{E}\Omega^*(G|\Theta) \leq 1 + \delta + |\mathcal{E}| M \exp(-\delta^2/(2M^2))$$

and the result follows for $\delta = M\sqrt{2\log(|\mathcal{E}|M)}$. ■

In particular, when $\Theta = (0,\infty)^d$, the norm (5.2) is the ℓ_1^d-norm and according to Proposition 5.2 we obtain that the Gaussian mean width of its unit ball is of the order of $\sqrt{\log(ed)}$ and we recover the result from Section 5.1. In the case of the group Lasso, the size of the Gaussian mean width is $\sqrt{\log T}$ where T is the number of groups. Other examples of applications can be found in [28, 27].

6 Regularization methods in $\mathbb{R}^{m\times T}$

In this section, the X_i's belong to the set of $m \times T$ matrices $\mathbb{R}^{m\times T}$ endowed with the inner product $\langle A, B\rangle = \sum_{u,v} A_{uv}B_{uv}$. As in Section 5, we consider $A^* \in \operatorname{argmin}_{A\in\mathbb{R}^{m\times T}} \mathbb{E}\big(Y - \langle X, A\rangle\big)^2$ so that $\langle X, A^*\rangle$ is the best linear approximation of Y. Usually the dimension mT will be larger than the number of observations N but we believe that A^* have some low-dimensional structure characterized by some function $\|\cdot\|$ satisfying Assumption 1.1 so that $\|A^*\|$ should be small.

In this context, we may again apply Theorem 4.2 when X is L-subgaussian for some $L > 0$ and $Y - \langle X, A^*\rangle \in L_q$ for $q > 2$ and consider the regularization procedure:

$$\hat{A} \in \operatorname*{argmin}_{A\in\mathbb{R}^{m\times T}} \Big(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \langle X_i, A\rangle)^2 + 2\eta_1^3\kappa_1(u)\sigma_q \|A\| \frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}\Big) \tag{6.1}$$

where $\|\cdot\|$ is a functions on $\mathbb{R}^{m\times T}$ satisfying Assumption (4.6). It follows from Theorem 4.2 that $\hat{A}$ satisfies, with probability at least $1-2\exp(-c_0N)-5(c_1/u)^q$,

$$\big\|\langle X, \hat{A} - A^*\rangle\big\|_{L_2} \leq c_0 s(C_0\|A^*\|) \text{ and } \big\|\hat{A}\big\| \leq \eta_1 C_0 \|A^*\|$$

where $s^2(\rho) \sim \kappa_1(u)\sigma_q\rho\ell^*(B_{\|\cdot\|})/\sqrt{N}$ when $N \gtrsim \ell^*(\mathcal{E})^2$ and when $N \lesssim \ell^*(\mathcal{E})^2$,

$$s^2(\rho) \sim \max\Big(\kappa_1(u)\sigma_q\rho\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}, \rho^2\frac{\ell^*(B_{\|\cdot\|})^2}{N}\Big).$$

In the following section, we provide Gaussian mean widths associated with some classical regularization functions that have been used in fields like matrix completion and collaborative filtering.

6.1 S_p-regularization for $p > 0$

In this section, we consider the Schatten (quasi)-norms $\|\cdot\|_{S_p}$ as regularization function defined for every A in $\mathbb{R}^{m\times T}$ by

$$\|A\|_{S_p} = \Big(\sum_{j=1}^{m\wedge T}\sigma_j(A)^p\Big)^{1/p}$$

where $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{m\wedge T}(A)$ are the ordered singular values of A and $m \wedge T = \min(m, T)$.

Those norms have been extensively used for the matrix completion and collaborative filtering problems. Exact reconstruction properties of procedures based on minimizing the S_1-norm constrained to matching the data have been proved for instance in [5, 7, 6, 17, 9]. In the noisy setup, statistical properties of regularized procedures based on the S_1-norm have been obtained in [20, 38, 19, 33, 15, 18].

A result closely related to our is Theorem 9.2 from [19]. It shows that in the statistical model $Y = \big\langle X, A^*\big\rangle + \zeta$ where X is sub-gaussian, isotropic (i.e. $\mathbb{E}\big\langle X, A\big\rangle^2 = \|A\|_{S_2}^2$ for every $A \in \mathbb{R}^{m\times T}$) and ζ is in the Orlicz space ψ_α for some $\alpha \geq 1$, the regularization procedure $\hat{A}$ with regularizing function $\|\cdot\|_{S_1}$ satisfies for every $t > 0$, with probability larger than $1 - \exp(-t)$,

$$\left\|\hat{A} - A^*\right\|_{S_2}^2 \leq C\min\big(\lambda\|A^*\|_{S_1}, \lambda^2\mathrm{rank}(A^*)\big) \qquad (6.2)$$

when $N \gtrsim m\mathrm{rank}(A^*)$ and the regularization parameter is such that

$$\lambda \gtrsim \max\Big[\|\zeta\|_2\sqrt{\frac{m(t+\log m)}{N}}, \|\zeta\|_{\psi_\alpha}\log^{1/\alpha}\Big(\frac{\|\zeta\|_{\psi_\alpha}}{\|\zeta\|_{L_2}}\Big)\frac{\sqrt{m}(t+\log N)(t+\log m)}{N}\Big]$$

where $\|\zeta\|_{\psi_\alpha}$ is the ψ_α-Orlicz norm of ζ (cf. [35]).

An estimation result also follows from the next well-known result (cf. for instance Proposition 1.4.4 in [8]) for (6.1) for S_p-norm regularization, $p > 0$

without assuming a statistical model, for a noise in L_q, $q > 2$ and without assuming isotropicity of X.

Proposition 6.1 *There exists an absolute constant c_0 such that the following holds. Let $p > 0$ and denote by B_p^{mT} the unit ball of $\|\cdot\|_{S_p}$. The Gaussian mean width of B_p^{mT} satisfies*

$$\ell^*(B_p^{mT}) \le c_0 \left\{ \begin{array}{cc} \sqrt{m+T} & \text{when } p \le 1 \\ (m \wedge T)^{1-1/p}\sqrt{m+T} & \text{when } p > 1. \end{array} \right.$$

6.2 Max-norm regularization

Constrained empirical risk minimization procedures using the max-norm have been used in [42, 32] and [22]. This norm is defined by

$$\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\to\infty} \|V\|_{2\to\infty}\,.$$

Let B_{max} be the unit ball relative to that norm. We have

$$\ell^*(B_{max}) \lesssim \sqrt{(mT)(m+T)}.$$

Indeed, an application of Grothendieck's inequality (see, e.g., [32]) shows that

$$\mathrm{conv}\big(\mathcal{X}_\pm\big) \subset B_{max} \subset K_G \mathrm{conv}\big(\mathcal{X}_\pm\big)$$

where K_G is the Grothendieck constant and $\mathcal{X}_\pm = \{uv^\top : u \in \{\pm 1\}^m, v \in \{\pm 1\}^T\}$. If $\mathfrak{G} = (g_{ij})_{1\le u\le m : 1 \le v \le T}$ is a standard $m \times T$ Gaussian matrix, it follows from a Gaussian maximal inequality (cf. Chapter 3 in [24]) that

$$\begin{aligned} \ell^*(B_{max}) &= \mathbb{E} \sup_{A\in B_{max}} |\big\langle \mathfrak{G}, A\big\rangle| \le K_G \mathbb{E} \sup_{A\in \mathrm{conv}(\mathcal{X}_\pm)} |\big\langle \mathfrak{G}, A\big\rangle| \\ &= K_G \mathbb{E} \sup_{A\in\mathcal{X}_\pm} |\big\langle \mathfrak{G}, A\big\rangle| \lesssim \max_{A\in\mathcal{X}_\pm} \|A\|_F \sqrt{\log|\mathcal{X}_\pm|} \lesssim \sqrt{(mT)(m+T)}. \end{aligned}$$

6.3 Atomic-norm regularization

Atomic-norm have been studied in [9] for the exact and robust recovery problem from few Gaussian linear measurements. Minimal numbers of Gaussian measurements are obtained which insures exact and robust recovery of constrained and regularized procedures. The analysis from [9] follows from some computations of the Gaussian mean width of the intersection of the tangent cone at the target point of the unit ball associated with the atomic norm with the unit Euclidean sphere.

We recall the construction of atomic regularization functions in $\mathbb{R}^{m\times T}$. Let $\mathcal{A}\subset\mathbb{R}^{m\times T}$. The elements in $\mathcal{A}$ are called the *atoms*. Denote by $\mathrm{conv}(\mathcal{A})$ the convex hull of $\mathcal{A}$. The **gauge function** associated with $\mathrm{conv}(\mathcal{A})$ is

$$\|A\|_{\mathcal{A}}=\inf\big(t>0:A\in t\mathrm{conv}(\mathcal{A})\big). \tag{6.3}$$

Even though, $\|\cdot\|_{\mathcal{A}}$ is not a norm in general it always satisfies that: for every $A,B\in\mathbb{R}^{m\times T}$ and $\lambda\geq 0$:

$$\|A+B\|_{\mathcal{A}}\leq\|A\|_{\mathcal{A}}+\|B\|_{\mathcal{A}}\text{ and }\|\lambda A\|_{\mathcal{A}}=\lambda\|A\|_{\mathcal{A}}$$

therefore, if we further assume that $\mathrm{conv}(\mathcal{A})$ is symmetric around 0 then conditions (4.6) is satisfied and we can applied Theorem 4.2. It only remains to compute the Gaussian mean width of the unit ball associated with $\|\cdot\|_{\mathcal{A}}$ which follows from the computation of $\ell^*(\mathcal{A})$ since

$$\ell^*\big(B_{\|\cdot\|_{\mathcal{A}}}\big)=\ell^*(\mathrm{conv}(\mathcal{A}))=\ell^*(\mathcal{A}).$$

For instance, when $m=T$ and $\mathcal{A}$ is the set of all orthogonal matrices, we have $\|\cdot\|_{\mathcal{A}}=\|\cdot\|_{S_2}$. Then $\ell^*\big(B_{\|\cdot\|_{\mathcal{A}}}\big)=\mathbb{E}\|\mathfrak{G}\|_{S_2}\leq\sqrt{m}\mathbb{E}\|\mathfrak{G}\|_{S_\infty}\lesssim m$ because the spectral norm ball is the convex hull of the set of orthogonal matrices. Note that we recover the same order of the Gaussian mean width obtained in Proposition 3.13 in [9].

7 Regularization method by RKHS norm

In this section, we consider regularizing by the norm of a Reproducing Kernel Hilbert Space (RKHS). Important facts on RKHS may be found in Chapter 4 from [39] or in [10].

Recall that if $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a positive definite kernel, then by Mercer's theorem, there is an orthogonal basis $\big(\phi_i\big)_{i\in\mathbb{N}}$ of $L_2=L_2(\mu)$ (where we recall that μ is the probability distribution of X) such that $\mu\otimes\mu$-almost surely, $K(x,y)=\sum_{i=1}^\infty\lambda_i\phi_i(x)\phi_i(y)$ where $(\lambda_i)_{i\in\mathbb{N}}$ is the sequence of eigenvalues of the integral operator T_K (arranged in a non-increasing order) defined for every $f\in L_2(\mu)$ and $x\in\mathcal{X}$ by

$$(T_Kf)(x)=\int K(x,y)f(y)d\mu(y)$$

so that for all $i\in\mathbb{N}$, ϕ_i is the eigenvector corresponding to λ_i.

The reproducing kernel Hilbert space $\mathcal{H}_K$ is the set of all function series $\sum_{i=1}^{\infty} a_i K(x_i, \cdot)$ converging in L_2 endowed with the inner product

$$\Big\langle \sum a_i K(x_i, \cdot), \sum b_i K(y_i, \cdot) \Big\rangle = \sum_{i,j} a_i a_j K(x_i, x_j)$$

where a_i, b_i's are real numbers and x_i, y_i's belong to $\mathcal{X}$. The unit ball of $\mathcal{H}_K$ can be constructed from the eigenvalue decomposition of T_K by considering the feature map $\Phi : \mathcal{X} \to \ell_2$ defined by $\Phi(x) = \big(\sqrt{\lambda_i}\phi_i(x)\big)_{i \in \mathbb{N}}$ and then

$$B_{\mathcal{H}_K} = \big\{ f_\beta(\cdot) = \big\langle \beta, \Phi(\cdot) \big\rangle : \|\beta\|_{\ell_2} \leq 1 \big\}.$$

There is an isometry between the two Hilbert spaces $\mathcal{H}_K$ and ℓ_2 endowed with the norm $\|\beta\|_K = \big(\sum \beta_i^2/\lambda_i\big)^{1/2}$ whose unit ball is an ellipsoid denoted by $\mathcal{E}_K$. So that, we obtain

$$\ell^*(B_{\mathcal{H}_K}) = \ell^*(\mathcal{E}_K) \sim \Big(\sum_{j \in \mathbb{N}} \lambda_j \Big)^{1/2}$$

where the last inequality follows from Talagrand ellipsoid Theorem (cf. Chapter 2 in [40]).

We can therefore apply Theorem 4.2 to obtain estimation results for the regularization method based on the norm of a RKHS.

Note that classical procedures in RKHS are mostly developed in the classification framework. They are usually based on the hinge loss and the regularization function is the square of the RKHS norm. For such procedures, oracle inequalities have been obtained in Chapter 7 from [39] under the margin assumption (cf. [43]).

8 Proof of Theorem A

We consider the function

$$H_N(f) = P_N \mathcal{L}_f + \lambda\big(\|f\| - \|f^*\| \big) \tag{8.1}$$

where we recall that $P_N \mathcal{L}_f = P_N(\ell_f - \ell_{f^*}) = N^{-1} \sum_{i=1}^N (Y_i - f(X_i))^2 - (Y_i - f^*(X_i))^2$. It follows from the definition of $\hat{f}$ that $H_N(\hat{f}) \leq 0$. Therefore, any function $f \in F$ such that $H_N(f) > 0$ cannot be a RERM as defined in (1.3).

The scheme of the proof is as follows. We want to prove results like

$$\Big\| \hat{f} - f^* \Big\|_{L_2} \leq \Box \text{ and } \Big\| \hat{f} - f^* \Big\| \leq \triangle \tag{8.2}$$

for appropriate quantities $\square$ and $\triangle$. Our strategy is to prove that functions f such that $\|f-f^*\|_{L_2} > \square$ or $\|f-f^*\| > \triangle$ satisfy $H_N(f) > 0$ and therefore cannot be a RERM.

The proof of Theorem A is based on proving that H_N is positive on different parts of the set F so that should only remain the set of functions f for which $\|f-f^*\|_{L_2} \leq \square$ and $\left\|\hat{f}-f\right\| \leq \triangle$ on which H_N may take non-positive values.

We start by proving a result on the linear process under assumption $(\mathbf{L}(\eta_1^2 C_0 \|f^*\|))$.

Proposition 8.1 *Let $C_0 \geq 1$. Assume that ($\boldsymbol{L}(\eta_1^2 C_0 \|f^*\|)$) holds for some $s_L(\eta_1^2 C_0 \|f^*\|) \geq 0$ and $\kappa_1 \geq 0$. Then, for every $f \in F_{\eta_1 C_0 \|f^*\|}$,*

$$P_N(f^*-f)(f^*-Y) \leq \kappa_1 \max\big(s_L(\eta_1^2 C_0 \|f^*\|)\|f-f^*\|_{L_2}, s_L^2\big(\eta_1^2 C_0 \|f^*\|\big)\big).$$

Proof. We denote $\alpha_N = s_L(\eta_1^2 C_0 \|f^*\|)$. First assume that $\alpha_N > 0$. It follows from Definition 1.3 that $\phi_N(\eta_1 C_0 \|f^*\|, \alpha_N) \leq \kappa_1 \alpha_N^2$.

Let $f \in F_{\eta_1 C_0 \|f^*\|}$. When $\|f-f^*\|_{L_2} > \alpha_N$ then, by convexity of F and both properties (N1) and (N2) in Assumption 1.1,

$$\frac{\alpha_N(f-f^*)}{\|f-f^*\|_{L_2}} + f^* \in F_{\eta_1^2 C_0 \|f^*\|} \cap \big(f^* + \alpha_N \mathcal{D}\big).$$

Therefore, by definition of α_N,

$$\frac{1}{N}\sum_{i=1}^N (f^*(X_i)-Y_i)\frac{\alpha_N(f^*-f)(X_i)}{\|f^*-f\|_{L_2}} \leq \kappa_1 \alpha_N^2$$

and, since $\alpha_N > 0$,

$$\frac{1}{N}\sum_{i=1}^N (f^*(X_i)-Y_i)(f^*-f)(X_i) \leq \kappa_1 \alpha_N \|f^*-f\|_{L_2}.$$

In the other case, when $\|f-f^*\|_{L_2} \leq \alpha_N$ then $f \in F_{\eta_1^2 C_0 \|f^*\|} \cap \big(f^* + \alpha_N \mathcal{D}\big)$ therefore, by definition of α_N,

$$\frac{1}{N}\sum_{i=1}^N (f^*(X_i)-Y_i)(f^*-f)(X_i) \leq \kappa_1 \alpha_N^2.$$

In both cases, we have

$$\frac{1}{N}\sum_{i=1}^N (f^*(X_i)-Y_i)(f^*-f)(X_i) \leq \kappa_1 \max\big(\alpha_N \|f-f^*\|_{L_2}, \alpha_N^2\big).$$

Finally, when $\alpha_N = 0$ then for all $f \in F_{\eta_1 C_0 \|f^*\|}$, $P_N(f^* - f)(f^* - Y) \leq 0$ and the result holds as well. ■

We start with the case $\|f^*\| > 0$ – the case $\|f^*\| = 0$ follows an identical path and will be studied after. First, we prove that functions f in $F_{\eta_1 C_0 \|f^*\|}$ such that $\|f - f^*\|_{L_2} > c_0 s^2(\eta_1^2 C_0 \|f^*\|)$ satisfy $H_N(f) > 0$, for some well chosen c_0 and C_0. The idea is that in this part of F, the quadratic term $P_N(f^* - f)^2$ is larger than both the linear term $-2P_N(f^* - f)(f^* - Y)$ and the regularization term $-\lambda \|f^*\|$, thanks to point *i)* in Assumption 1.2 – in particular, the regularization term $\lambda \|f\|$ does not help to show that H_N is positive, somehow because f is in the "true" model $F_{\eta_1 C_0 \|f^*\|}$.

Lemma 8.2 *Assume that* ($\mathbf{L}(\eta_1^2 C_0 \|f^*\|)$) *and* ($\mathbf{Q}(\eta_1^2 C_0 \|f^*\|)$) *hold and that* $\|f^*\| > 0$. *Let* f *be in* $F_{\eta_1 C_0 \|f^*\|}$. *If* $\|f - f^*\|_{L_2} > c_0 s(\eta_1^2 C_0 \|f^*\|)$ *for* $c_0 \geq \max\big(4\kappa_1/\kappa_0, 1\big)$ *then* $H_N(f) > 0$.

Proof. Since $\|f - f^*\|_{L_2} > c_0 s(\eta_1^2 C_0 \|f^*\|)$ and $c_0 \geq 1$, according to ($\mathbf{L}(\eta_1^2 C_0 \|f^*\|)$) – together with Proposition 8.1 – and ($\mathbf{Q}(\eta_1^2 C_0 \|f^*\|)$) we have both:

1. $P_N(f - f^*)^2 \geq \kappa_0 \|f - f^*\|_{L_2}^2$
2. $P_N(f^* - f)(f^* - Y) \leq \kappa_1 s_L(\eta_1^2 C_0 \|f^*\|) \|f - f^*\|_{L_2}$.

Therefore, since $c_0 \geq 4\kappa_1/\kappa_0$ then

$$\begin{aligned} H_N(f) &= P_N(f - f^*)^2 - 2P_N(f^* - f)(f^* - Y) + \lambda\big(\|f\| - \|f^*\|\big) \\ &\geq \kappa_0 \|f - f^*\|_{L_2}^2 - 2\kappa_1 s_L(\eta_1^2 C_0 \|f^*\|) \|f - f^*\|_{L_2} - \lambda \|f^*\| \\ &\geq \frac{\kappa_0}{2} \|f - f^*\|_{L_2}^2 - \lambda \|f^*\| > \frac{\kappa_0 c_0^2}{2} s^2(\eta_1^2 C_0 \|f^*\|) - \lambda \|f^*\| \geq 0 \end{aligned}$$

where the last inequality follows from point i) in Assumption 1.2. ■

Lemma 8.2 is the first "excluding lemma": it follows from this result that inside the sub-model $F_{\eta_1 C_0 \|f^*\|}$, all functions $f \in F_{\eta_1 C_0 \|f^*\|}$ such that $\|f - f^*\|_{L_2} > c_0 s(\eta_1^2 C_0 \|f^*\|)$ have a positive $H_N(f)$ and therefore cannot be a RERM. In particular, if one proves that $\hat{f} \in F_{\eta_1 C_0 \|f^*\|}$ then, it follows from Lemma 8.2 that $\left\|\hat{f} - f^*\right\|_{L_2} \leq c_0 s(\eta_1^2 C_0 \|f^*\|)$. We are now proving that $\hat{f}$ cannot be outside of $F_{\eta_1 C_0 \|f^*\|}$.

To show that $\hat{f}$ belongs to $F_{\eta_1 C_0 \|f^*\|}$ is based again on an excluding lemma showing that all functions f outside of $F_{\eta_1 C_0 \|f^*\|}$ are such that $H_N(f) > 0$. In fact, we prove below a stronger result saying that H_N is positive outside of $K = \{f \in F : \|f - f^*\| \leq (C_0 - 1)\|f^*\|\}$ for some $C_0 > 2\eta_1 + 1$. We obtain this result by first obtaining an intermediate result on the boundary of K:

$$\partial K = \big\{ f \in F : \|f - f^*\| = (C_0 - 1)\|f^*\| \big\}.$$

Lemma 8.3 *Assume that* $(\mathbf{L}(\eta_1^2 C_0 \|f^*\|))$ *and* $(\mathbf{Q}(\eta_1^2 C_0 \|f^*\|))$ *hold and that* $\|f^*\| > 0$. *Let* $C_0 > 2\eta_1 + 1$. *For any* f *in* ∂K,

$$P_N \mathcal{L}_f + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\| > 0.$$

Proof. Let f be in ∂K. First assume that $\|f - f^*\|_{L_2} \geq c_0 s(\eta_1^2 C_0 \|f^*\|)$. Since $f \in \partial K \subset F_{\eta_1 C_0 \|f^*\|}$ and $c_0 \geq 1$, it follows from $(\mathbf{L}(\eta_1^2 C_0 \|f^*\|))$ – together with Proposition 8.1 – and $(\mathbf{Q}(\eta_1^2 C_0 \|f^*\|))$ that:

a) $P_N(f - f^*)^2 \geq \kappa_0 \|f - f^*\|_{L_2}^2$,

b) $P_N(f^* - f)(f^* - Y) \leq \kappa_1 s_L(\eta_1^2 C_0 \|f^*\|)\|f - f^*\|_{L_2}$.

Therefore,

$$\begin{aligned} P_N \mathcal{L}_f &= P_N(f - f^*)^2 - 2P_N(f^* - f)(f^* - Y) \\ &\geq \kappa_0 \|f - f^*\|_{L_2}^2 - 2\kappa_1 s_L(\eta_1^2 C_0 \|f^*\|)\|f - f^*\|_{L_2} \geq \frac{\kappa_0}{2}\|f - f^*\|_{L_2}^2 > 0 \end{aligned}$$

because $c_0 \geq 4\kappa_1/\kappa_0$ and $\|f - f^*\|_{L_2} > 0$ (because $\|f - f^*\| = (C_0 - 1)\|f^*\|$ and $\|f^*\| > 0$). So the result holds in this case because $C_0 > 2\eta_1 + 1$.

Now, assume that $\|f - f^*\|_{L_2} < c_0 s(\eta_1^2 C_0 \|f^*\|)$. It follows from point *ii)* in Assumption 1.2 that

$$\begin{aligned} 2P_N(f^* - f)(f^* - Y) &\leq 2\phi_N(\eta_1 C_0 \|f^*\|, c_0 s(\eta_1^2 C_0 \|f^*\|)) \\ &\leq \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\| \end{aligned} \tag{8.3}$$

because $\|f\| \leq \eta_1(\|f - f^*\| + \|f^*\|) \leq \eta_1 C_0 \|f^*\|$ and $\|f - f^*\| = (C_0 - 1)\|f^*\|$. Finally, if $P_N(f - f^*)^2 = 0$ then $f(X_i) = f^*(X_i)$ for every $i = 1, \ldots, N$ so $P_N(f^* - f)(f^* - Y) = 0$ then $P_N \mathcal{L}_f = 0$ and

$$P_N \mathcal{L}_f + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\| > 0$$

because $C_0 > 2\eta_1 + 1$ and $\|f - f^*\| > 0$. When $P_N(f - f^*)^2 > 0$, then, it follows from (8.3) that

$$\begin{aligned} &P_N \mathcal{L}_f + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\| \\ &> -2P_N(f^* - f)(f^* - Y) + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\| \geq 0. \end{aligned}$$

■

Now, we are in position to prove that H_N is positive outside of K.

Lemma 8.4 *Assume that ($\mathbf{L}(\eta_1^2 C_0 \|f^*\|)$) and ($\mathbf{Q}(\eta_1^2 C_0 \|f^*\|)$) hold and that $\|f^*\| > 0$. Let $C_0 > 2\eta_1 + 1$. For any $f \in F$ such that $\|f - f^*\| \geq (C_0 - 1)\|f^*\|$, $H_N(f) > 0$.*

Proof. Let $f \in F$ be such that $\|f - f^*\| \geq (C_0 - 1)\|f^*\|$. First we prove that there exists $g \in \partial K$ for which

$$H_N(f) \geq P_N \mathcal{L}_g + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|g - f^*\| \tag{8.4}$$

For any $\theta \in [0,1]$, denote $f_\theta = \theta f + (1-\theta) f^*$. We have $\|f_\theta - f^*\| = \|\theta(f - f^*)\|$. Therefore, according to (N2) in Assumption 1.1, there exists $\theta_0 \in (0,1]$ such that $\|f_{\theta_0} - f^*\| = (C_0 - 1)\|f^*\|$ – note that $\theta_0 \neq 0$ because $\|f^*\| > 0$ (and $\|0\| = 0$).

It follows from (N1) in Assumption 1.1 that

$$\|f\| - \|f^*\| \geq \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\|f - f^*\|. \tag{8.5}$$

This implies that

$$H_N(f) = P_N \mathcal{L}_f + \lambda\big(\|f\| - \|f^*\|\big) \geq P_N \mathcal{L}_f + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\|.$$

Therefore, we obtain

$$\begin{aligned} H_N(f) &\geq P_N \mathcal{L}_f + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f - f^*\| \\ &\geq P_N \mathcal{L}_f + \theta_0^{-1}\Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f_{\theta_0} - f^*\|, \end{aligned}$$

because, according to (N2) in Assumption 1.1, $\|f_{\theta_0} - f^*\| = \|\theta_0(f - f^*)\| \leq \theta_0 \|f - f^*\|$ and $C_0 > 2\eta_1 + 1$. Since $0 < \theta_0 \leq 1$, we also have

$$\begin{aligned} &P_N \mathcal{L}_f = P_N(f - f^*)^2 - 2P_N(f^* - f)(f^* - Y) \\ &= \theta_0^{-2} P_N(f_{\theta_0} - f^*)^2 - 2\theta_0^{-1} P_N(f^* - f_{\theta_0})(f^* - Y) \geq \theta_0^{-1} P_N \mathcal{L}_{f_{\theta_0}} \end{aligned}$$

Therefore,

$$H_N(f) \geq \theta_0^{-1}\Big(P_N \mathcal{L}_{f_{\theta_0}} + \Big(\frac{1}{\eta_1} - \frac{2}{C_0 - 1}\Big)\lambda \|f_{\theta_0} - f^*\|\Big)$$

and the result (8.4) holds for $g = f_{\theta_0} \in \partial K$ since $\theta_0^{-1} \geq 1$. Then the result follows from Lemma 8.3. ■

Now, we study **the case** $\|f^*\| = 0$. In this case, we have

$$H_N(f) = P_N(f - f^*)^2 - 2P_N(f^* - f)(f^* - Y) + \lambda \|f\| . \qquad (8.6)$$

The argument is merely the same as for the other case $\|f^*\| > 0$. It is based on some excluding lemmas proving that $H_N(f) > 0$ for any $f \in F$ such that $\|f - f^*\|_{L_2} > c_0 s(0)$ or $\|f - f^*\| > 0$. The main difference with the previous case is that we don't have to deal with the negative term $-\lambda \|f^*\|$ which is equal to zero and we cannot work with the sub-model $F_{\eta_1 C_0 \|f^*\|}$ which is F_0 for which the argument used in Lemma 8.4 does not work. We therefore have to work with the somehow "artificial" sub-model $F_{\eta_1 C_0 R^*}$ for $R^* > 0$ introduced in Assumption 1.2.

We start with a result inside model $F_{\eta_1 C_0 R^*}$ saying that the only functions $f \in F_{C_0 R^*}$ that may have a non-positive $H_N(f)$ are such that $\|f\| = 0$.

Lemma 8.5 *Assume that* ($\mathbf{L(}\eta_1^2 C_0 R^*\mathbf{)}$) *and* ($\mathbf{Q(}\eta_1^2 C_0 R^*\mathbf{)}$) *hold and that* $\|f^*\| = 0$*. Let* $f \in F_{\eta_1 C_0 R^*}$*. If* $\|f\| > 0$ *then* $H_N(f) > 0$*.*

Proof. It follows from ($\mathbf{L(}\eta_1^2 C_0 R^*\mathbf{)}$) – together with Proposition 8.1 – and ($\mathbf{Q(}\eta_1^2 C_0 R^*\mathbf{)}$) that for any $f \in F_{\eta_1 C_0 R^*}$,

a) $P_N(f - f^*)^2 \geq \kappa_0 \|f - f^*\|_{L_2}^2$ when $\|f - f^*\|_{L_2} \geq s_Q(\eta_1^2 C_0 R^*)$,

b) $P_N(f^* - f)(f^* - Y) \leq \kappa_1 \max\left(s_L(\eta_1^2 C_0 R^*) \|f - f^*\|_{L_2}, s_L^2(\eta_1^2 C_0 R^*)\right)$.

Let $f \in F$ be such that $\|f\| \leq \eta_1 C_0 R^*$. If $\|f - f^*\|_{L_2} > c_0 s(\eta_1^2 C_0 R^*)$ then, according to *point a)* and *b)* above, the quadratic term $P_N(f - f^*)^2$ is strictly larger than the linear term $-2P_N(f^* - f)(f^* - Y)$ and therefore, given the form of $H_N(\cdot)$ in (8.6), $H_N(f) > 0$. Now, assume that $\|f - f^*\|_{L_2} \leq c_0 s(\eta_1^2 C_0 R^*)$. It follows from point *iii)* in Assumption 1.2 that when $\|f\| > 0$

$$\begin{aligned} &2P_N(f^* - f)(f^* - Y) \leq 2\phi_N(\|f\|, c_0 s(\eta_1^2 C_0 R^*)) \\ &\leq 2\eta_1 \phi_N(\eta_1 \|f\|, c_0 s(\eta_1^2 C_0 R^*)) \leq \lambda \|f\| . \end{aligned}$$

Then by studying the cases $P_N(f - f^*)^2 = 0$ or $P_N(f - f^*)^2 > 0$ it is easy to see that $H_N(f) > 0$ when $\|f\| > 0$. ■

Now, we obtain an intermediate result for functions on the border $\{f \in F : \|f - f^*\| = C_0 R^*\}$ that will allow us to prove that all functions f such that $\|f - f^*\| \geq C_0 R^*$ have a positive $H_N(f)$ (and therefore cannot be a RERM $\hat{f}$).

Lemma 8.6 *Assume that* ($\mathbf{L}(\eta_1^2 C_0 R^*)$) *and* ($\mathbf{Q}(\eta_1^2 C_0 R^*)$) *hold and that* $\|f^*\| = 0$. *Let* $f \in F$. *If* $\|f - f^*\| = C_0 R^*$ *then*

$$P_N \mathcal{L}_f + \frac{\lambda}{\eta_1} \|f - f^*\| > 0.$$

Proof. Let $f \in F$ be such that $\|f - f^*\| = C_0 R^*$, in particular, $f \in F_{\eta_1 C_0 R^*}$ since $\|f\| \leq \eta_1 \|f - f^*\|$ so ($\mathbf{L}(\eta_1^2 C_0 R^*)$) – together with Proposition 8.1 – and ($\mathbf{Q}(\eta_1^2 C_0 R^*)$) apply. Therefore, if $\|f - f^*\|_{L_2} > c_0 s(\eta_1^2 C_0 R^*)$ then the quadratic term is strictly larger than the linear term and so $P_N \mathcal{L}_f > 0$. Then, if $\|f - f^*\|_{L_2} \leq c_0 s(\eta_1^2 C_0 R^*)$. In this case, it follows from point *iii)* in Assumption 1.2 that the linear term is such that

$$2P_N(f^* - f)(f^* - Y) \leq 2\phi_N\big(\eta_1 \|f - f^*\|, c_0 s(\eta_1^2 C_0 R^*)\big) \leq \frac{\lambda}{\eta_1} \|f - f^*\|$$

because $\|f\| \leq \eta_1 \|f - f^*\|$ and $\|f - f^*\| \leq \eta_1 C_0 R^*$. Then the result follows by studying the cases $P_N(f - f^*)^2 = 0$ or $P_N(f - f^*)^2 > 0$ and by noting that $\|f - f^*\| > 0$. ■

Lemma 8.7 *Assume that* ($\mathbf{L}(\eta_1^2 C_0 R^*)$) *and* ($\mathbf{Q}(\eta_1^2 C_0 R^*)$) *hold and that* $\|f^*\| = 0$. *Let* $f \in F$. *If* $\|f - f^*\| \geq C_0 R^*$ *then* $H_N(f) > 0$.

Proof. When $\|f - f^*\| \geq C_0 R^*$ then thanks to the same argument as the one used in the proof of Lemma 8.4, there exists $g \in F$ such that $\|g - f^*\| = C_0 R^*$, $P_N \mathcal{L}_f \geq P_N \mathcal{L}_g$ and

$$\lambda \|f\| \geq \frac{\lambda}{\eta_1} \|f - f^*\| \geq \frac{\lambda}{\eta_1} \|g - f^*\|.$$

Therefore, $H_N(f) \geq P_N \mathcal{L}_g + (\lambda/\eta_1) \|g - f^*\|$. Then it follows from Lemma 8.6 that $H_N(f) > 0$. ■

End of the proof of Theorem A: First assume that $\|f^*\| > 0$. Lemma 8.4 shows that if $\|f - f^*\| \geq (C_0 - 1)\|f^*\|$ then $H_N(f) > 0$ therefore, $\left\|\hat{f} - f^*\right\| < (C_0 - 1)\|f^*\|$. In particular, $\hat{f} \in F_{\eta_1 C_0 \|f^*\|}$, therefore, it follows from Lemma 8.2 that $\left\|\hat{f} - f^*\right\|_{L_2} \leq c_0 s(\eta_1^2 C_0 \|f^*\|)$. This proves the result of Theorem A when $\|f^*\| > 0$.

When $\|f^*\| = 0$. It follows from Lemma 8.7 that $\left\|\hat{f} - f^*\right\| \leq C_0 R^*$ hence $\left\|\hat{f}\right\| \leq \eta_1 C_0 R^*$ and so, according to Lemma 8.5, $\left\|\hat{f}\right\| = 0$. Now, we apply

($\mathbf{Q}(0)$) and ($\mathbf{L}(0)$) to show that when $\|f\| = \|f^*\| = 0$, if $\|f - f^*\|_{L_2} > c_0 s(0)$ then $H_N(f) = P_N(f - f^*)^2 - 2P_N(f^* - f)(f^* - Y) > 0$ therefore, $\left\|\hat{f} - f\right\|_{L_2} \leq c_0 s(0)$.

Finally, it follows from (N1) in Assumption 1.1 that

$$\left\|\hat{f}\right\| = \left\|\hat{f} - f^* + f^*\right\| \leq \eta_1\big(\left\|\hat{f} - f^*\right\| + \|f^*\|\big) \leq \eta_1 C_0 \|f^*\|.$$

References

[1] Peter L. Bartlett, Shahar Mendelson, and Joseph Neeman. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *To appear in Probab. Theory Related Fields*, 2011.

[2] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

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

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

[5] Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[6] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[7] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[8] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses].* Société Mathématique de France, Paris, 2012.

[9] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[10] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[11] 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.

[12] David L. Donoho and Iain M. Johnstone. Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, 99(2):277–303, 1994.

[13] D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996.

[14] Simon Foucart, Alain Pajor, Holger Rauhut, and Tino Ullrich. The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. *J. Complexity*, 26(6):629–640, 2010.

[15] Stéphane Gaïffas and Guillaume Lecué. Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Trans. Inform. Theory*, 57(10):6942–6957, 2011.

[16] Yehoram Gordon, Alexandre E. Litvak, Shahar Mendelson, and Alain Pajor. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *J. Approx. Theory*, 149(1):59–73, 2007.

[17] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011.

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

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

[20] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[21] Rafał Latała. Estimation of moments of sums of independent real random variables. *Ann. Probab.*, 25(3):1502–1513, 1997.

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

[23] Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. Technical report, CNRS, Ecole Polytechnique and Technion, 2014.

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

[25] Po-Ling Loh and Martin J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. 2013.

[26] Pascal Massart and Caroline Meynet. The Lasso as an ℓ_1-ball model selection procedure. *Electron. J. Stat.*, 5:669–687, 2011.

[27] Andreas Maurer, Charles Micchelli, and Massimiliano Pontil. A family of penalty functions for structured sparsity. *NIPS*, 2010.

[28] Andreas Maurer and Massimiliano Pontil. Structured sparsity and generalization. *J. Mach. Learn. Res.*, 13:671–690, 2012.

[29] Shahar Mendelson. Learning without concentration. Technical report, Technion, 2013. arXiv:1401.0304.

[30] Shahar Mendelson. On the geometry of subgaussian coordinate projections. Technical report, Technion, I.I.T., 2013.

[31] Shahar Mendelson and Vladimir Koltchinskii. Bounding the smallest singular value of a random matrix without concentration. Technical report, Technion and Georgia Tech, 2013. arXiv:1312.3580.

[32] Srebro Nathan and Shraibman Adi. Rank, trace-norm and max-norm. *18th Annual Conference on Learning Theory (COLT)*, 2005.

[33] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012.

[34] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[35] M. M. Rao and Z. D. Ren. *Theory of Orlicz spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1991.

[36] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

[37] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

[38] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

[39] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[40] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[41] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[42] Cai Toni and Zhou Wenxin. Matrix completion via max-norm constrained optimization. Technical report, Wharton University, 2013.

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

[44] Sara A. van de Geer. The deterministic lasso. Technical report, ETH Zürich, 2007. http://www.stat.math.ethz.ch/ geer/lasso.pdf.

[45] Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[46] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[47] Nicolas Verzelen. Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electron. J. Stat.*, 6:38–90, 2012.

[48] Zhan Wang, Sandra Paterlini, Frank Gao, and Yuhong Yang. Adaptive minimax estimation over sparse ℓ_q-hulls. Technical report, arXiv:1108.1961, 2012.

[49] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[50] Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593, 2012.