

# Regularization and the small-ball method II: complexity dependent error rates

**Guillaume Lecué**

*ENSAE*

*5 avenue Henry Le Chatelier*

*91120 Palaiseau, France*

GUILLAUME.LECUE@ENSAE.FR

**Shahar Mendelson**

*Department of Mathematics*

*Technion, Haifa, Israel*

SHAHAR@TX.TECHNION.AC.IL

**Editor:** Zhihua Zhang

## Abstract

We study estimation properties of regularized procedures of the form

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

for a convex class of functions  $F$ , regularization function  $\Psi(\cdot)$  and some well chosen regularization parameter  $\lambda$ , where the given data is an independent sample  $(X_i, Y_i)_{i=1}^N$ .

We obtain bounds on the  $L_2$  estimation error rate that depend on the complexity of the true model  $F^* := \{f \in F : \Psi(f) \leq \Psi(f^*)\}$ , where  $f^* \in \operatorname{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$  and the  $(X_i, Y_i)$ 's are independent and distributed as  $(X, Y)$ . Our estimate holds under weak stochastic assumptions – one of which being a small-ball condition satisfied by  $F$  – and for rather flexible choices of regularization functions  $\Psi(\cdot)$ . Moreover, the result holds in the learning theory framework: we do not assume any a-priori connection between the output  $Y$  and the input  $X$ .

As a proof of concept, we apply our general estimation bound to various choices of  $\Psi$ , for example, the  $\ell_p$  and  $S_p$ -norms (for  $p \geq 1$ ), weak- $\ell_p$ , atomic norms, max-norm and SLOPE. In many cases, the estimation rate almost coincides with the minimax rate in the class  $F^*$ .

**Keywords:** Empirical processes theory, high-dimensional Statistics, regularization, learning theory, minimax rates.

## 1. Introduction

In the standard learning framework, one would like to approximate / predict an unknown random variable  $Y$  using functions from a given class  $F$ , and to do so using only random data. To be more accurate, let  $(\mathcal{X}, \mu)$  be a probability space and consider a class of functions  $F$  on  $(\mathcal{X}, \mu)$ . Let  $X$  be distributed according to  $\mu$  and set  $X_1, \dots, X_N \in \mathcal{X}$  to be  $N$  independent copies of  $X$ .

Given an unknown random variable  $Y$ , let  $\mathcal{D} = (X_i, Y_i)_{i=1}^N$  be a sample selected according to the joint distribution of  $(X, Y)$ . One would like to use the data  $\mathcal{D}$  and construct a (random) function  $\hat{f}(\cdot) = \hat{f}(\mathcal{D}, \cdot) \in F$ , with  $\hat{f}(X)$  serving as a good guess of  $Y$ .

While there are various interpretations of the meaning of ‘a good guess’, the notion we will focus on here is as follows.

In a typical problem, very little is assumed on the target  $Y$  or on the measure  $\mu$ ; on the other hand, the class  $F$  is known and a typical assumption is that  $F$  is **convex and closed** in  $L_2(\mu)$ . Therefore, the functional  $f \rightarrow \mathbb{E}(f(X) - Y)^2$  has a unique minimizer in  $F$ ,

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

The notion of ‘a good guess’ is that  $\hat{f}$  is close to  $f^*$  in  $L_2(\mu)$ , and one would like to obtain a high probability bound on the  $L_2(\mu)$  distance of the form

$$\|\hat{f} - f^*\|_{L_2}^2 = \mathbb{E} \left[ (f^*(X) - \hat{f}(X))^2 | \mathcal{D} \right] \leq \alpha_N^2. \quad (1.2)$$

In this case,  $\alpha_N^2$  is called a *rate of convergence*, the *error rate* or the  $L_2(\mu)$ -*estimation rate* of the problem.

Clearly, one has to pay a price for allowing a rather general target  $Y$ . Also, to have any hope that  $f^*(X)$  is reasonably close to  $Y$ , one has to consider large classes, leading to an error  $\alpha_N^2$  that is often too large to be of any use.

A possible way of bypassing the fact that  $F$  may be very large, is the classical approach to *regularization*, where a certain property one believes  $f^*$  to possess is emphasized by penalizing functions that do not have that property. The penalty is endowed via a *regularization function*  $\Psi(\cdot)$ , defined on an appropriate subspace  $E \subset L_2(\mu)$  that contains  $F$ , and for which  $\Psi(f^*)$  is believed to be small (though one does not know that for certain). As a consequence, regularization procedures are designed to fit the data and to have a small  $\Psi$  value at the same time. One way of achieving that is to search for functions in  $F$  that realize a good tradeoff between fitting that data, which is measured via an empirical loss function  $P_N \ell_f$ , and the size of the regularization term  $\lambda \Psi(f)$ .

**Definition 1.1** *The Regularized Empirical Risk Minimization procedure (RERM) is defined by*

$$\hat{f} \in \operatorname{argmin}_{f \in F} (P_N \ell_f + \lambda \Psi(f)), \quad (1.3)$$

where here and throughout the article,  $P_N h$  denotes the empirical mean of  $h$ ,  $\ell_f$  is the loss function associated with  $f$  and  $\lambda$  is the so-called regularization parameter.

We only consider the square loss  $\ell_f(x, y) = (y - f(x))^2$ , and thus,

$$P_N \ell_f = \frac{1}{N} \sum_{i=1}^N (Y_i - f(X_i))^2.$$

A well known example to this, the ‘classical approach’ to regularization, is the cubic smoothing spline that can be obtained with the choice

$$\Psi(f) = \int (f''(t))^2 dt.$$

Another well-studied example is of the form

$$\Psi(f) = \int_{\mathbb{R}^d} \frac{|\bar{f}(t)|^2}{\bar{G}(t)} dt$$

where the integration is with respect to the Lebesgue measure,  $\bar{f}$  is the Fourier transform of  $f$  and  $\bar{G}$  is some positive function tending to zero when  $|t|$  goes to infinity (cf. [Girosi et al. \(1995\)](#)). In fact, this type of regularization methods dates back to Tikhonov ([Tikhonov \(1943\)](#)) and is sometimes called Tikhonov regularization; it is also known as  $L_2$ -regularization or Ridge regularization ([Golub et al. \(1979\)](#)).

These methods and others like them have been used to “smooth” estimators that have poor generalization capability because of their tendency to over-fit the data, and for the corresponding regularization functions, having a small  $\Psi$  value is a guarantee of smoothness. We refer to [Hastie et al. \(2009\)](#) for other examples of regularization functions that have been used to “smooth” estimators.

We said “classical approach to regularization” because in the more modern approach the aim is somewhat different. One uses a penalty that seemingly has little to do with the property one wishes to emphasize (usually, some notion of *sparsity*). Yet somehow, almost magically, the penalty enhances a hidden property and the resulting error rate does not depend on  $\Psi(f^*)$  but on that hidden property of  $f^*$ . We call such error rates **sparsity-dependent error rates**.

The first part of this article ([Lecué and Mendelson \(to appear\)](#)) has dealt with the modern approach to regularization. Here we would like to complete the picture by exploring bounds that depend on  $\Psi(f^*)$  rather than on some hidden sparsity structure of  $f^*$ . Such error rates will be called **complexity-dependent error rates**, since the aim is to obtain rates of convergence that depend on the complexity of the unknown true model  $\{f \in F : \Psi(f) \leq \Psi(f^*)\}$ . Of course, the two approaches may sometimes be combined advantageously (see some examples below).

In this context, we will consider regularization functions that satisfy the following properties, which are more general than the ones considered in [Lecué and Mendelson \(to appear\)](#).

**Assumption 1.1** *A function  $\Psi : E \rightarrow \mathbb{R}_+$  is a regularization function if*

- *It is nonnegative, even, convex and  $\Psi(0) = 0$ .*
- *There is a constant  $\eta \geq 1$ , for which, for every  $f, h \in E$ ,*

$$\Psi(f + h) \leq \eta(\Psi(f) + \Psi(h)).$$

- *For every  $0 \leq \alpha \leq 1$  and  $h \in E$ ,  $\Psi(\alpha h) \leq \alpha\Psi(h)$ .*

**Remark 1.2** *Classical Model Selection regularization functions, such as the cardinality of the support of a vector or the rank of a matrix, are usually not convex and do not satisfy Assumption 1.1. Such examples are therefore not considered in what follows.*

### 1.1 Classical vs. modern

As mentioned above, the direction we take here is closely related to the classical approach to regularization and is rather different from the modern approach. To explain the differences we shall use the celebrated LASSO estimator (cf. Tibshirani (1996); Donoho and Johnstone (1994)) as an example.

Let  $F$  be a class of linear functionals on  $\mathbb{R}^d$  of the form  $\langle t, \cdot \rangle$ . Set  $t^* \in \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2$ , and consider the RERM (1.3) with the  $\ell_1^d$ -norm,  $\|t\|_1 = \sum_{i=1}^d |t_i|$ , serving as a regularization function. Let

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + \lambda \|t\|_1 \right),$$

and the resulting minimizer is the LASSO estimator.

Estimation, de-noising, prediction and support recovery results have been obtained for the LASSO in the last decades (see, for example, Tibshirani (1996), Bickel et al. (2009), and the books Giraud (2015); Bühlmann and van de Geer (2011) and Koltchinskii (2011) for additional references).

The LASSO has been used in high-dimensional problems, in which the aim was to enhance a low-dimensional structure. The hope was that if the signal  $t^*$  were sparse (that is, supported on relatively few coordinates), the regularization procedure  $\hat{t}$  would estimate  $t^*$  with an error rate depending on the cardinality of the support of  $t^*$ , denoted by  $\|t^*\|_0 = |\{j \in \{1, \dots, d\} : t_j^* \neq 0\}|$ .

However, if  $t^*$  happens to be ‘well-spread’ rather than sparse, though with a reasonable  $\ell_1^d$  norm, the sparsity-dependent error rate is useless, while a complexity-dependent error rate, which yields bounds in terms of  $\|t^*\|_1$ , is sharper. The obvious example is  $t_1^* = (1/d, \dots, 1/d)$  and  $t_2^* = (1, 0, \dots, 0)$ : although  $\|t_1^*\|_1 = \|t_2^*\|_1 = 1$ , the cardinalities of their supports are very different, and sparsity-dependent error rates when  $t^* = t_1^*$  are likely to be bad.

Examples of that nature are the reason why error rates combining both sparsity and complexity have been obtained for the LASSO. A typical example is Corollary 9.1 in Koltchinskii (2011). To formulate it, Let  $W_1, \dots, W_N$  be  $N$  independent, centered subgaussian variables with variance  $\sigma^2$  and set  $x_1, \dots, x_N$  to be  $N$  deterministic vectors in  $\mathbb{R}^d$ . Assume that the design matrix,  $\Gamma = N^{-1/2} \sum_{i=1}^N \langle x_i, \cdot \rangle e_i$ , whose rows are  $x_i/\sqrt{N}$ , satisfies some Restricted Isometry Property (cf. Candès et al. (2006)). If  $Y_i = \langle x_i, t^* \rangle + W_i$ ,  $i = 1, \dots, N$ , then for a well chosen regularization parameter  $\lambda$ , one has, with high probability,

$$\mathbb{E} \langle X, \hat{t} - t^* \rangle^2 \leq C \min \left\{ \frac{\sigma^2 \|t^*\|_0 \log d}{n}, \sigma \|t^*\|_1 \sqrt{\frac{\log d}{n}} \right\} \quad (1.4)$$

for a suitable absolute constant  $C$ .

The error rate from (1.4) consists of two components: the sparsity-dependent error term  $\sigma^2 (\|t^*\|_0 \log d)/n$ , and the complexity-dependent error term  $\sigma \|t^*\|_1 \sqrt{(\log d)/n}$ , and in what follows we will present a few other examples that combine the two rates – because the procedure one uses to obtain both types of rate is the same.

The aim of this article is to address the “complexity-based” aspect of the problem: to study regularization problems in which one believes that the  $\Psi(f^*)$  is relatively small, and obtain an error rate that depends on  $\Psi(f^*)$  rather than on some sparsity property of  $f^*$ .

## 1.2 Attaining Minimax rates

A natural benchmark for measuring the success of a regularization method is the minimax error rate, assuming that  $\Psi(f^*)$  is known: if one is given additional information on  $\Psi(f^*)$ , e.g., that  $f^* \in \{f : \Psi(f) \leq R\}$ , one may consider the estimation problem in  $\{f : \Psi(f) \leq R\}$  using the given random data. Such a problem has an optimal error rate (called the minimax rate): it is the best rate any learning procedure may achieve in the class  $\{f : \Psi(f) \leq R\}$  given the random data  $(X_i, Y_i)_{i=1}^N$ . This minimax rate will serve as our benchmark, and will be compared with the error rates that we obtain.

Of course, one *is not* given additional information on  $\Psi(f^*)$  and it is reasonable to expect that the error rate of the regularization procedure will be significantly slower than this benchmark. The question we shall study here focuses on that gap. In fact, we will show that the price one has to pay for not knowing  $\Psi(f^*)$  is surprisingly small, under rather weak assumptions.

From a technical perspective, all regularization-based procedures share one crucial aspect: the calibration of the regularization parameter  $\lambda$ . That choice is very important as  $\lambda$  is an essential component in ensuring that the error rate of the estimator  $\hat{f}$  is well-behaved. Thus, to study the gap between the regularization error rate and the minimax rate, one has to identify the right choice of  $\lambda$ .

**Question 1.3** *What is the ‘correct choice’ of the regularization parameter  $\lambda$ , and given that choice, what is the rate of convergence of RERM? Specifically, how far is the resulting rate from the one that could have been achieved had  $\Psi(f^*)$  been given in advance?*

An answer to Question 1.3 requires one to identify  $\lambda$ ; to find a high probability upper bound on  $\|\hat{f} - f^*\|_{L_2(\mu)}^2$  where  $\hat{f}$  is defined in Definition 1.1 for that choice of  $\lambda$ ; and then to compare the error rate to the minimax rate of the estimation problem in the class  $\{f \in F : \Psi(f) \leq \Psi(f^*)\}$  (the so-called true model).

The strategy we use below follows a similar path to [Lecué and Mendelson \(to appear\)](#) and is based on the small ball method, introduced in [Mendelson \(2015, 2014\)](#); [Koltchinskii and Mendelson \(2015\)](#); [Mendelson \(To appear\)](#).

## 1.3 The small-ball method

Given a closed and convex class  $F$  and an unknown target  $Y$ , recall that  $f^* \in F$  is a minimizer in  $F$  of the functional  $f \rightarrow \mathbb{E}(f(X) - Y)^2$ .

The excess loss functional associated with  $f \in L_2(\mu)$  is

$$\begin{aligned} f \rightarrow \mathcal{L}_f(X, Y) &= \ell_f(X, Y) - \ell_{f^*}(X, Y) = (f(X) - Y)^2 - (f^*(X) - Y)^2 \\ &= (f - f^*)^2(X) + 2(f^*(X) - Y)(f - f^*)(X). \end{aligned} \tag{1.5}$$

Moreover, since  $F$  is closed and convex, then by the characterization of the nearest point map in a Hilbert space,

$$\mathbb{E}(f^*(X) - Y)(f - f^*)(X) \geq 0 \text{ for every } f \in F;$$

thus

$$\frac{1}{N} \sum_{i=1}^N (f^*(X_i) - Y_i)(f - f^*)(X_i) \geq \frac{1}{N} \sum_{i=1}^N (f^*(X_i) - Y_i)(f - f^*)(X_i) - \mathbb{E}(f^*(X) - Y)(f - f^*)(X). \quad (1.6)$$

Let  $E$  be a subspace that contains  $F$  and set  $\Psi(\cdot)$  to be a regularization function on  $E$  (i.e., a functional that satisfies Assumption 1.1). Set  $\rho \geq 0$  and put

$$K_\rho(f^*) = \{h \in E : \Psi(h - f^*) \leq \rho\},$$

which, by the convexity of  $\Psi$ , is a convex set.

**Definition 1.4** For every  $\lambda > 0$  and any  $f \in L_2(\mu)$ , define the regularized excess loss by

$$\mathcal{L}_f^\lambda = (\ell_f + \lambda\Psi(f)) - (\ell_{f^*} + \lambda\Psi(f^*)) = \mathcal{L}_f + \lambda(\Psi(f) - \Psi(f^*)).$$

Note that for every sample  $(X_i, Y_i)_{i=1}^N$ , a minimizer  $\hat{f}$  of the empirical regularized loss functional (1.3) also minimizes in  $F$  the empirical regularized excess loss  $f \rightarrow P_N \mathcal{L}_f^\lambda$ . Hence, since  $\mathcal{L}_{f^*}^\lambda = 0$ , it follows that for every  $(X_i, Y_i)_{i=1}^N$ , the empirical regularized excess loss in  $\hat{f}$  is non-positive:

$$P_N \mathcal{L}_{\hat{f}}^\lambda \leq 0. \quad (1.7)$$

This observation is at the heart of our analysis, as it allows one to exclude functions  $f$  in  $F$  that satisfy  $P_N \mathcal{L}_f^\lambda > 0$  as potential minimizers of the empirical regularized loss function. Our strategy is therefore to show that if  $f \in F$  and  $\|f - f^*\|_{L_2(\mu)}$  is not ‘too small’, then necessarily  $P_N \mathcal{L}_f^\lambda > 0$  (for the right choice of  $\lambda$ ); hence, functions cannot be minimizers of the empirical regularized (excess) loss function.

To simplify notation, set  $\xi = Y - f^*(X)$ ,

$$\mathcal{M}_{f-f^*}(X, Y) = \xi(f - f^*)(X) - \mathbb{E}\xi(f - f^*)(X) \text{ and } \mathcal{Q}_{f-f^*}(X) = (f - f^*)^2(X);$$

therefore, combining (1.5) and (1.6),

$$P_N \mathcal{L}_f \geq P_N \mathcal{Q}_{f-f^*} - 2|P_N \mathcal{M}_{f-f^*}|. \quad (1.8)$$

The main step in the small-ball method is to find a lower bound on the quadratic process  $f \rightarrow P_N \mathcal{Q}_{f-f^*}$  and an upper bound on  $f \rightarrow |P_N \mathcal{M}_{f-f^*}|$ . The two estimates should hold with high probability on certain subsets of  $F$ . Then, they have to be compared with the behaviour of the regularization term  $\lambda(\Psi(f) - \Psi(f^*))$  on those sets to ensure that  $P_N \mathcal{L}_f^\lambda > 0$ .

A uniform lower bound on the quadratic component  $P_N \mathcal{Q}_{f-f^*}$  can be obtained under the following small-ball condition:

**Assumption 1.2** Assume that there are constants  $\kappa > 0$  and  $0 < \varepsilon \leq 1$ , for which, for every  $f, h \in F$ ,

$$Pr\left(|f(X) - h(X)| \geq \kappa \|f - h\|_{L_2(\mu)}\right) \geq \varepsilon.$$

There are numerous examples in which Assumption 1.2 may be verified for  $\kappa$  and  $\varepsilon$  that are absolute constants and we refer the reader to Mendelson (2014, 2015); Lecué and Mendelson (2017); Mendelson (To appear); Koltchinskii and Mendelson (2015); Rudelson and Vershynin (2015) for some of them.

To put Assumption 1.2 in some perspective, recall that the class  $F = \{f_t = \langle \cdot, t \rangle : t \in \mathbb{R}^d\}$  is *identifiable* if for every  $t_1, t_2 \in \mathbb{R}^d$ ,  $Pr(f_{t_1} \neq f_{t_2}) > 0$ , (where the probability is taken with respect to the underlying measure  $\mu$ ). By linearity, this condition is equivalent to assuming that for every  $t \in \mathbb{R}^d$ ,  $Pr(|\langle X, t \rangle| > 0) > 0$ . Thus, the small-ball condition (which in this case implies that for every  $t \in \mathbb{R}^d$ ,  $Pr(|\langle X, t \rangle| \geq \kappa \|\langle X, t \rangle\|_{L_2}) \geq \varepsilon$ ) is simply a uniform estimate on the degree of identifiability of class  $F$  and is therefore a rather weak assumption from a statistical point of view.

Now, let us introduce two complexity parameters that play a central role in our analysis. Let  $D$  be the unit ball in  $L_2(\mu)$  and for  $r > 0$  set

$$rD_{f^*} = \{f \in L_2(\mu) : \|f - f^*\|_{L_2(\mu)} \leq r\} = f^* + rD.$$

**Definition 1.5** Given a class  $F$  of functions and  $\tau > 0$ , let

$$r_Q(F, \tau) = r_Q(F, f^*, \tau) = \inf \left\{ r > 0 : \mathbb{E} \sup_{f \in F \cap rD_{f^*}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right| \leq \tau r \right\},$$

where  $(\varepsilon_i)_{i=1}^N$  are independent, symmetric,  $\{-1, 1\}$ -valued random variables that are also independent of  $(X_i, Y_i)_{i=1}^N$ .

Set

$$\phi_N(F, f^*, s) = \sup_{f \in F \cap sD_{f^*}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i (f - f^*)(X_i) \right| \quad (1.9)$$

and put

$$r_M(F, \tau, \delta) = r_M(F, f^*, \tau, \delta) = \inf \left\{ s > 0 : Pr\left(\phi_N(F, f^*, s) \leq \tau s^2 \sqrt{N}\right) \geq 1 - \delta \right\}.$$

One may show the following (see Theorem 3.1 in Mendelson (2015)):

**Theorem 1.6** Let  $F$  be a closed, convex class of functions that satisfies Assumption 1.2 with constants  $\kappa$  and  $\varepsilon$ , and set  $\theta = \kappa^2 \varepsilon / 16$ . For every  $\delta \in (0, 1)$ , with probability at least  $1 - \delta - 2 \exp(-N\varepsilon^2/2)$  one has both:

- for every  $f \in F$ ,

$$|P_N \mathcal{M}_{f-f^*}| \leq \frac{\theta}{4} \max \left\{ \|f - f^*\|_{L_2(\mu)}^2, r_M^2(F, \theta/5, \delta/4) \right\},$$

- for every  $f \in F$  with  $\|f - f^*\|_{L_2(\mu)} \geq r_Q(F, \kappa\varepsilon/32)$ ,

$$P_N \mathcal{Q}_{f-f^*} \geq \theta \|f - f^*\|_{L_2(\mu)}^2.$$

In particular, with probability at least  $1 - \delta - 2 \exp(-N\varepsilon^2/2)$ ,  $P_N \mathcal{L}_f \geq \frac{\theta}{2} \|f - f^*\|_{L_2(\mu)}^2$  for every  $f \in F$  that satisfies

$$\|f - f^*\|_{L_2(\mu)} \geq \max\{r_M(F, \theta/5, \delta/4), r_Q(F, \kappa\varepsilon/32)\}.$$

**Remark 1.7** An immediate outcome of Theorem 1.6 is that with high probability, a minimizer in  $F$  of the empirical excess-loss functional  $P_N \mathcal{L}_f$  must satisfy

$$\|\tilde{f} - f^*\|_{L_2(\mu)} \leq \max\{r_M(F, \theta/5, \delta/4), r_Q(F, \kappa\varepsilon/32)\}. \quad (1.10)$$

In fact, results from Lecué and Mendelson show that (1.10) is optimal in the minimax sense under additional mild technical assumptions on  $F$  when the data is assumed to satisfy the Gaussian regression model, that is, when the targets are of the form  $Y = f_0(X) + W$  for  $f_0 \in F$  and  $W$  that is a centered Gaussian random variable, independent of  $X$ . *Empirical risk minimization* performed in the set

$$F^* = \{f \in F : \Psi(f) \leq \Psi(f^*)\}$$

yields

$$\|\tilde{f} - f^*\|_{L_2} \leq \max\{r_M(F^*, \theta/5, \delta/4), r_Q(F^*, \kappa\varepsilon/32)\}, \quad (1.11)$$

and the r.h.s. of (1.11) is the minimax rate of the estimation problem in  $F^*$  (up to some mild technical assumptions). This will serve as a benchmark for the performance of the regularization procedure (1.3).

#### 1.4 The Main result

Let  $F \cap K_\rho(f^*) = \{f \in F : \Psi(f - f^*) \leq \rho\}$  and observe that these are convex subsets of  $F$ . To simplify notation, set

$$r_M(\rho) = r_M\left(F \cap K_\rho(f^*), \frac{\kappa^2\varepsilon}{80}, \frac{\delta}{4}\right) \quad \text{and} \quad r_Q(\rho) = r_Q\left(F \cap K_\rho(f^*), \frac{\kappa\varepsilon}{32}\right). \quad (1.12)$$

Let  $r(\cdot)$  be a function that satisfies that for every  $\rho \geq 0$ , one has

$$r(\rho) \geq \max\{r_M(\rho), r_Q(\rho)\}. \quad (1.13)$$

It should be noted that  $r(\rho)$  may depend on  $f^*$ , and that it also depends on other parameters – like  $\delta$ ,  $\kappa$  and  $\varepsilon$ . We will not specify the dependence on those parameters, but rather, only on the radius  $\rho$ .

When sharp estimates on  $r_M(\rho)$  and  $r_Q(\rho)$  can be obtained, one may set  $r(\rho)$  to be of the order of  $\max\{r_M(\rho), r_Q(\rho)\}$  for the class  $F \cap K_\rho(f^*)$ . This choice leads to the rate appearing in Theorem 1.6 and in (1.10). As we mentioned previously, the rate  $\max\{r_M(\rho), r_Q(\rho)\}$  is optimal in the minimax sense under additional mild technical assumptions on  $F \cap K_\rho(f^*)$ .



More information on the  $r(\rho)$  may be found in Section 1 of [Lecué and Mendelson](#) or in [Mendelson \(2015\)](#).

AS it happens, the geometry of the sets  $F \cap K_\rho(f^*)$  (see Figure 1) determines both the error rate and the regularization parameter  $\lambda$ , while  $r(\rho)$  measures the sets' 'sizes'.

The choice of  $\lambda$  is made as follows:

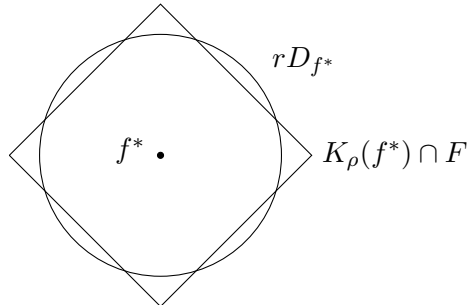


Figure 1: Localization of the set  $F \cap K_\rho(f^*)$ , i.e. its intersection with  $L_2(\mu)$ -balls of various radii  $r$  for the right choice of radius  $\rho$ , plays a central role in the analysis of the quadratic and multiplier processes.

Let

$$\mathcal{O}(\rho) = \sup \left( |P_N \mathcal{M}_{f-f^*}| : f \in F \cap K_\rho(f^*) \cap r(\rho)D_{f^*} \right)$$

and for  $\tau > 0$  and  $0 < \delta < 1$ , set

$$\gamma_{\mathcal{O}}(\rho, \tau, \delta) = \inf \{x > 0 : Pr(\mathcal{O}(\rho) \leq \tau x) \geq 1 - \delta\}$$

and

$$\gamma_{\mathcal{O}}(\rho) = \gamma_{\mathcal{O}}(\rho, 3/(80\eta^3), \delta).$$

In other words,  $\gamma_{\mathcal{O}}(\rho)$  is proportional to the smallest possible upper estimate on  $\mathcal{O}(\rho)$  that still holds with probability at least  $1 - \delta$ .

**Definition 1.8** For any  $\tau > 0$  and  $0 < \delta < 1$ , set

$$\lambda_0(\delta, \tau) = \sup_{\rho > 0, f^* \in F} \frac{\gamma_{\mathcal{O}}(\rho, \tau, \delta)}{\rho}.$$

To compare  $\lambda_0(\delta, \tau)$  with  $r_Q$  and  $r_M$ , first note that  $r_M(\rho)$  and  $\mathcal{O}(\rho)$  both depend on properties of the multiplier processes indexed by localizations of  $F \cap K_\rho(f^*)$ , and recall that symmetrized and centered processes are essentially equivalent. Second, if  $r(\rho) = r_M(\rho)$  then  $\gamma_{\mathcal{O}}(\rho) \sim r_M^2(\rho)$ ; moreover,  $\gamma_{\mathcal{O}}(\rho)$  is trivially bounded by  $\sim r^2(\rho)$  for the right choice of  $\tau$  and  $\delta$ . However, if  $r_M(\rho) \leq r_Q(\rho)$ , that is, when  $r(\rho) = r_Q(\rho)$  – which is the case when  $\rho$  is very large – one may find that  $\gamma_{\mathcal{O}}(\rho)$  is actually significantly smaller than  $\sim r^2(\rho)$ . This observation is of crucial importance because of the choice of the regularization parameter: for the right choice of  $\tau$ ,  $\gamma_{\mathcal{O}}(\rho) \leq r^2(\rho)$  and

$$\lambda_0(\delta, \tau) \leq \sup_{\rho > 0, f^* \in F} r^2(\rho)/\rho;$$

thus, one may be tempted to select the latter as a regularization term. However, there are natural examples in which  $\sup_{\rho>0} r^2(\rho)/\rho = \infty$ , rendering that choice impossible, whereas  $\sup_{\rho>0, f^* \in F} \gamma_{\mathcal{O}}(\rho)/\rho$  turns out to be finite. Of course, there are still cases in which  $\sup_{\rho>0, f^* \in F} r^2(\rho)/\rho$  is finite, and  $\lambda_0(\delta, \tau)$  is of the same order as  $\sup_{\rho>0, f^* \in F} r_M^2(\rho)/\rho$ , though that is not the generic situation.

We now come to the main result of the article.

**Theorem 1.9** *Let  $F$  be a closed, convex class of functions that satisfies Assumption 1.2 with constants  $\kappa$  and  $\varepsilon$ . Set  $\Psi(\cdot)$  to be a regularization function that satisfies Assumption 1.1 with constant  $\eta$ . Furthermore, assume that  $\lim_{\rho \rightarrow 0} r(\rho) = 0$  and put  $\lambda > \lambda_0(\delta, 3/(80\eta^3))$ .*

*If  $\hat{f}$  is the RERM with a regularization parameter  $\lambda$  as in (1.3), then with probability at least  $1 - 2\delta - 2\exp(-N\varepsilon^2/2)$ ,*

$$\|\hat{f} - f^*\|_{L_2(\mu)}^2 \leq \max\left\{r^2(10\eta\Psi(f^*)), \left(\frac{32}{\kappa^2\varepsilon}\right)\lambda\Psi(f^*)\right\}. \quad (1.14)$$

Observe that  $\lambda_0$  depends only on the oscillations of the multiplier process. Hence, if the problem is noise-free then  $\lambda_0 = 0$ , showing that any regularization parameter  $\lambda > 0$  would do. Moreover, in that case  $r_M(\rho) = 0$  and so one can choose  $r(\rho) \geq r_Q(\rho)$  obtaining an error rate that depends only on  $r_Q^2(10\eta\Psi(f^*))$ .

As noted previously, if one considers empirical risk minimization performed in  $F^* = \{f \in F : \Psi(f) \leq \Psi(f^*)\}$ , the resulting error rate is  $\|\hat{f} - f^*\|_{L_2(\mu)}^2 \leq c_0 r^2(c\Psi(f^*))$  for a suitable absolute constant  $c_0$  and a constant  $c$  that depends on  $\kappa$ ,  $\varepsilon$  and  $\delta$ ; moreover, under some minor additional assumptions, that rate is optimal in the minimax sense (cf. [Lecué and Mendelson](#)) when one takes  $r(\rho) \sim \max\{r_M(\rho), r_Q(\rho)\}$ . Hence, up to constants involved, the first term in Theorem 1.9 is essentially the minimax rate that one can obtain if  $\Psi(f^*)$  were known.

If one chooses  $\lambda \sim \lambda_0(\delta, \tau)$  for  $\tau = 3/(80\eta^3)$  then the second term in (1.14) is of the order of

$$\lambda\Psi(f^*) = \left(\sup_{\rho, f^*} \frac{\gamma_{\mathcal{O}}(\rho, \tau, \delta)}{\rho}\right) \cdot \Psi(f^*).$$

Note that for  $\rho$  that is of the order of  $\Psi(f^*)$ , one has

$$\frac{\gamma_{\mathcal{O}}(\rho, \tau, \delta)}{\rho} \cdot \Psi(f^*) \leq c_1 \gamma_{\mathcal{O}}(\rho, \tau, \delta) \leq c_2 r^2(c_3 \Psi(f^*)),$$

which coincides with the first term, up to the constants involved. Thus, the price that one has to pay for not knowing  $\Psi(f^*)$  is manifested in the need to take the supremum over all possible choices of  $\rho$  in the second term, rather than considering only the level  $\rho \sim \Psi(f^*)$ .

Thankfully, there are many natural cases in which that price is rather small, allowing for satisfactory outcomes of Theorem 1.9 that are close to the minimax rate.

## 2. Proof of Theorem 1.9

Before we present the proof, let us introduce some notation. Throughout, absolute constants or constants that depend on other parameters are denoted by  $c, C, c_1, c_2$ , etc., (and,

of course, we will specify when a constant is absolute and when it depends on other parameters). The values of these constants may change from line to line. The notation  $x \sim y$  (resp.  $x \lesssim y$ ) means that there exist absolute constants  $0 < c < C$  for which  $cy \leq x \leq Cy$  (resp.  $x \leq Cy$ ). If  $b > 0$  is a parameter, then  $x \lesssim_b y$  means that  $x \leq C(b)y$  for some constant  $C(b)$  that depends only on  $b$ .

The normed space  $\ell_p^d$  is  $\mathbb{R}^d$  endowed with the norm  $\|x\|_p = (\sum_j |x_j|^p)^{1/p}$ ; the corresponding unit ball is denoted by  $B_p^d$  and the unit Euclidean sphere in  $\mathbb{R}^d$  is  $S^{d-1}$ .

Finally, from here on we write  $Pr$  and  $\|\cdot\|_{L_2}$  without specifying the underlying measure.

The proof of Theorem 1.9 follows an almost identical path as the proof of Theorem 3.2 from [Lecué and Mendelson \(to appear\)](#). The differences between the two arguments are minor and their source is the fact that unlike [Lecué and Mendelson \(to appear\)](#), here we do not assume that  $\Psi$  is a norm. In Remark 2.2 we will outline how a version of Theorem 1.9 may be derived directly from Theorem 3.2 in [Lecué and Mendelson \(to appear\)](#) when  $\Psi$  is a norm.

Theorem 1.9 is an immediate outcome of the following lemma:

**Lemma 2.1** *Let  $\lambda_0 = \lambda_0(\delta, 3/(80\eta^3))$  and set  $\lambda > \lambda_0$ . If  $\lim_{\rho \rightarrow 0} r(\rho) = 0$ ,  $\rho \geq 5\eta\Psi(f^*)$  and  $\rho > 0$ , then with probability at least  $1 - 2\delta - 2\exp(-N\varepsilon^2/2)$ ,*

$$\|\hat{f} - f^*\|_{L_2}^2 \leq \max\{r^2(\rho), (32/(\kappa^2\varepsilon))\lambda\Psi(f^*)\}.$$

To see how Lemma 2.1 can be used to conclude the proof of Theorem 1.9, observe that if  $\Psi(f^*) > 0$ , one may simply select  $\rho = 5\eta\Psi(f^*)$  in the lemma. If, on the other hand,  $\Psi(f^*) = 0$ , let  $(\gamma_n)_{n=1}^\infty$  be a positive sequence decreasing to 0 and set  $\mathcal{A}_n = \{\|\hat{f} - f^*\|_{L_2} \leq \gamma_n\}$ , which is a decreasing sequence of events. If  $Pr(\mathcal{A}_n) \geq 1 - \nu$  for some  $0 \leq \nu \leq 1$  and every  $n$  then  $Pr(\{\hat{f} = f^*\}) \geq 1 - \nu$ . Since  $\lim_{\rho \rightarrow 0} r(\rho) = 0$ , one may apply Lemma 2.1 to each member of a nonnegative sequence  $\rho_n$  that decreases to zero and for which  $\gamma_n = r(\rho_n)$  decreases to zero. By Lemma 2.1,  $Pr(\mathcal{A}_n) \geq 1 - 2\delta - 2\exp(-N\varepsilon^2/2)$  for every  $n$  and the proof of Theorem 1.9 follows.

**Proof of Lemma 2.1.** Fix  $f^*$  and set  $\rho > 0$  that satisfies  $\rho \geq 5\eta\Psi(f^*)$ . Let

$$F_1 = \{f \in F : \Psi(f - f^*) \leq \rho\} = F \cap K_\rho(f^*),$$

and

$$F_2 = \{f \in F : \Psi(f - f^*) = \rho\}.$$

Clearly,  $F_1$  is a convex set that contains  $f^*$ , and by the continuity of the real-valued function  $t \rightarrow \Psi(f^* + t(f - f^*))$ , every ray  $[f^*, f)$  that originates in  $f^*$  and passes through some  $f \in F \setminus F_1$  intersects  $F_2$ .

Let  $\theta = \kappa^2\varepsilon/16$  and set

$$r_Q(\rho) = r_Q(F_1, \kappa\varepsilon/32) \quad \text{and} \quad r_M(\rho) = r_M(F_1, \theta/5, \delta/4).$$

There is an event  $\mathcal{A}_0$  of probability at least  $1 - \delta - 2\exp(-N\varepsilon^2/2)$ , and for every  $(X_i, Y_i)_{i=1}^N \in \mathcal{A}_0$  the following holds:

- If  $f \in F_1$  and  $\|f - f^*\|_{L_2} \geq r_Q(\rho)$  then

$$\frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i) \geq \theta \|f - f^*\|_{L_2}^2.$$

- If  $f \in F_1$  then

$$\left| \frac{1}{N} \sum_{i=1}^N \xi_i(f - f^*)(X_i) - \mathbb{E}\xi(f - f^*)(X) \right| \leq \frac{\theta}{4} \max\{\|f - f^*\|_{L_2}^2, r_M^2(\rho)\}.$$

In particular, if  $f \in F_1$  and  $\|f - f^*\|_{L_2} \geq r(\rho) \geq \max\{r_M(\rho), r_Q(\rho)\}$  then

$$P_N \mathcal{L}_f \geq \frac{\theta}{2} \|f - f^*\|_{L_2}^2.$$

By the choice of  $\lambda$ , there is an event  $\mathcal{A}_1$  of probability at least  $1 - \delta$  on which if  $f \in F_1$  and  $\|f - f^*\|_{L_2} \leq r(\rho)$ , then

$$\left| \frac{2}{N} \sum_{i=1}^N \xi_i(f - f^*)(X_i) - \mathbb{E}\xi(f - f^*)(X) \right| < \frac{3}{80\eta^3} \lambda \rho < \frac{3}{5\eta} \lambda \rho. \quad (2.1)$$

Set  $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1$  and let  $(X_i, Y_i)_{i=1}^N \in \mathcal{A}$ . The proof now follows in three steps:

- (1) Show that the functional  $f \rightarrow P_N \mathcal{L}_f^\lambda$  is bounded from below – away from zero – in  $F_2$ .
- (2) An outcome of (1) is that if  $f \in F \setminus F_1$ ,  $P_N \mathcal{L}_f^\lambda > 0$ ; hence,  $\hat{f} \notin F \setminus F_1$ .
- (3) Finally, pin-point  $\hat{f}$  within  $F_1 = \{f \in F : \Psi(f - f^*) \leq \rho\}$ .

**Step 1.** Fix  $f \in F_2$  and note that by the ‘triangle inequality’ satisfied by  $\Psi$ ,

$$\Psi(f) \geq \eta^{-1} \Psi(f - f^*) - \Psi(f^*).$$

Recall that  $\eta^{-1} \Psi(f - f^*) \geq \eta^{-1} \rho \geq 5\Psi(f^*)$  and thus,  $\Psi(f) - \Psi(f^*) \geq (3/5)\eta^{-1} \rho$ . Hence, if  $\|f - f^*\|_{L_2} \geq r(\rho)$  then

$$P_N \mathcal{L}_f^\lambda \geq (\theta/2) \|f - f^*\|_{L_2}^2 + \lambda \rho \cdot \frac{3}{5\eta} > 0.$$

On the other hand, by the choice of  $\lambda$ , if  $\|f - f^*\|_{L_2} \leq r(\rho)$  then

$$\begin{aligned} P_N \mathcal{L}_f^\lambda &\geq - \left| \frac{1}{N} \sum_{i=1}^N \xi_i(f - f^*)(X_i) - \mathbb{E}\xi(f - f^*)(X) \right| + \lambda (\Psi(f) - \Psi(f^*)) \\ &\geq - \left| \frac{1}{N} \sum_{i=1}^N \xi_i(f - f^*)(X_i) - \mathbb{E}\xi(f - f^*)(X) \right| + \lambda \rho \cdot \frac{3}{5\eta} > 0. \end{aligned}$$

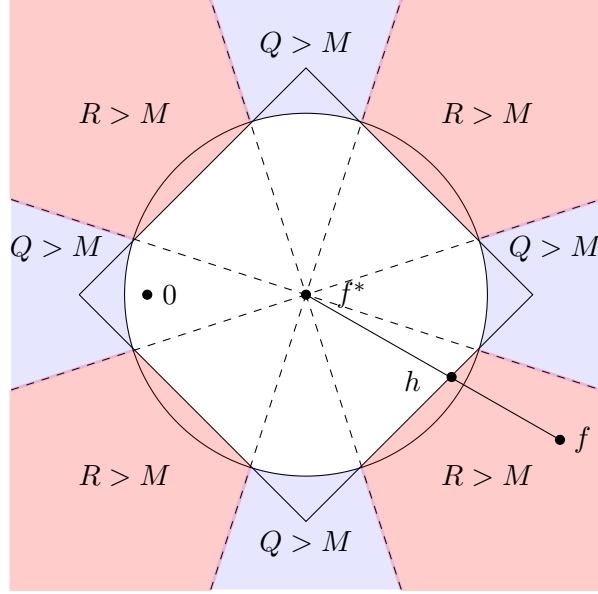


Figure 2:  $P_N \mathcal{L}_f^\lambda > 0$  for two different reasons: either  $Q > M$  – the quadratic component dominates the multiplier component, or  $R > M$  – the regularization component dominates the multiplier component. Unlike Theorem 3.2 in [Lecué and Mendelson \(to appear\)](#), here we choose  $\rho \sim \Psi(f^*)$  to ensure that  $0 \in F \cap K_\rho(f^*)$ .

It should be noted that the same proof shows that on the event  $\mathcal{A}$ , for every  $f \in F_2$ ,

$$P_N \mathcal{L}_f + \frac{\lambda}{2\eta^2} (\Psi(f) - \Psi(f^*)) > 0, \quad (2.2)$$

a fact that will be used below. Indeed,  $(\lambda/2\eta^2) \cdot (\Psi(f) - \Psi(f^*)) \geq (\lambda/2\eta^2) \cdot (3\rho/5\eta)$  and by (2.1), if  $\|f - f^*\|_{L_2} \leq r(\rho)$  then  $P_N \mathcal{L}_f \geq -(3/80) \cdot (\lambda\rho/\eta^3)$ .

**Step 2.** Let  $f \in F \setminus F_1$  and note that by the convexity of  $F$  and the continuity of  $\Psi$  on rays, there is some  $h \in F_2$  and  $R > 1$  for which  $f = f^* + R(h - f^*)$ . Thus,

$$P_N \mathcal{L}_f^\lambda = \frac{R^2}{N} \sum_{i=1}^N (h - f^*)^2(X_i) + \frac{2R}{N} \sum_{i=1}^N \xi_i(h - f^*)(X_i) + \lambda (\Psi(f) - \Psi(f^*)).$$

Observe that

$$\Psi(f) - \Psi(f^*) \geq \frac{R}{2\eta^2} (\Psi(h) - \Psi(f^*)); \quad (2.3)$$

indeed,

$$\Psi(f^* + R(h - f^*)) \geq \eta^{-1} \Psi(R(h - f^*)) - \Psi(f^*) \geq R\eta^{-1} \Psi(h - f^*) - \Psi(f^*),$$

and thus it suffices to show that

$$\frac{R}{\eta} \Psi(h - f^*) \geq \frac{R}{2\eta^2} \Psi(h) + 2\Psi(f^*).$$

But since  $\Psi(h - f^*) \geq 5\eta\Psi(f^*)$ ,  $\eta \geq 1$  and  $R \geq 1$ , one has

$$\begin{aligned} \frac{R}{\eta}\Psi(h - f^*) &\geq \frac{R}{2\eta}\Psi(h - f^*) + \frac{R}{2}\Psi(f^*) + 2R\Psi(f^*) \\ &\geq \frac{R}{2\eta}(\Psi(h - f^*) + \Psi(f^*)) + 2\Psi(f^*) \geq \frac{R}{2\eta^2}\Psi(h) + 2\Psi(f^*), \end{aligned}$$

and (2.3) follows.

Finally, applying (2.2) to  $h \in F_2$ ,

$$\begin{aligned} P_N\mathcal{L}_f^\lambda &\geq \frac{R^2}{N} \sum_{i=1}^N (h - f^*)^2(X_i) + \frac{2R}{N} \sum_{i=1}^N \xi_i (h - f^*)(X_i) + \lambda \frac{R}{2\eta^2} (\Psi(h) - \Psi(f^*)) \\ &\geq R \left( P_N\mathcal{L}_h + \frac{\lambda}{2\eta^2} (\Psi(h) - \Psi(f^*)) \right) > 0, \end{aligned}$$

and  $\hat{f} \notin F \setminus F_1$ .

**Step 3.** Turning to  $F_1 = \{f \in F : \Psi(f - f^*) \leq \rho\} = F \cap K_\rho(f^*)$ , recall that if  $f \in F_1$  and  $\|f - f^*\|_{L_2} \geq r(\rho)$ , then  $P_N\mathcal{L}_f \geq (\theta/2)\|f - f^*\|_{L_2}^2$ ; hence, if  $f$  is a potential minimizer and  $\|f - f^*\|_{L_2} \geq r(\rho)$  then

$$0 \geq P_N\mathcal{L}_f^\lambda \geq (\theta/2)\|f - f^*\|_{L_2}^2 + \lambda(\Psi(f) - \Psi(f^*)) \geq (\theta/2)\|f - f^*\|_{L_2}^2 - \lambda\Psi(f^*),$$

and

$$\|\hat{f} - f^*\|_{L_2}^2 \leq \frac{2\lambda}{\theta}\Psi(f^*),$$

as claimed. ■

**Remark 2.2** *If  $\Psi$  happens to be a norm (which is an assumption slightly stronger than Assumption 1.1), one may apply Theorem 3.2 from [Lecué and Mendelson \(to appear\)](#) directly. Indeed, and using the notation from [Lecué and Mendelson \(to appear\)](#) if  $\rho \gtrsim \Psi(f^*)$  then the set  $K = \{f : \Psi(f - f^*) \leq \rho/20\}$  contains a  $\Psi$ -ball around 0, and  $\Gamma_{f^*}(\rho)$  – the collection of norming functionals (i.e., the sub-gradient of  $\Psi$ ) of any  $h \in K$  – is the entire unit ball in the dual space to  $(E, \Psi)$ . Recall that*

$$\Delta(\rho) = \inf_h \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(h - f^*),$$

where the infimum is taken in the set

$$\{h \in F : \Psi(h - f^*) = \rho \text{ and } \|h - f^*\|_{L_2} \leq r(\rho)\}.$$

Since  $\Gamma_{f^*}(\rho)$  is the entire dual unit ball, it follows that  $\Delta(\rho) = \rho$ , and Theorem 3.2 in [Lecué and Mendelson \(to appear\)](#) may be applied. The desired version of Theorem 1.9 now follows from Remark 3.3 in [Lecué and Mendelson \(to appear\)](#).

### 3. Towards the examples - preliminary estimates

It is rather obvious that any implementation of Theorem 1.9 requires specific estimates on  $r_M$ ,  $r_Q$  and  $\lambda_0$ . This section is devoted to some preliminary facts that will play an instrumental part in establishing such estimates.

Our main interest is the study of upper bounds on the three processes used to define the parameter  $r_M$ ,  $r_Q$  and  $\gamma_{\mathcal{O}}$ , and which have the following forms:

$$(*) = \sup_{f \in F} \left| \sum_{i=1}^N \varepsilon_i \xi_i f(X_i) \right|, \quad (**) = \sup_{f \in F} \left| \sum_{i=1}^N (\xi_i f(X_i) - \mathbb{E} \xi f(X)) \right| \quad \text{and} \quad \mathbb{E} \sup_{f \in F} \left| \sum_{i=1}^N \varepsilon_i f(X_i) \right|,$$

where  $X_1, \dots, X_N$  are independent and distributed according to the underlying measure  $\mu$ ,  $\xi_1, \dots, \xi_N$  are independent copies of  $\xi \in L_q$  for some  $q > 2$  (though  $(\xi_i)_{i=1}^N$  need not be independent of  $(X_i)_{i=1}^N$ ), and  $(\varepsilon_i)_{i=1}^N$  are independent, symmetric  $\{-1, 1\}$ -valued random variables that are independent of  $(X_i)_{i=1}^N$  and  $(\xi_i)_{i=1}^N$ .

Standard symmetrization methods (see, e.g., [Giné and Zinn \(1984\)](#); [Ledoux and Talagrand \(1991\)](#); [van der Vaart and Wellner \(1996\)](#)) show that  $(*)$  and  $(**)$  are equivalent in expectation and in deviation (up to a slight restriction on the deviation parameter). We will present one example in which this symmetrization argument is carried out in full (Theorem 4.2), but in the other examples we will only consider the symmetrized case.

#### 3.1 Estimates for subgaussian classes

The first result is from [Mendelson \(2016\)](#), under the assumption that  $F$  is an  $L$ -subgaussian class of functions.

**Definition 3.1** *A class of functions  $F \subset L_2(\mu)$  is  $L$ -subgaussian if for every  $f, h \in F \cup \{0\}$  and every  $u \geq 1$ ,*

$$Pr(|f(X) - h(X)| \geq Lu \|f - h\|_{L_2(\mu)}) \leq 2 \exp(-u^2/2),$$

where  $X$  is distributed according to  $\mu$ .

Let  $F \subset L_2(\mu)$  and set  $\{G_f : f \in F\}$  to be the centered, canonical Gaussian process indexed by  $F$  (i.e., the covariance operator of the process is  $\mathbb{E}G_f G_g = \mathbb{E}f(X)g(X)$  for every  $f, g \in F$ ). Put

$$\ell^*(F) = \mathbb{E} \sup_{f \in F} G_f, \quad \text{and} \quad d_2(F) = \sup_{f \in F} \|f\|_{L_2(\mu)}. \quad (3.1)$$

**Theorem 3.2 (Corollary 1.10 in [Mendelson \(2016\)](#))** *Let  $X$  be distributed according to  $\mu$ , set  $\xi \in L_q$  for some  $q > 2$  and assume that  $F \subset L_2(\mu)$  is an  $L$ -subgaussian class. There are constants  $c, c_0, c_1, c_2$  and  $c_3$  that depend only on  $q$ , for which, for any  $w, u > c$ , with probability at least*

$$1 - \frac{c_0 \log^q N}{w^q N^{q/2-1}} - 2 \exp(-c_1 u^2 (\ell^*(F)/d_2(F))^2),$$

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \xi_i f(X_i) \right| \leq c_2 L w u \|\xi\|_{L_q} \frac{\ell^*(F)}{\sqrt{N}}$$

and

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N \xi_i f(X_i) - \mathbb{E} \xi f(X) \right| \leq c_2 L w u \|\xi\|_{L_q} \frac{\ell^*(F)}{\sqrt{N}}.$$

**Corollary 3.3** *Using the notation and assumptions of Theorem 3.2, let  $\xi = Y - f^*(X)$  and assume that  $\xi \in L_q$  for some  $q > 2$ . Fix  $\tau > 0$  and  $0 < \delta < 1$ , and set  $A > 0$  for which*

$$c_2 L w u \|\xi\|_{L_q} \ell^*(F \cap AD_{f^*}) \leq \tau A^2 \sqrt{N}. \quad (3.2)$$

If

$$\delta \geq \frac{c_0 \log^q N}{w^q N^{q/2-1}} + 2 \exp(-c_1 a_0 u^2) \quad (3.3)$$

then  $r_M(F, \tau, \delta) \leq A$ .

**Proof.** Clearly, it follows from Theorem 3.2 that if

$$\delta \geq \frac{c_0 \log^q N}{w^q N^{q/2-1}} + 2 \exp\left(-c_1 u^2 \left(\frac{\ell^*(F \cap AD_{f^*})}{d_2(F \cap AD_{f^*})}\right)^2\right)$$

then  $r_M(F, \tau, \delta) \leq A$ . The claim follows because if  $F \cap AD_{f^*}$  is nonempty,

$$\frac{\ell^*(F \cap AD_{f^*})}{d_2(F \cap AD_{f^*})} \geq a_0$$

for a suitable absolute constant  $a_0$ . ■

**Remark 3.4** *The estimate in Corollary 3.3 can be rather loose. The reason for the sub-optimal estimate is that usually, the Gaussian mean-width  $\ell^*(F \cap AD_{f^*})$  is much larger than  $d_2(F \cap AD_{f^*})$ . For example, let  $F = \{\langle t, \cdot \rangle : t \in S^{d-1}\}$  be the class of linear functionals on  $\mathbb{R}^d$  indexed by the Euclidean unit ball. Assume that  $X$  is an isotropic vector – that is, its covariance structure coincides with the standard Euclidean structure on  $\mathbb{R}^d$ ; that  $f^* = 0$ ; and that  $A \leq 1$ . Then  $F \cap AD_{f^*} = \{\langle t, \cdot \rangle : \|t\|_2 \leq A\}$ ,  $d_2(F \cap AD_{f^*}) = A$  and  $\ell^*(F \cap AD_{f^*}) = A\sqrt{d}$ , implying that*

$$\frac{\ell^*(F \cap AD_{f^*})}{d_2(F \cap AD_{f^*})} \geq \sqrt{d} \quad (3.4)$$

which is significantly larger than an absolute constant.

Having said that, the question of an accurate probability estimate is not the main issue of this article and we will not explore that point further.

Next, we provide an estimate on  $\gamma_{\mathcal{O}}(\rho, \tau, \delta)$  that follows from Theorem 3.2 when  $F$  is  $L$ -subgaussian and  $\xi \in L_q$  for some  $q > 2$ . The proof is identical to the one of Corollary 3.3 and is omitted.



**Corollary 3.5** *Let  $F$  be a closed, convex  $L$ -subgaussian class of functions and let  $\xi = Y - f^*(X) \in L_q$  for some  $q > 2$ . Set  $w, u > c$ ,  $\tau > 0$ ,  $0 < \delta < 1$  and  $\rho > 0$ . If  $A > 0$  satisfies*

$$c_2 L w u \|\xi\|_{L_q} \ell^*(F \cap K_\rho(f^*) \cap r(\rho) D_{f^*}) \leq \tau A \sqrt{N}.$$

and

$$\delta \geq \frac{c_0 \log^q N}{w^q N^{q/2-1}} + 2 \exp(-c_1 a_0 u^2)$$

then  $\gamma_{\mathcal{O}}(\rho, \tau, \delta) \leq A$ .

Finally, when  $F$  is an  $L$ -subgaussian class it follows from a standard chaining argument (cf. [Talagrand \(2005\)](#) or [Mendelson \(2016\)](#)) that

$$\mathbb{E} \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i f(X_i) \right| \leq \frac{c_0 L \ell^*(F)}{\sqrt{N}}. \quad (3.5)$$

This observation will be used in what follows to upper bound  $r_Q$ .

### 3.2 Estimates under a limited moment condition

In this section we shall consider the case of a class that need not be subgaussian, but rather, the growth of moments of class members is well-behaved up to some point. More accurately, we will assume that there is some  $p_0$  for which, for every  $f, h \in F \cup \{0\}$  and  $2 \leq p \leq p_0$ ,

$$\|f - h\|_{L_p} \leq L \sqrt{p} \|f - h\|_{L_2}. \quad (3.6)$$

In contrast, a subgaussian condition is equivalent to having  $\|f - h\|_{L_p} \leq L \sqrt{p} \|f - h\|_{L_2}$  for every  $p \geq 2$ .

The motivation behind this type of limited moment assumption is the LASSO estimator. Recent results on properties of the *basis pursuit algorithm* in  $\mathbb{R}^d$  [Lecué and Mendelson \(2017\)](#); [Dirksen et al. \(2014\)](#) indicate that the limited moment assumption (3.6) for  $p_0 \sim \log d$  should suffice for an optimal estimate on the performance of the LASSO – as if the class were subgaussian.

When analyzing the LASSO via the computation of the fixed points  $r_M$  and  $r_Q$ , one encounters the following scenario. Let  $X = (x_j)_{j=1}^d$  be a random vector in  $\mathbb{R}^d$  and set  $X_1, \dots, X_N$  to be independent copies of  $X$ . Let  $X_i(j)$  be the  $j$ -th coordinate of  $X_i$  and thus  $(X_i(j))_{i=1}^N$  is a random vector with independent coordinates, distributed as  $x_j$ .

Consider the random variables appearing in the definition of  $r_M$  and  $r_Q$  in the LASSO case:

$$\max_{1 \leq j \leq d} \left| \sum_{i=1}^N \varepsilon_i X_i(j) \right|, \quad (3.7)$$

and

$$\max_{1 \leq j \leq d} \left| \sum_{i=1}^N \varepsilon_i \xi_i X_i(j) \right| \quad (3.8)$$

The aim of this section is to derive upper bounds on (3.7) in expectation and (3.8) in deviation when each  $x_j$  satisfies that

$$\|x_j\|_{L_p} \leq L\sqrt{p}\|x_j\|_{L_2}$$

for  $p \lesssim \log d$ . Note that an upper bound on the centered empirical process involved in the definition of  $\gamma_{\mathcal{O}}(\rho)$  will follow from a symmetrization argument and a bound on (3.8).

The obvious difference between (3.7) and (3.8) are the multipliers  $(\xi_i)_{i=1}^N$ : although the  $x_j$ 's have  $\sim \log d$  moments,  $\xi$  may be heavy-tailed, in the sense that it only belongs to  $L_q$  for some fixed  $q > 2$ ; this difference makes the analysis of (3.8) more difficult.

Upper bounds on (3.7) and (3.8) are obtained under the following assumption.

**Assumption 3.1** *Let  $N \leq d$ ,  $t \geq 4$  and set  $p_0 = t \log d$ . Assume that  $p_0 \lesssim N$  (and note that  $p_0 \geq \log N$ ) and that for every  $1 \leq j \leq d$  and  $p \leq p_0$ ,  $\|x_j\|_{L_p} \leq L\sqrt{p}\|x_j\|_{L_2}$ . Consider  $\xi \in L_q$  for some  $q > 2$ ; let  $r = \min\{1/2 + q/4, 2\}$ ; set  $r'$  to be the conjugate index of  $r$ ; and assume that  $4r' \max\{2, 1 + a_0/a_1\} \leq t \log N$  (where  $a_0$  and  $a_1$  are two absolute constants to be specified later – in Lemma 7.3 and Lemma 7.4).*

Under this assumption we will prove the following:

**Theorem 3.6** *Let the random vector  $X$  and  $\xi = Y - f^*(X)$  satisfy Assumption 3.1. Then,*

$$\mathbb{E} \max_{1 \leq i \leq d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i(j) \right| \leq c_0 \sqrt{\log d} \cdot L \max_{1 \leq j \leq d} \|x_j\|_{L_2}. \quad (3.9)$$

Also, for every  $u > 2, v > 0, w \geq 2$  and for  $p = p_0/2$  and  $m = p/\log(eN/p)$ , one has that with probability at least

$$1 - \frac{\exp(-p/2)}{u^{2p}} - \frac{4 \exp(-p/2)}{u^{c_1 m}} - \frac{c_2 \log^q N}{w^q N^{q/2-1}} - 2 \exp(-v^2 t \log d), \quad (3.10)$$

$$\max_{1 \leq j \leq d} \left| \sum_{i=1}^N \varepsilon_i \xi_i X_i(j) \right| \leq c_3(q)(uw + u^2 v)L \|\xi\|_{L_q} \sqrt{N} \sqrt{t \log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}. \quad (3.11)$$

The proofs of both estimates in Theorem 3.6 follow from a more general result, established in Mendelson (2016), on the supremum of a centered multiplier process under a limited moment assumption like (3.6). Although the estimate in Mendelson (2016) is stated for the centered empirical process (cf. Section 4 there) its proof is actually based on an estimate on the symmetrized process. The proof of Theorem 3.6 will be presented in a final section of this article.

#### 4. The LASSO under a limited moment assumption

In this section, we obtain complexity-dependent error rates for the LASSO. Our aim is to show that the LASSO (almost) achieves the minimax rates of convergence in the true model. Observe that in this case, the true model is the smallest  $\ell_1^d$ -ball centered at 0 that contains  $t^*$ . Thus, the price one has to pay for not knowing  $\|t^*\|_1$  is rather minimal.

The rate we shall be comparing the LASSO's performance to is the minimax rate of the following problem. Let  $X \sim \mathcal{N}(0, I_{d \times d})$  and set  $\xi \sim \mathcal{N}(0, \sigma^2)$  to be independent of  $X$ . Let  $\rho > 0$ , consider an unknown  $t_0 \in \rho B_1^d$  and put  $Y = \langle X, t_0 \rangle + \xi$ .

Let  $c_0, c_1$  and  $c_2$  be well-chosen absolute constants and consider the cases  $\log d \leq N \leq c_0 d$  or  $c_1 d \leq N$ . Following [Lecué and Mendelson](#), if

$$s_M^2(\rho) = c_2 \begin{cases} \frac{\sigma^2 d}{N} & \text{if } \rho^2 N \geq \sigma^2 d^2 \\ \rho \sigma \sqrt{\frac{1}{N} \log \left( \frac{e \sigma d}{\rho \sqrt{N}} \right)} & \text{if } \sigma^2 \log d \leq \rho^2 N \leq \sigma^2 d^2 \\ \rho \sigma \sqrt{\frac{\log d}{N}} & \text{if } \rho^2 N \leq \sigma^2 \log d \end{cases} \quad \text{and } s_Q^2(\rho) \begin{cases} = 0 & \text{if } N \geq c_0 d \\ \lesssim \rho^2 / d & \text{if } c_0 d \leq N \leq c_1 d \\ \sim \frac{\rho^2}{N} \log \left( \frac{d}{N} \right) & \text{if } N \leq c_1 d, \end{cases}$$

then the minimax rate of convergence in the class  $\rho B_1^d$  is

$$\max \left\{ s_M^2(\rho), s_Q^2(\rho) \right\} \quad (4.1)$$

when  $\rho \geq \sigma \sqrt{(\log d)/N}$  and  $\rho^2$  when  $\rho \leq \sigma \sqrt{(\log d)/N}$ . Note that when  $c_0 d \leq N \leq c_1 d$  (i.e.  $N \sim d$ ),  $s_Q^2(\rho)$  decays rapidly from  $\frac{\rho^2}{N} \log(d/N)$  to 0 and there are no precise estimates on the minimax rate in that range.

It turns out that for this problem – the so-called Gaussian linear model – the minimax rate in  $\rho B_1^d$  is achieved by the Empirical Risk Minimization procedure (see, e.g., [Lecué and Mendelson](#)); however, an underlying assumption is that  $\rho$  is given. Thanks to regularization, and specifically, thanks to the LASSO, one does not need to know the value of  $\|t_0\|_1$  in advance to achieve the minimax rate, at least up to a logarithmic term. In fact, as will be explained below, the optimal rate can be achieved using regularization in a much more general framework than just the Gaussian linear model.

In what follows we focus on the performance of the LASSO in the high-dimensional case, that is, when  $N \leq c_1 d$ . One may do the same when  $N \geq c_0 d$  and we leave that to the reader.

Let  $X$  be a random vector in  $\mathbb{R}^d$  and consider the class of linear functionals  $F = \{f_t = \langle \cdot, t \rangle : t \in \mathbb{R}^d\}$ . In particular, if  $Y \in L_2$  is an arbitrary target random variable then  $f^* = f_{t^*} = \langle \cdot, t^* \rangle$  satisfies

$$t^* \in \operatorname{argmin}_{t \in \mathbb{R}^d} \mathbb{E}(Y - \langle X, t \rangle)^2. \quad (4.2)$$

As noted in the introduction, the regularization function associated with the LASSO is the  $\ell_1^d$ -norm: for every  $t = (t_j)_{j=1}^d \in \mathbb{R}^d$ ,

$$\Psi(f_t) = \|t\|_1 = \sum_{j=1}^d |t_j|.$$

Clearly, as a norm, the  $\ell_1^d$ -regularization function satisfies Assumption [1.1](#) for  $\eta = 1$ .

The LASSO with regularization parameter  $\lambda$  produces

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + \lambda \|t\|_1 \right), \quad (4.3)$$

and one would like to control  $\|f_{\hat{t}} - f_{t^*}\|_{L_2}^2 = \mathbb{E}\langle X, \hat{t} - t^* \rangle^2$ , where the expectation is taken with respect to  $X$  conditionally on the data.

It should be noted that despite the LASSO's popularity, there are relatively few results in the random design scenario we are interested in (see, e.g., [Bartlett et al. \(2012\)](#), [Massart and Meynet \(2011\)](#) and chapter 8.2 in [Koltchinskii \(2011\)](#)). The overwhelming majority of existing results have been obtained for the linear model with subgaussian noise and a fixed design (i.e., each data point is of the form  $Y_i = \langle t^*, z_i \rangle + \xi_i$ ), and the deterministic design matrix, whose rows are the vectors  $z_i$ , which satisfies some form of the *restricted isometry property* – for example, the *Restricted Eigenvalue Condition* (REC) from [Bickel et al. \(2009\)](#) or the *Compatibility Condition* (CC) from [van de Geer \(2007\)](#)).

To define the restricted eigenvalue condition, let us introduce the following notation: for  $x \in \mathbb{R}^d$  and a set  $S_0 \subset \{1, \dots, d\}$  of cardinality  $|S_0| \leq s$ , let  $S_1$  be the set of indices of the  $m$  largest coordinates of  $(|x_i|)_{i=1}^d$  that are outside  $S_0$ . Let  $x_{S_0^c}$  be the restriction of  $x$  to the set  $S_0^c = S_0^c \cup S_1$ .

**Definition 4.1** ([Bickel et al. \(2009\)](#)) *Let  $\Gamma$  be an  $N \times d$  matrix. For  $c_0 \geq 1$  and an integer  $1 \leq s \leq m \leq d$  for which  $m + s \leq d$ , the **restricted eigenvalue constant** is*

$$\kappa(s, m, c_0) = \min \left\{ \frac{\|\Gamma x\|_2}{\|x_{S_0^c}\|_2} : S_0 \subset \{1, \dots, d\}, |S_0| \leq s, \|x_{S_0^c}\|_1 \leq c_0 \|x_{S_0^c}\|_1 \right\}.$$

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

One can show (see, [Bickel et al. \(2009\)](#), [Bühlmann and van de Geer \(2011\)](#)) that if  $\Gamma$  satisfies REC and  $\lambda \gtrsim \sigma \sqrt{(\log d)/N}$ , then with high probability (with respect to the noise), simultaneously for every  $1 \leq p \leq 2$ ,

$$\|\hat{t} - t^*\|_p^p \lesssim_p \|t^*\|_0 \left( \frac{\sigma}{\kappa(s, s, 3)} \sqrt{\frac{\log d}{N}} \right)^p \tag{4.4}$$

where  $\|t^*\|_0$  is the cardinality of the support of  $t^*$ .

The main result in this section is an estimate on  $\|\hat{f} - f^*\|_{L_2}^2$  that depends on  $\|t^*\|_1$  rather than on the cardinality of the support of  $t^*$  (we refer to [Lecué and Mendelson \(to appear\)](#) for sparsity-dependent rates of convergence for the LASSO in the same framework as we consider here). Such a result follows from [Theorem 1.9](#), and to that end, one has to construct a function  $r(\cdot)$  as in [\(1.13\)](#) and to compute  $\lambda_0(\delta, \gamma)$  as in [Definition 1.8](#). We will do so in the following case: Set  $a_2 \geq 4$ ,  $2 \leq p_0 = a_2 \log d \lesssim N$ ,  $q > 2$ ,  $r = \min\{1/2 + q/4, 2\}$  and  $r'$  that is the conjugate index of  $r$ . Assume that  $4r' \max\{2, 1 + a_0/a_1\} \leq a_2 \log N$  (which is equivalent to assuming that  $q > 2 + c_1/\log N$  for some constant  $c_1 = c_1(a_0, a_1, a_2)$ ). Let  $X = (x_j)_{j=1}^d$  be a random vector and note that the coordinates  $x_1, \dots, x_d$  need not be independent.

**Assumption 4.1** *Using the above notation, assume that there are constants  $\kappa_0, \kappa$  and  $\varepsilon$  for which the following holds:*

- For every  $1 \leq j \leq d$  and every  $2 \leq p \leq p_0$ ,  $\|x_j\|_{L_p} \leq \kappa_0 \sqrt{p} \|x_j\|_{L_2}$ .
- $X$  satisfies a small-ball condition with constants  $\kappa$  and  $\varepsilon$ ; that is, for every  $t \in \mathbb{R}^d$ ,

$$\Pr \left( |\langle X, t \rangle| \geq \kappa \|\langle X, t \rangle\|_{L_2} \right) \geq \varepsilon. \quad (4.5)$$

- $\xi = Y - f^*(X) \in L_q$ .

To put this assumption in some perspective, note that an obvious underlying condition in any estimation problem with respect to the squared loss is that  $\mathbb{E}(f(X) - Y)^2$  is defined for any  $f \in F$ , and in particular, that  $\xi = Y - f^*(X) \in L_2$ . Thus, assuming that  $\xi \in L_q$  for some  $q > 2 + c_1/\log N$  is not very restrictive. Also, as noted previously, the small-ball assumption is rather minimal.

The most restrictive component of Assumption 4.1 is the moment assumption on the coordinates of  $X$ , namely that their moments exhibit a subgaussian behavior, up to, roughly,  $p \sim \log d$ .

While this assumption can be weakened to other types of moment growth condition (e.g.,  $\|x_j\|_{L_p} \leq \kappa_0 p^\alpha \|x_j\|_{L_2}$  for some  $\alpha \geq 1/2$  and up to  $p \sim \log d$ ), the resulting analysis is more involved (see [Lecué and Mendelson \(2017\)](#)), and will not be explored here.

Finally, [Lecué and Mendelson \(2017\)](#) shows that even if one assumes a subgaussian behavior of the coordinates  $x_i$ , but only up to  $p \sim (\log d)/(\log \log d)$ , Basis Pursuit may fail to recover even a 1-sparse vector, implying that the choice of  $p_0$  in Assumption 4.1 can not be relaxed significantly.

Given any  $\rho \geq 0$ , set  $M = \max_{1 \leq j \leq d} \|x_j\|_{L_2}$ , let  $\sigma_q = \|\xi\|_q$  and put

$$\Lambda(\rho) = \frac{\kappa_0 \rho M}{\kappa^2 \varepsilon} \sqrt{\frac{\log d}{N}}.$$

Moreover, for  $R(t) = \mathbb{E}(Y - \langle X, t \rangle)^2$ , one has

$$R(t) - R(t^*) = \mathbb{E} \langle X, t - t^* \rangle^2,$$

because  $\langle X, t^* \rangle$  is the best approximation of  $Y$  in a closed subspace of  $L_2$ . Thus, the estimation bounds also lead to excess risk bounds.

**Theorem 4.2** *There are absolute constants  $c_0, \dots, c_6$  for which the following holds. Assume that  $X$  and  $\xi = Y - f^*(X)$  satisfy Assumption 4.1 and that  $N \leq d$ . Let  $u > 2, v > 0$  and  $w \geq 2$ , and set  $p = (a_2/2) \log d$  and  $m = p/\log(eN/p)$ . Put*

$$\delta = \frac{\exp(-p/2)}{u^{2p}} - \frac{4 \exp(-p/2)}{u^{c_0 m}} - \frac{c_1 \log^q N}{w^q N^{q/2-1}} - 2 \exp(-v^2 t \log d) \quad (4.6)$$

and set

$$r^2(\rho) = c_2 \begin{cases} (uw + u^2 w) \sigma_q \Lambda(\rho) & \text{if } N \geq (\kappa \varepsilon / 32)^2 d \\ \max \left\{ (uw + u^2 v) \sigma_q \Lambda(\rho), \kappa^2 \Lambda^2(\rho) \right\} & \text{otherwise.} \end{cases}$$

If  $\hat{t}$  is produced by the LASSO for a regularization parameter

$$\lambda > c_4(uw + u^2v)\kappa_0 \|\xi\|_{L_q} \eta^3 M \sqrt{\frac{\log d}{N}},$$

then with probability at least  $1 - 5\delta - 2 \exp(-\varepsilon^2 N/2)$ ,

$$R(\hat{t}) - R(t^*) = \|\langle X, \hat{t} - t^* \rangle\|_{L_2}^2 \leq c_5 \max \left\{ r^2(c_6 \|t^*\|_1), \frac{\lambda}{\kappa^2 \varepsilon} \|t^*\|_1 \right\}.$$

Observe that like known estimates on the LASSO, and despite imposing considerably weaker assumptions on  $X$  and  $Y$ , the regularization parameter in Theorem 4.2 is of the order of  $\|\xi\|_{L_q} \sqrt{(\log d)/N}$ . And, when  $\|\xi\|_{L_q}$  is equivalent to  $\sigma$  – the  $L_2$  norm of  $\xi$  – then for  $N \lesssim d$ , the rate of convergence is

$$c(M) \max \left\{ \sigma \|t^*\|_1 \sqrt{\frac{\log d}{N}}, \|t^*\|_1^2 \frac{\log d}{N} \right\}$$

for a constant  $c(M)$  that depends only on  $M$ .

Hence, up to a logarithmic factor, the LASSO attains the minimax rate in  $\|t^*\|_1 B_1^d$  when  $\log d \leq N \lesssim d$  and when  $\|t^*\|_1 \geq \sigma \sqrt{\log d/N}$ ; moreover, it does so without knowing in advance the identity of the true model  $\|t^*\|_1 B_1^d$ .

Note that one may want to combine the sparsity-dependent error rate from Theorem 1.3 in Lecué and Mendelson (to appear) and the complexity-dependent error rate from Theorem 4.2. To simplify the formulation we assume that  $X$  is subgaussian, in which case the probability estimate in Theorem 4.2 can be improved and the third condition in Assumption 4.1 (i.e. that  $q > 2 + c_1/\log N$ ) can be relaxed to only  $q > 2$  (see more details in the next section, and, in particular, Theorem 5.4). Combining the two approaches, one has that when  $X$  is isotropic and  $L$ -subgaussian, and when  $\xi \in L_q$  for some  $q > 2$  then for any  $u, w > c$  with probability larger than  $1 - \delta$  for

$$\delta = 2 \exp(-c_2 N/L^8) - \frac{c_0 \log^q N}{w^q N^{q/2-1}} - c_0 \exp(-c_1 u^2/L^2),$$

the LASSO estimator  $\hat{t}$  with the universal regularization parameter  $\|\xi\|_{L_q} \sqrt{(\log d)/N}$  satisfies that

$$\|\hat{t} - t^*\|_2^2 \lesssim_{L,q} \min \left\{ \frac{\|t^*\|_0 \sigma^2 \log d}{N}, \max \left\{ \sigma \|t^*\|_1 \sqrt{\frac{\log d}{N}}, \|t^*\|_1^2 \frac{\log d}{N} \right\} \right\} \quad (4.7)$$

when  $N \gtrsim \|t^*\|_0 \log(d/\|t^*\|_0)$ .

Observe that (4.7) seemingly exhibits a different rate than Corollary 9.1 in Koltchinskii (2011) (see also (1.4)), the difference being the extra (and necessary)  $\|t^*\|_1^2 \frac{\log d}{N}$  term in (4.7). This extra term appears only in the random design scenario, and the rates of convergence of the LASSO appear to deteriorate when

$$\sigma \sqrt{\frac{N}{\log d}} \leq \|t^*\|_1 \leq \sigma \sqrt{\|t^*\|_0}. \quad (4.8)$$

However, the sparsity-dependent error rate, and therefore Equation (4.7), holds only when  $N \gtrsim \|t^*\|_0 \log(d/\|t^*\|_0)$ . And, when  $N \gtrsim \|t^*\|_0 \log d$  (which is only slightly larger than  $\|t^*\|_0 \log(d/\|t^*\|_0)$ ), the error rates in the two scenarii (random and deterministic design) are the same and are given by

$$\min \left\{ \frac{\sigma^2 \|t^*\|_0 \log d}{N}, \sigma \|t^*\|_1 \sqrt{\frac{\log d}{N}} \right\}.$$

**Proof of Theorem 4.2.** As noted previously, since  $\|\cdot\|_1$  is a norm,  $\Psi(t) = \|t\|_1$  satisfies Assumption 1.1 for  $\eta = 1$  and Theorem 1.9 may be applied. To that end, one has to control  $r(\rho) \equiv \max\{r_M(\rho), r_Q(\rho)\}$  and  $\lambda_0(\delta, \tau)$ . In what follows we will invoke the results of Section 3 and estimate these parameters.

Set  $F(f^*, \rho) = F \cap K_\rho(f^*) - f^*$  and recall that  $r_Q(\rho) = r_Q(F \cap K_\rho(f^*), \kappa\varepsilon/32)$  is determined by the behavior of

$$(\star) = \mathbb{E} \sup_{f \in F(f^*, \rho) \cap rD} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i f(X_i) \right|; \quad (4.9)$$

as a consequence, it suffices to upper bound  $(\star)$ . Let  $\mathcal{E} = \{t \in \mathbb{R}^d : \mathbb{E}\langle X, t \rangle^2 \leq 1\}$ , put  $\mathcal{E}^\circ$  to be the polar of  $\mathcal{E}$  (that is,  $\mathcal{E}^\circ = \{u : \sup_{t \in \mathcal{E}} |\langle u, t \rangle| \leq 1\}$ ), and set  $\|t\|_{\mathcal{E}} = \sup_{x \in \mathcal{E}} \langle x, t \rangle$ . Thus,

$$\begin{aligned} (\star) &= \mathbb{E} \sup_{t \in \rho B_1^d \cap r\mathcal{E}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \langle X_i, t \rangle \right| \leq \min \left\{ \mathbb{E} \sup_{t \in \rho B_1^d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \langle X_i, t \rangle \right|, \mathbb{E} \sup_{t \in r\mathcal{E}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \langle X_i, t \rangle \right| \right\} \\ &= \min \left\{ \rho \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i \right\|_{\ell_\infty^d}, r \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i \right\|_{\mathcal{E}^\circ} \right\}. \end{aligned}$$

It is standard to verify (see, for instance, the proof of Lemma 2.2 in [Lecué and Mendelson \(2016\)](#)) that

$$\mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i \right\|_{\mathcal{E}^\circ} \leq \sqrt{d}.$$

Moreover, by (3.9),

$$\mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i \right\|_{\ell_\infty^d} = \mathbb{E} \max_{1 \leq j \leq d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i(j) \right| \leq c_0 \kappa_0 \sqrt{\log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}.$$

Therefore,

$$(\star) \leq \min \left\{ c_0 \rho \kappa_0 \sqrt{\log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}, r \sqrt{d} \right\}, \quad (4.10)$$

and setting  $\gamma = \kappa\varepsilon/32$ , one has

$$r_Q(\rho) \leq \begin{cases} 0 & \text{if } N \geq \gamma^2 d \\ \frac{c_0 \rho \kappa_0}{\gamma} \sqrt{\frac{\log d}{N}} M & \text{otherwise.} \end{cases}$$

Next, let us establish a high probability upper bound on  $r_M(\rho) = r_M(F \cap K_\rho(f^*), \kappa^2\varepsilon/80, \delta/4)$ . Note that

$$\phi_N(F \cap K_\rho(f^*), f^*, s) = \sup_{t \in \rho B_1^d \cap s\mathcal{E}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i \langle X_i, t \rangle \right| \leq \rho \max_{1 \leq j \leq d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i X_i(j) \right|.$$

Applying the second part of Theorem 3.6 for  $u > 2, v > 0, w \geq 2, p = (a_2/2) \log d, m = p/\log(eN/p)$  and

$$\delta = \frac{\exp(-p/2)}{u^{2p}} - \frac{4 \exp(-p/2)}{u^{c_0 m}} - \frac{c_1 \log^q N}{w^q N^{q/2-1}} - 2 \exp(-v^2 t \log d), \quad (4.11)$$

it follows that with probability at least  $1 - \delta$ ,

$$\phi_N(F \cap K_\rho(f^*), f^*, s) \leq c_2 \kappa_0 (uw + u^2 v) \|\xi\|_{L_q} \rho M \sqrt{\log d};$$

thus,

$$r_M^2(\rho) \leq \frac{c_3 \kappa_0 (uw + u^2 v)}{\kappa^2 \varepsilon} \|\xi\|_{L_q} \rho M \sqrt{\frac{\log d}{N}}.$$

Finally, let us identify an upper bound on  $\lambda_0(\delta, \tau)$  for  $\tau = 3/(80\eta^3)$ . Let  $\{e_1, \dots, e_d\}$  be the canonical basis of  $\mathbb{R}^d$ . Since  $K_\rho(f^*) = \{t : \|t - t^*\|_1 \leq \rho\}$ , we have

$$\begin{aligned} (\star_1) &= \sup_{f \in F \cap K_\rho(f^*) \cap r(\rho) D_{f^*}} \left( \frac{1}{N} \sum_{i=1}^N \xi_i (f - f^*)(X_i) - \mathbb{E} \xi (f - f^*)(X) \right) \\ &\leq \rho \max_{t - t^* \in \{\pm e_1, \dots, \pm e_d\}} \left( \frac{1}{N} \sum_{i=1}^N \xi_i \langle X_i, t - t^* \rangle - \mathbb{E} \xi \langle X, t - t^* \rangle \right) \\ &= \rho \max_{t \in \{\pm e_1, \dots, \pm e_d\}} \left( \frac{1}{N} \sum_{i=1}^N \xi_i \langle X_i, t \rangle - \mathbb{E} \xi \langle X, t \rangle \right). \end{aligned}$$

Recall that  $X = (x_j)_{j=1}^d$ . By a standard symmetrization argument (see, for example, Lemma 2.3.7 in [van der Vaart and Wellner \(1996\)](#)), if  $z \geq 4 \max_{1 \leq j \leq d} \sqrt{\text{Var}(\xi x_j)}/N$  then

$$Pr \left( \max_{t \in \{\pm e_1, \dots, \pm e_d\}} \frac{1}{N} \sum_{i=1}^N \xi_i \langle X_i, t \rangle - \mathbb{E} \xi \langle X, t \rangle \geq z \right) \leq 4 Pr \left( \max_{t \in \{\pm e_1, \dots, \pm e_d\}} \frac{1}{N} \sum_{i=1}^N \varepsilon_i \xi_i \langle X_i, t \rangle \geq \frac{z}{4} \right).$$

Note that  $\sqrt{\text{Var}(\xi x_j)} \leq \sqrt{\mathbb{E} \xi^2 x_j^2} \leq \|\xi\|_{L_q} \|x_j\|_{L_{2q'}}$  where  $q'$  is the conjugate index of  $q/2$ . Therefore,  $\sqrt{\text{Var}(\xi x_j)} \leq \kappa_0 \sqrt{2q'} \|\xi\|_{L_q} M$  as long as  $2q' \leq a_2 \log d$ , i.e., when  $q \geq 2 +$



$2/(a_2 \log d - 1)$  – which is the case under Assumption 4.1. Therefore, applying the second part of Theorem 3.6 for  $\delta$  as in (4.11), it follows that with probability at least  $1 - \delta$ ,

$$(\star_1) \leq \rho c_0 \kappa_0 (uw + u^2v) \|\xi\|_{L_q} M \sqrt{\frac{\log d}{N}},$$

and, for  $\tau = 3/(80\eta^3)$  one may select

$$\lambda_0(\delta, \tau) = c_4 \kappa_0 (uw + u^2v) \|\xi\|_{L_q} \eta^3 w M \sqrt{\frac{\log d}{N}}.$$

■

## 5. Regularization methods for subgaussian classes

In this section, we assume that  $X$  is a random vector that takes its values in a Hilbert space  $\mathcal{H}$ . The main examples we consider are when  $\mathcal{H}$  is the  $d$ -dimensional Euclidean space and when it is the space of  $m \times T$  matrices endowed with the Frobenius norm.

The inner product in  $\mathcal{H}$  is denoted by  $\langle \cdot, \cdot \rangle$ , and the norm and unit ball endowed by the inner product are denoted by  $\|\cdot\|_{\mathcal{H}}$  and  $B_{\mathcal{H}} = \{t \in \mathcal{H} : \|t\|_{\mathcal{H}} \leq 1\}$  respectively.

There is another natural Hilbertian structure on  $\mathcal{H}$ , endowed by  $\Sigma = \mathbb{E}XX^\top$ , the covariance operator associated with the random vector  $X$ . The corresponding unit ball  $\mathcal{E} = \{t \in \mathcal{H} : \mathbb{E}\langle X, t \rangle^2 \leq 1\}$ , is an ellipsoid in  $\mathcal{H}$ .

Let  $T \subset \mathcal{H}$  be a closed and convex set and put

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

Let  $\Psi(\cdot)$  be a regularization function on  $\mathcal{H}$  that satisfies Assumption 1.1 and set

$$\hat{t} \in \operatorname{argmin}_{t \in T} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + \lambda \Psi(t) \right) \quad (5.1)$$

for a well-chosen regularization parameter  $\lambda$ .

Unlike the results of the previous section, in what follows we assume that  $F = \{\langle t, \cdot \rangle : t \in T\}$  is an  $L$ -subgaussian class (see Definition 3.1). Moreover,  $F$  satisfies a small-ball property with constants that depend only on  $L$ . Indeed, observe that for every  $t \in T$

$$\|\langle X, t \rangle\|_{L_4} \lesssim L \|\langle X, t \rangle\|_{L_2},$$

and applying the Paley-Zygmund inequality (see, e.g., Corollary 3.3.2 in [de la Peña and Giné \(1999\)](#)),

$$\Pr \left( |\langle X, t \rangle| \geq \kappa \|\langle X, t \rangle\|_{L_2} \right) \geq \varepsilon \text{ for } \kappa = 1/2 \text{ and } \varepsilon = c/L^4. \quad (5.2)$$

From here on we say that the random vector  $X$  taking its values in  $\mathcal{H}$  is  $L$ -subgaussian if the class consisting of all the linear functionals on  $\mathcal{H}$ , i.e.,  $\{\langle t, \cdot \rangle : t \in \mathcal{H}\}$ , is  $L$ -subgaussian. Also, throughout this section, we assume that  $\xi = Y - f^*(X) \in L_q$  for some  $q > 2$ , and denote  $\sigma_q = \|\xi\|_{L_q}$  and

$$T \cap K_\rho(t^*) = \{t \in T : \Psi(t - t^*) \leq \rho\}.$$

### 5.1 ‘Heavy tailed’ noise

Thanks to the subgaussian assumption, both  $r(\rho)$  and  $\lambda_0 = \lambda_0(\delta, 3/(80\eta^3))$  may be determined using the Gaussian mean-widths of the sets  $T \cap K_\rho(t^*)$  for all  $\rho > 0$ . Recall that for  $T_0 \subset \mathcal{H}$  the Gaussian mean-width of  $T_0$  is  $\ell^*(T_0) = \mathbb{E} \sup_{t \in T_0} G_t$ , where  $(G_t)_{t \in T_0}$  is the centered canonical Gaussian process indexed by  $T_0$  with covariance structure given by  $\mathbb{E} G_{t_1} G_{t_2} = \mathbb{E} \langle X, t_1 \rangle \langle X, t_2 \rangle$  for every  $t_1, t_2 \in T$ .

**Definition 5.1** Let  $r\mathcal{E}_{t^*} = \{t \in \mathcal{H} : \|\langle t - t^*, \cdot \rangle\|_{L_2(\mu)} \leq r\} = t^* + r\mathcal{E}$ , and for  $\alpha, \beta > 0$  set

$$\tilde{r}_Q(\rho, \alpha) = \inf \left\{ r > 0 : \ell^*(T \cap K_\rho(t^*) \cap r\mathcal{E}_{t^*}) \leq \alpha r \sqrt{N} \right\}$$

and

$$\tilde{r}_M(\rho, \beta) = \inf \left\{ r > 0 : \ell^*(T \cap K_\rho(t^*) \cap r\mathcal{E}_{t^*}) \leq \beta r^2 \sqrt{N} \right\}.$$

Let  $c_0$  be an absolute constant to be specified later. Fix  $u, w > c$ , and  $\varepsilon$  and  $\kappa$  as in (5.2). Consider

$$\alpha = \frac{\kappa\varepsilon}{c_0 L}, \quad \beta = \frac{\kappa^2 c_1 \varepsilon}{L w u \|\xi\|_{L_q}}, \quad \text{and} \quad \gamma = c_0 \eta^3 L w u \|\xi\|_{L_q}, \quad (5.3)$$

put

$$r(\rho) \geq \max \{ \tilde{r}_Q(\rho, \alpha), \tilde{r}_M(\rho, \beta) \} \quad (5.4)$$

and set

$$\lambda_0(\gamma) = \gamma \sup_{\rho > 0, t^* \in T} \frac{\ell^*(T \cap K_\rho(t^*) \cap r(\rho)\mathcal{E}_{t^*})}{\rho \sqrt{N}}. \quad (5.5)$$

The first result we present is rather general and holds for any closed and convex subset  $T \subset \mathcal{H}$  and any regularization function satisfying Assumption 1.1. It allows one to take into account an additional constraint on the ‘signal’  $t^* \in T$ .

**Theorem 5.2** *There are absolute constants  $c, c_1$  and  $c_2$  for which the following holds. Let  $\Psi$  be a regularization function satisfying Assumption 1.1. Assume that  $X$  is  $L$ -subgaussian for some  $L > 0$  and that  $\xi = Y - \langle X, t^* \rangle$  is in  $L_q$  for some  $q > 2$ .*

*If  $\hat{t}$  is given by (5.1) for a regularization parameter  $\lambda > \lambda_0(\gamma)$  as in (5.5), then with probability larger than*

$$1 - 2 \exp(-N\varepsilon^2/8) - \frac{c_0 \log^q N}{w^q N^{q/2-1}} - c_0 \exp(-c_1 u^2/L^2), \quad (5.6)$$

$$\|\langle X, \hat{t} - t^* \rangle\|_{L_2}^2 \leq \max \{ r(10\eta\Psi(t^*))^2, (32/\kappa^2\varepsilon)\lambda\Psi(t^*) \}$$

for  $r(\cdot)$  given by (5.4).

**Proof.** The proof follows from Theorem 1.9 by estimating  $r(\rho)$  and  $\lambda_0$  using the ‘local’ Gaussian mean-widths of the sets  $T \cap K_\rho(t^*)$ .

Since  $X$  is  $L$ -subgaussian, the process  $\{\langle X, t \rangle : t \in \mathcal{H}\}$  is  $L$ -subgaussian. Setting  $F = \{\langle t, \cdot \rangle : t \in \mathcal{H}\}$  and  $f^* = \langle t^*, \cdot \rangle$ , a standard chaining argument shows that

$$\mathbb{E} \sup_{f \in F \cap K_\rho(f^*) \cap rD_{f^*}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right| \leq c_0 L \frac{\ell^*(T \cap K_\rho(t^*) \cap r\mathcal{E}_{t^*})}{\sqrt{N}}.$$

Thus,

$$r_Q \left( F \cap K_\rho(f^*), \frac{\kappa\varepsilon}{32} \right) \leq \tilde{r}_Q(\rho, \alpha). \quad (5.7)$$

As for the fixed point associated with the multiplier process, it follows from Corollary 3.3 that

$$r_M \left( F \cap K_\rho(f^*), \frac{\kappa^2\varepsilon}{160}, \frac{\delta}{4} \right) \leq \tilde{r}_M(\rho, \beta) \quad (5.8)$$

for  $\beta$  as defined in (5.3), and as long as

$$\frac{\delta}{4} \geq \frac{c_0 \log^q N}{w^q N^{q/2-1}} + 2 \exp(-c_1 u^2 / L^2).$$

Finally by Corollary 3.5,  $\lambda_0(\delta, \gamma) \leq \lambda_0(\gamma)$ . The claim now follows from Theorem 1.9. ■

If one is to apply Theorem 5.2, an essential component is an upper bound on  $\ell^*(T \cap K_\rho(t^*) \cap r\mathcal{E}_{t^*})$ , leading to estimates on  $r$  and  $\lambda$ . To simplify the analysis we shall use an additional assumption on  $\Psi$ :

**Assumption 5.1** *Assume that for every  $x, y \in \mathcal{H}$  and  $\lambda \geq 0$ ,*

$$\Psi(x) = \Psi(-x), \quad \Psi(x + y) \leq \eta(\Psi(x) + \Psi(y)) \text{ and } \Psi(\lambda x) \leq \lambda \Psi(x). \quad (5.9)$$

Also, recall that  $\mathcal{E} = \{t \in \mathcal{H} : \mathbb{E}\langle X, t \rangle^2 \leq 1\}$ ,  $\sigma_q = \|\xi\|_q$  and set  $K = \{t \in \mathcal{H} : \Psi(t) \leq 1\}$ .

**Theorem 5.3** *Assume that  $\Psi$  satisfies Assumption 5.1 and that the assumptions of Theorem 5.2 hold. Let  $\Lambda(\rho) \geq \rho \ell^*(K) / \sqrt{N}$  for every  $\rho > 0$ ,  $w, u > c$  and consider the RERM*

$$\hat{t} \in \operatorname{argmin}_{t \in \mathcal{H}} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_0 \eta^3 L w u \sigma_q \Lambda(\Psi(t)) \right).$$

Then, with probability larger than the one in (5.6),

$$R(\hat{t}) - R(t^*) = \|\langle X, \hat{t} - t^* \rangle\|_{L_2}^2 \leq c_0 r^2 (10\eta \Psi(t^*))$$

where, for  $\alpha$  and  $\beta$  defined in (5.3) and, for any  $\rho \geq 0$ ,

$$r^2(\rho) = \begin{cases} \frac{\Lambda(\rho)}{\beta} & \text{if } N \geq (\ell^*(\mathcal{E})/\alpha)^2 \\ \max \left\{ \frac{\Lambda(\rho)}{\beta}, \frac{\Lambda^2(\rho)}{\alpha^2} \right\} & \text{otherwise.} \end{cases} \quad (5.10)$$

**Proof.** The result follows immediately from Theorem 5.2. Indeed, for every  $\rho > 0$ ,  $r > 0$  and  $t^* \in T = \mathcal{H}$ ,

$$\ell^*(T \cap K_\rho(t^*) \cap r\mathcal{E}_{t^*}) = \ell^*(K_\rho(0) \cap r\mathcal{E}) \leq \ell^*(\rho K \cap r\mathcal{E}) \leq \min \{\rho \ell^*(K), r \ell^*(\mathcal{E})\},$$

because  $K_\rho(0) \subset \rho K = \{\rho t : t \in K\}$ . ■

Note that in a  $d$ -dimensional space, the trivial bound  $\ell^*(\mathcal{E}) \leq \sqrt{d}$  holds (see, e.g., Lemma 2.2 in [Lecué and Mendelson \(2016\)](#)). Therefore, one only needs to control  $\ell^*(K)$ . In the next section, we provide several examples of applications of Theorem 5.3 that follow from estimates on  $\ell^*(K)$ . We will simplify the analysis further by assuming that there is some compatibility between the norm  $\|\cdot\|_{\mathcal{H}}$  and the one endowed by the covariance structure of  $X$ :

**Assumption 5.2** *Assume that  $X$  is isotropic; that is, for every  $t \in \mathcal{H}$ ,  $(\mathbb{E}\langle X, t \rangle^2)^{1/2} = \|t\|_{\mathcal{H}}$ .*

Observe that Assumption 5.2 implies that  $\ell^*(K) = \mathbb{E} \sup_{t \in K} G_t$ , where  $(G_t)_{t \in K}$  is the canonical Gaussian process indexed by  $K$  with the covariance  $\mathbb{E}G_{t_1}G_{t_2} = \langle t_1, t_2 \rangle$  for every  $t_1, t_2 \in K$ . Indeed, this follows because the inner-product in  $\mathcal{H}$  coincides with the one endowed by  $L_2(\mu)$ .

## 5.2 Regularization methods in $\mathbb{R}^d$

Consider a regularization function  $\Psi(\cdot)$  satisfying Assumption 5.1. Assume that  $X$  is  $L$ -subgaussian and isotropic in  $\mathbb{R}^d$  with respect to the standard Euclidean inner-product, and that  $\xi \in L_q$  for some  $q > 2$ . Let  $u, w > c$ . For any  $\rho \geq 0$  set  $\Lambda(\rho) \geq \rho \ell^*(K) / \sqrt{N}$  and put

$$r^2(\rho) \sim_{L,q} \begin{cases} wu\sigma_q\Lambda(\rho) & \text{when } N \gtrsim_L d \\ \max\{wu\sigma_q\Lambda(\rho), \Lambda^2(\rho)\} & \text{otherwise.} \end{cases}$$

It follows from Theorem 5.3 that if

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_0 \eta^3 L w u \sigma_q \Lambda(\Psi(t)) \right) \quad (5.11)$$

then with probability larger than the one in (5.6)

$$\|\langle \hat{t} - t^*, \cdot \rangle\|_{L_2(\mu)}^2 \lesssim r^2(10\eta\Psi(t^*)).$$

As a consequence, one can derive an estimation result for (5.11) whenever  $\ell^*(K)$  may be controlled from above. In the following section, we shall apply this observation to some classical problems and compare the error rates obtained by the RERM (5.11) to the minimax rate in the true model  $\{t \in T : \Psi(t) \leq \Psi(t^*)\}$ .

**Example:  $\ell_p$ -regularization for  $1 \leq p \leq \infty$ .** Consider the regularization function  $\Psi(t) = \|t\|_p$  for some  $p \geq 1$ . Assumption 5.1 holds with  $\eta = 1$  because  $\|\cdot\|_p$  is a norm. In order to apply the general result for the RERM in (5.11), one has to estimate the Gaussian mean-width of the unit ball associated with the regularization function  $\Psi(\cdot) = \|\cdot\|_p$ .

In the range  $1 \leq p \leq 1 + (\log d)^{-1}$ , we obtain the same performance as the LASSO, because  $B_1^d \subset B_p^d \subset cB_1^d$  for a suitable absolute constant  $c$ , implying that  $\ell^*(B_p^d) \sim \ell^*(B_1^d) \sim \sqrt{\log(ed)}$ .

When  $1 + (\log(ed))^{-1} \leq p$ , set  $r$  to be the conjugate index for  $p$  and one may easily verify that  $\ell^*(B_p^d) \sim \sqrt{r}d^{1/r}$ .

Applying Theorem 5.3, one has the following:

**Theorem 5.4** *Under the assumptions of Theorem 5.3 and using its notation,*

- *If  $1 \leq p \leq 1 + 1/(\log d)$  and*

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_2 \eta_p^3 L w u \sigma_q \|t\|_p \sqrt{\frac{\log d}{N}} \right)$$

*then with probability larger than the one in (5.6),*

$$\|\hat{t} - t^*\|_2^2 \lesssim_{p,L,q} \begin{cases} w u \sigma_q \|t^*\|_p \sqrt{\frac{\log d}{N}} & \text{if } N \gtrsim_L d, \\ \max \left\{ w u \sigma_q \|t^*\|_p \sqrt{\frac{\log d}{N}}, \|t^*\|_p^2 \frac{\log d}{N} \right\} & \text{otherwise.} \end{cases}$$

- *If  $p \geq 1 + 1/(\log d)$  and*

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_2 \sigma_q L w u \|t\|_p \frac{\sqrt{p/(p-1)} d^{(p-1)/p}}{\sqrt{N}} \right),$$

*then with probability larger than the one in (5.6)*

$$\|\hat{t} - t^*\|_2^2 \lesssim_{L,q} \begin{cases} w u \sigma_q \|t^*\|_p \frac{d^{(p-1)/p}}{p\sqrt{N}} & \text{if } N \gtrsim_L d, \\ \max \left\{ w u \sigma_q \|t^*\|_p \frac{d^{(p-1)/p}}{\sqrt{N}}, \|t^*\|_p^2 \frac{d^{2(p-1)/p}}{N} \right\} & \text{otherwise.} \end{cases}$$

**Remark.** [the case  $0 < p < 1$ ] Despite being a non-convex function,  $\ell_p$ -regularization for  $0 < p < 1$  has attracted much attention in the context of Signal Processing and High-Dimensional Statistics. Among the problems studied using  $\ell_p$  regularization were the linear regression model with a deterministic design (cf. [Raskutti et al. \(2011\)](#); [Rigollet and Tsybakov \(2011\)](#); [Verzelen \(2012\)](#)); the sequence space model [Donoho and Johnstone \(1994\)](#); [Abramovich et al. \(2006\)](#); and the random design linear regression model [Wang et al. \(2014\)](#).

From our point of view, there is no particular restriction on  $p$  as long as the regularization function satisfies Assumption 5.1. We can therefore consider the regularization function  $\Psi(t) = \|t\|_p$  for any  $0 < p < 1$ . In that range of  $p$ , Assumption 5.1 holds for  $\eta = \eta_p = 2^{1/p}$  (see, for example, page 2 in [Edmunds and Triebel \(1996\)](#)) and the Gaussian mean width of the “unit ball” associated with  $\Psi(\cdot) = \|\cdot\|_p$  for  $0 < p < 1$  can also be computed.

To that end, let  $\{e_1, \dots, e_d\}$  be the canonical basis of  $\mathbb{R}^d$ . Since  $\{\pm e_1, \dots, \pm e_d\} \subset B_p^d \subset B_1^d$  for  $p < 1$ , it is evident that  $\ell^*(B_p^d) \sim \sqrt{\log d}$ . Thus, the error rates of the LASSO from Theorem 4.2 dominate all the  $\ell_p$ -regularization rates when  $0 < p \leq 1$ .

However, the resulting rate is not the minimax rate in the true model, as can be seen from [Raskutti et al. \(2011\)](#). Indeed, fix  $0 < p \leq 1$ . Consider an unknown  $t^* \in \rho B_p^d$  and the corresponding Gaussian linear model  $Y_i = \langle x_i, t^* \rangle + W_i, i = 1, \dots, N$ , where the matrix whose rows are  $(x_i)_{i=1}^N$  satisfies some RIP property and  $W_1, \dots, W_N$  are independent, centered Gaussian variables with variance  $\sigma^2$ . For specific asymptotic rates in  $N$  and  $d$  (see

Raskutti et al. (2011) for a precise formulation), the authors show that minimax rate of the problem is given by

$$\sigma^2 \rho \left( \frac{\log d}{N} \right)^{1-\frac{p}{2}},$$

and similar results have been obtained in Wang et al. (2014). Thus, our estimate recovers the minimax rate in the true model only when  $p = 1$ . When  $0 < p < 1$ , it is possible that the choice of the  $\Psi(t) = \|t\|_p$  is suboptimal, and instead one should use  $\Psi(t) = \|t\|_p^p$  as was suggested in Rohde and Tsybakov (2011) for the problem of  $S_p$ -regularization for  $0 < p \leq 1$ .

**Example: weak- $\ell_p$ -regularization for  $0 < p \leq 1$ .** Weak- $\ell_p$  norms have been used to model sparsity in High-Dimensional Statistics (see, for instance, Abramovich et al. (2006); Wang et al. (2014)). To define those norms, let  $t_1^* \geq t_2^* \geq \dots \geq t_d^*$  be the non-increasing rearrangement of  $(|t_i|)_{i=1}^d$ . Set  $\|t\|_{p\infty} = \max_{1 \leq j \leq d} j^{1/p} t_j^*$  and put  $B_{p\infty}^d = \{t \in \mathbb{R}^d : t_j^* \leq j^{-1/p} \text{ for every } 1 \leq j \leq d\}$ .

One can use the following well-known fact (see, e.g., Theorem B in Gordon et al. (2007)) to control the Gaussian mean-width of the unit ball associated with  $\|\cdot\|_{p\infty}$ .

**Proposition 5.5** For  $0 < p \leq 1$ ,

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

Applying Theorem 5.3, one has the following:

**Theorem 5.6** Under the assumptions of Theorem 5.3 and using its notation,

- If  $p < 1$  and

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_2 \eta_p^3 \sigma_q L w u \|t\|_{p\infty} \sqrt{\frac{\log d}{N}} \right),$$

then with probability larger than the one in (5.6)

$$\|\hat{t} - t^*\|_2^2 \lesssim_{p,L,q} \begin{cases} \sigma_q w u \|t^*\|_{p\infty} \sqrt{\frac{\log d}{N}} & \text{if } N \gtrsim_L d, \\ \max \left\{ \sigma_q w u \|t^*\|_p \sqrt{\frac{\log d}{N}}, \|t^*\|_{p\infty}^2 \frac{\log d}{N} \right\} & \text{otherwise.} \end{cases}$$

- If  $p = 1$  and

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_2 \eta_1^3 \sigma_q \|t\|_{1\infty} \sqrt{\frac{\log^3 d}{N}} \right),$$

then with probability larger than the one in (5.6)

$$\|\hat{t} - t^*\|_2^2 \lesssim_{L,\delta,q} \begin{cases} \sigma_q w u \|t^*\|_{1\infty} \sqrt{\frac{\log^3 d}{N}} & \text{if } N \gtrsim_L d, \\ \max \left\{ \sigma_q w u \|t^*\|_{1\infty} \sqrt{\frac{(\log d)^3}{N}}, \|t^*\|_{1\infty}^2 \frac{\log^3 d}{N} \right\} & \text{otherwise.} \end{cases}$$

**Example: the Micchelli, Morales and Pontil’s regularization functions.**

Let  $\Theta$  be a nonempty convex cone in  $[0, \infty)^d$ , and for every  $t \in \mathbb{R}^d$  set

$$\Omega(t|\Theta) = \inf_{\theta \in \Theta} \frac{1}{2} \sum_{j=1}^d \left( \frac{t_j^2}{\theta_j} + \theta_j \right). \quad (5.12)$$

It was shown in [Maurer et al. \(2010\)](#) that  $\Omega(\cdot|\Theta)$  is a norm on  $\mathbb{R}^d$ .

This family of norms captures several classical regularization functions, by an appropriate choice of the cone  $\Theta$ . For instance, the  $\ell_1^d$ -norm is obtained by selecting  $\Theta = [0, \infty)^d$ . Also, the *group LASSO* introduced in [Yuan and Lin \(2006\)](#) is generated by a cone: indeed, if  $(G_1, \dots, G_T)$  is a partition of  $\{1, \dots, d\}$  and

$$\Theta = \{\theta \in [0, \infty)^d \text{ that is constant within each group } G_\ell\}, \quad (5.13)$$

then

$$\Omega(t|\Theta) = \sum_{\ell=1}^T \sqrt{|G_\ell|} \|t_{|G_\ell}\|_2,$$

where  $|G_\ell|$  is the cardinality of the set of coordinates  $G_\ell$  and  $t_{|G_\ell}$  is the restriction of  $t$  to  $G_\ell$ .

Error bounds for procedures that use  $\Psi(t) = \Omega(t|\Theta)$  as regularization functions have been established in [Maurer and Pontil \(2012\)](#), under the assumption that the loss functions is bounded and Lipschitz (see Theorem 1 there). Naturally, the squared loss is not covered by such a result because it is not bounded in  $\mathbb{R}^d$ , nor is it Lipschitz. Our aim is to provide a version of the result from [Maurer and Pontil \(2012\)](#) for the quadratic loss and when  $X$  is subgaussian and the noise belongs to  $L_q$  for some  $q > 2$ . The first step in that direction is the following estimate on the Gaussian mean width of the corresponding unit ball.

**Proposition 5.7** *Let  $\Theta$  be a nonempty convex cone in  $[0, \infty)^d$  and set  $\mathcal{B} = \{t : \Omega(t|\Theta) \leq 1\}$ . Let  $S_1^{d-1}$  be the unit sphere of  $\ell_1^d$  and put  $\mathcal{E}x$  to be the set of extreme points of  $\Theta \cap S_1^{d-1}$ . If  $M = \max_{a \in \mathcal{E}x} \|a\|_\infty^{1/2}$ , then, for an absolute constant  $c$ ,*

$$\ell^*(\mathcal{B}) \leq 1 + cM \sqrt{2 \log(|\mathcal{E}x|)}. \quad (5.14)$$

The proof of Proposition 5.7 may be derived in various ways (see a similar result in [Maurer and Pontil \(2012\)](#)), though we will use a chaining argument which actually leads to a stronger estimate than (5.14).

**Definition 5.8** *Let  $T \subset \mathbb{R}^d$  and set  $\|\cdot\|$  to be a norm on  $\mathbb{R}^d$ . For every  $\alpha > 1$  let*

$$\gamma_\alpha(T, \|\cdot\|) = \inf_{(T_s)} \sup_{t \in T} \sum_{s=0}^{\infty} 2^{s/\alpha} \|\pi_{s+1}t - \pi_s t\|;$$

*the infimum is taken with respect to all sequences  $(T_s)$  of subsets of  $T$  for which  $|T_0| = 1$  and for  $s \geq 1$ ,  $|T_s| \leq 2^{2^s}$ , and  $\pi_s t$  denotes the nearest point to  $t$  in  $T_s$  with respect to  $\|\cdot\|$ .*

Clearly, if  $T$  is finite then  $\gamma_\alpha(T, \|\cdot\|) \lesssim \sup_{t \in T} \|t\| \cdot \log^{1/\alpha} |T|$ .

**Proof of Proposition 5.7.** It is straightforward to verify (see, e.g., [Maurer et al. \(2010\)](#)) that the dual norm to  $\Omega(\cdot|\Theta)$  is

$$\Omega^*(t|\Theta) = \max_{a \in \mathcal{E}x} \left( \sum_{j=1}^d a_j t_j^2 \right)^{1/2}. \quad (5.15)$$

Let  $g_1, \dots, g_d$  be independent, standard Gaussian random variables, Applying a Bernstein type inequality for a sum of independent  $\psi_1$  random variables (see Corollary 2.10 in [Talagrand \(1994\)](#)), it follows that for every  $a_1, \dots, a_N$ , every  $u > 0$  and any  $s \in \mathbb{N}$ ,

$$\Pr \left[ \left| \sum_{j=1}^d a_j (g_j^2 - 1) \right| \geq u 2^{s/2} \|a\|_2 + u^2 2^s \|a\|_\infty \right] \leq 2 \exp(-c_1 2^s u^2).$$

Hence, using a standard chaining argument,

$$\mathbb{E} \sup_{a \in \mathcal{E}x} \sum_{j=1}^d a_j g_j^2 \leq 1 + c_2 (\gamma_2(\mathcal{E}x, \|\cdot\|_2) + \gamma_1(\mathcal{E}x, \|\cdot\|_\infty)).$$

Now one may apply the trivial estimates on  $\gamma_1$  and  $\gamma_2$ . Firstly,  $\gamma_1(\mathcal{E}x, \|\cdot\|_\infty) \lesssim M^2 \log(|\mathcal{E}x|)$ , and secondly, noting that  $|\sum_{j=1}^d a_j| \leq \|a\|_1 = 1$  and thus  $\|a\|_2 \leq \|a\|_\infty^{1/2}$ , one has  $\gamma_2(\mathcal{E}x, \|\cdot\|_2) \lesssim M \sqrt{\log(|\mathcal{E}x|)}$ . Therefore, by Jensen's inequality,

$$\mathbb{E} \sup_{a \in \mathcal{E}x} \left( \sum_{j=1}^d a_j g_j^2 \right)^{1/2} \leq 1 + cM \sqrt{\log(|\mathcal{E}x|)}.$$

■

**Theorem 5.9** *Using the notation above and of Theorem 5.3, let*

$$\Lambda(t) = \Omega(t|\Theta) M \sqrt{\frac{\log(|\mathcal{E}x|)}{N}}.$$

If

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_2 \sigma_q L w u \Lambda(t) \right)$$

then with probability larger than the one in (5.6)

$$\|\langle \hat{t} - t^*, \cdot \rangle\|_{L_2(\mu)}^2 \lesssim_{L,q} \begin{cases} \sigma_q w u \Lambda(t^*) & \text{if } N \gtrsim_L d, \\ \max \{ \sigma_q L w u \Lambda(t^*), \Lambda^2(t^*) \} & \text{otherwise.} \end{cases}$$



When  $\Theta = [0, \infty)^d$  then  $M\sqrt{\log(\mathcal{E}x)} \lesssim \sqrt{\log d}$ . Hence, Theorem 5.9 yields the same error rate as the one obtained for the LASSO in Theorem 4.2 and in Theorem 5.4, though under a stronger assumption on  $X$ . This is not surprising because when  $\Theta = [0, \infty)^d$ ,  $\Omega(t|\Theta) = \|t\|_1$  and the resulting RERM is just the LASSO.

In the case of the group LASSO, for  $\Theta$  as in (5.13), one has  $M\sqrt{\log(\mathcal{E}x)} \lesssim \sqrt{\log |T|}$ , and  $\Lambda(t^*) \sim \Omega(t^*|\Theta)M\sqrt{(\log |T|)/N}$ .

**Example: The SLOPE regularization**

In Su and Candès (2016); Bogdan et al. (2015), the authors introduced the regularization function:

$$\Psi(t) = \|t\|_{SLOPE} = \sum_{j=1}^d \lambda_j t_j^*$$

where  $\lambda_1 \geq \dots \geq \lambda_d \geq 0$  and  $t_1^* \geq \dots \geq t_d^* \geq 0$  is the non-increasing rearrangement of  $(|t_i|)_{i=1}^d$ .

In Su and Candès (2016) the given data is generated by the Gaussian linear model  $Y_i = \langle X_i, t^* \rangle + W_i, i = 1, \dots, N$  for a Gaussian design  $X_i \sim \mathcal{N}(0, N^{-1}I_{d \times d})$  (note that the covariance matrix is normalized by  $1/N$ ) and a centered Gaussian noise  $W_i$  with variance  $\sigma^2$  that is independent of the design  $X_i$ . Setting  $\Phi^{-1}(\alpha)$  to be the  $\alpha$ -th quantile of a standard normal distribution and  $q \in (0, 1)$ , the weights were chosen to be

$$\lambda_i = \Phi^{-1}(1 - iq/(2d)), \tag{5.16}$$

and, for this choice of weights, SLOPE was defined by

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{2N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + \sigma \frac{\|t\|_{SLOPE}}{\sqrt{N}} \right).$$

The result in Su and Candès (2016) is of an asymptotic nature:

**Theorem 5.10 (Theorem 1.2 Su and Candès (2016))** *Let  $0 < \varepsilon < 1$  and set  $1 \leq k \leq d$  that satisfy  $k/d = o(1)$  and  $(k \log d)/N = o(1)$  when  $N \rightarrow \infty$ . Then,*

$$\lim_{N \rightarrow \infty} \sup_{\|t^*\|_0 \leq k} Pr \left( \frac{N \|\hat{t} - t^*\|_2^2}{2\sigma^2 k \log(d/k)} > 1 + 3\varepsilon \right) = 0,$$

where the supremum is taken with respect to all vectors that are supported on at most  $k$  coordinates.

It was shown in Su and Candès (2016) that  $2\sigma^2 k \log(d/k)/N$  is the (asymptotic) minimax rate for  $t^*$  that is  $k$ -sparse.

The article Su and Candès (2016) (see Section 6 there) raises the question of extending Theorem 5.10 beyond the Gaussian case, especially when the coordinates of  $X$  may be correlated. We study this question in the context of sparse recovery and for an arbitrary choice of weights in Lecué and Mendelson (to appear), leading to error bounds that depend on  $\|t^*\|_0$ . Here, we obtain a complexity-dependent error rate that depends on  $\|t^*\|_{SLOPE}$ .

**Proposition 5.11** *Set  $\mathcal{B} = \{t \in \mathbb{R}^d : \|t\|_{SLOPE} \leq 1\}$ . There exists an absolute constant  $C$ , for which, if  $M = \max_{1 \leq j \leq d} \lambda_j^{-1} \sqrt{\log(ed/j)}$ , then  $\ell^*(\mathcal{B}) \leq CM$ .*

**Proof.** The proof is an outcome of a standard binomial estimate. Let  $G = (g_i)_{i=1}^d$  be a standard Gaussian vector and observe that

$$\ell^*(\mathcal{B}) = \mathbb{E} \sup_{t \in \mathcal{B}} \langle G, t \rangle \leq \mathbb{E} \sup_{t \in \mathcal{B}} \sum_{j=1}^d g_j^* t_j^* \leq \mathbb{E} \sup_{t \in \mathcal{B}} \sum_{j=1}^d \frac{g_j^*}{\lambda_j} \lambda_j t_j^* \leq \mathbb{E} \max_{1 \leq j \leq d} \frac{g_j^*}{\lambda_j}.$$

For  $u \geq 1$ ,

$$\begin{aligned} Pr\left(\max_{1 \leq j \leq d} \frac{g_j^*}{\lambda_j} \geq u\right) &\leq \sum_{j=1}^d Pr(g_j^* \geq u \lambda_j) \leq \sum_{j=1}^d \binom{d}{j} Pr^j(|g| \geq u \lambda_j) \\ &\leq 2 \sum_{j=1}^d \exp\left(j \log\left(\frac{ed}{j}\right) - c_1 j u^2 \lambda_j^2\right) \leq 2 \exp(-c_2 u^2), \end{aligned}$$

where the last inequality follows if one sets  $u^2 \geq \max_j \lambda_j^{-2} \log(ed/j)$ . The proof is concluded by integrating the tails.  $\blacksquare$

Theorem 5.3 leads to estimation properties of SLOPE.

**Theorem 5.12** *Using the notation of Theorem 5.3, if  $\Psi(t) = \|t\|_{SLOPE}$ ,  $\max_j \lambda_j^{-1} \sqrt{\log(ed/j)} \leq C$  and*

$$\hat{t} \in \operatorname{argmin}_{t \in \mathbb{R}^d} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, t \rangle)^2 + c_2 \sigma_q L w u \frac{\|t\|_{SLOPE}}{\sqrt{N}} \right),$$

then with probability larger than the one in (5.6),

$$\|\hat{t} - t^*\|_2^2 \lesssim_{L,q,C} \begin{cases} \frac{\sigma_q w u \|t^*\|_{SLOPE}}{\sqrt{N}} & \text{if } N \gtrsim_L d, \\ \max \left\{ \frac{\sigma_q L w u \|t^*\|_{SLOPE}}{\sqrt{N}}, \frac{\|t^*\|_{SLOPE}^2}{N} \right\} & \text{otherwise.} \end{cases}$$

As in the case of the LASSO, one may combine the sparsity-dependent error rate for SLOPE from [Lecué and Mendelson \(to appear\)](#) and the complexity-dependent error rate from Theorem 5.12. To that end, assume that  $X$  is isotropic,  $L$ -subgaussian and that the noise  $\xi$  is in  $L_q$  for some  $q > 2$ . Then, with probability larger than the one in (5.6)

$$\|\hat{t} - t^*\|_2^2 \lesssim_{L,q,C} \min \left\{ \frac{\sigma_q \|t^*\|_0}{N} \log\left(\frac{ed}{\|t^*\|_0}\right), \max \left\{ \frac{\sigma_q w u \|t^*\|_{SLOPE}}{\sqrt{N}}, \frac{\|t^*\|_{SLOPE}^2}{N} \right\} \right\},$$

for  $N \gtrsim \|t^*\|_0 \log(ed/\|t^*\|_0)$ .

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

In this section, we assume that  $X$  takes values in the set of  $m \times T$  matrices, endowed with the inner product  $\langle A, B \rangle = \sum_{u,v} A_{uv} B_{uv}$ . We consider  $A^* \in \operatorname{argmin}_{A \in \mathbb{R}^{m \times T}} \mathbb{E}(Y - \langle X, A \rangle)^2$  and thus  $\langle X, A^* \rangle$  is the best (linear) approximation of  $Y$  in the  $L_2$  sense.

Let  $\Lambda(\rho) \geq \rho \ell^*(K) / \sqrt{N}$  for all  $\rho > 0$ ,  $u, w > C$  and set

$$\hat{A} \in \operatorname{argmin}_{A \in \mathbb{R}^{m \times T}} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, A \rangle)^2 + c_2 \eta^3 \sigma_q L w u \Lambda(\Psi(A)) \right). \quad (5.17)$$

By Theorem 5.3, with probability larger than the one in (5.6)

$$\left\| \hat{A} - A \right\|_2^2 = \left\| \langle X, \hat{A} - A^* \rangle \right\|_{L_2}^2 \lesssim r(10\eta\Psi(A^*))^2$$

where for  $\rho \geq 0$ ,

$$r(\rho)^2 \sim_{L,q} \begin{cases} \sigma_q w u \Lambda(\rho) & \text{when } N \gtrsim_L mT \\ \max \{ \sigma_q w u \Lambda(\rho), \Lambda^2(\rho) \} & \text{otherwise.} \end{cases}$$

Let us turn to estimates on  $\ell^*(K)$  for the unit balls of the regularization functions used in the *matrix completion* and *collaborative filtering* problems.

**Example:  $S_p$ -regularization for  $p \geq 1$ .**

For any  $A \in \mathbb{R}^{m \times T}$ , let  $\sigma_1(A) \geq \sigma_2(A) \geq \dots \geq \sigma_{m \wedge T}(A)$  be the ordered singular values of  $A$  and set  $m \wedge T = \min\{m, T\}$ . Recall that the  $p$ -Schatten norm  $\|\cdot\|_{S_p}$  of  $A$  is defined by

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

Schatten norms have been used extensively in *matrix completion* and in *collaborative filtering*. Exact reconstruction properties of various procedures have been established via the minimization of the  $S_1$ -norm, constrained to matching the data (see, e.g., Candès and Plan (2011); Candès and Tao (2010); Candès and Recht (2009); Gross (2011); Chandrasekaran et al. (2012)).  $S_1$  regularization has also been used in the noisy setup for independent subgaussian noise and, in most cases, for subgaussian or deterministic designs, in Koltchinskii et al. (2011); Rohde and Tsybakov (2011); Koltchinskii (2011); Negahban and Wainwright (2012); Gaïffas and Lecué (2011); Klopp (2014).

A result that is closely related to ours is Theorem 9.2 from Koltchinskii (2011), in which  $X$  is isotropic and  $L$ -subgaussian;  $\xi$  is a symmetric random variable that is independent of  $X$  and for which  $\|\xi\|_{\psi_\alpha} < \infty$  for some  $\alpha \geq 1$  (cf. Rao and Ren (1991) for more details on the  $\psi_\alpha$ -norms); and the target is  $Y = \langle X, A^* \rangle + \xi$ .

Let  $N \gtrsim m \cdot \operatorname{rank}(A^*)$  and set

$$\lambda \gtrsim \max \left\{ \|\xi\|_2 \sqrt{\frac{m(t + \log m)}{N}}, \|\xi\|_{\psi_\alpha} \log^{1/\alpha} \left( \frac{\|\xi\|_{\psi_\alpha}}{\|\xi\|_{L_2}} \right) \frac{\sqrt{m}(t + \log N)(t + \log m)}{N} \right\}.$$

The  $S_1$ -regularization procedure with regularization parameter  $\lambda$  satisfies that for every  $t > 0$ , with probability larger than  $1 - 3 \exp(-t) - \exp(-c_0 N)$ ,

$$\left\| \hat{A} - A^* \right\|_{S_2}^2 \lesssim \min \left\{ \lambda \|A^*\|_{S_1}, \lambda^2 \text{rank}(A^*) \right\}. \quad (5.18)$$

In comparison, an estimation result for  $S_p$ -norm regularization (for any  $p \geq 1$ ) follows from Theorem 5.3, and does not require any assumptions on the “noise”  $\xi = Y - \langle A^*, X \rangle$ , other than  $\xi \in L_q$  for some  $q > 2$ . In particular,  $\xi$  need not belong to  $\psi_\alpha$ , nor does it have to be independent of  $X$ . The result uses the following estimate on the Gaussian mean-width of the unit ball of  $S_p$ -norms (see, for instance, Proposition 1.4.4 in Chafaï et al. (2012)):

**Proposition 5.13** *Let  $p \geq 1$  and set  $B_p^{mT}$  to be the unit ball of  $\|\cdot\|_{S_p}$ . Then*

$$\ell^*(B_p^{mT}) \sim \min\{m, T\}^{1-1/p} \sqrt{m+T}.$$

Combining the previous result with Theorem 5.3, one obtains the following:

**Theorem 5.14** *Assume that the assumptions of Theorem 5.3 hold. Let  $\Lambda_p(\rho) = \rho \min\{m, T\}^{1-1/p} \sqrt{\frac{m+T}{N}}$  and*

$$\hat{A} \in \underset{A \in \mathbb{R}^{m \times T}}{\text{argmin}} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, A \rangle)^2 + c_2 \sigma_q L w u \Lambda_p(\|A\|_{S_p}) \right).$$

*Then with probability at least  $1 - \delta$ ,*

$$\left\| \hat{A} - A^* \right\|_{S_2}^2 \lesssim_{p,L,q} \begin{cases} \sigma_q w u \Lambda_p(\|A^*\|_{S_p}) & \text{if } N \gtrsim_L mT, \\ \max \left\{ \sigma_q w u \Lambda_p(\|A^*\|_{S_p}), \Lambda_p^2(\|A^*\|_{S_p}) \right\} & \text{otherwise.} \end{cases}$$

**Remark.** As in the vector case mentioned earlier, Theorem 5.3 also applies for  $S_p$ -regularization for  $0 < p < 1$ . In that case, Assumption 1.1 is satisfied for  $\eta = 2^{1/p}$  and the Gaussian mean width of the  $S_p$ -unit ball satisfies  $\ell^*(B_p^{mT}) \lesssim \sqrt{m+T}$ . It therefore follows from Theorem 5.3 that under the same assumptions as in Theorem 5.3 and for  $\Lambda(\rho) = \rho \sqrt{(m+T)/N}$  for all  $\rho > 0$ , the RERM

$$\hat{A} \in \underset{A \in \mathbb{R}^{m \times T}}{\text{argmin}} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, A \rangle)^2 + c_2 \sigma_q L w u \Lambda(\|A\|_{S_p}) \right),$$

satisfies, with probability larger than the one in (5.6),

$$\left\| \hat{A} - A^* \right\|_{S_2}^2 \lesssim_{p,L,q} \begin{cases} \sigma_q w u \Lambda(\|A^*\|_{S_p}) & \text{if } N \gtrsim_L mT, \\ \max \left\{ \sigma_q w u \Lambda(\|A^*\|_{S_p}), \Lambda^2(\|A^*\|_{S_p}) \right\} & \text{otherwise.} \end{cases}$$

Observe that just as in the vector case, when  $0 < p < 1$  this rate is not the minimax rate in the true model  $\|A^*\|_{S_p} B_p^{mT}$ . Indeed, Rohde and Tsybakov (2011) provides the minimax rate, and, in fact, also shows that the minimax rate may be attained using  $\Psi(A) = \|A\|_{S_p}^p$

as a regularization function. To be more accurate, [Rohde and Tsybakov \(2011\)](#) considers the following problem: let  $x_1, \dots, x_N$  be  $N$  deterministic matrices in  $\mathbb{R}^{m \times T}$  satisfying some RIP property and set  $W_i$  to be  $N$  independent, standard Gaussian variables with variance  $\sigma^2$ . Set  $Y_i = \langle x_i, A^* \rangle + W_i, i = 1, \dots, N$ , leading to the so-called matrix regression model with Gaussian noise and a deterministic design. It is shown in [Rohde and Tsybakov \(2011\)](#) that when  $\rho B_p^{mT}$  for some  $0 < p \leq 1$ , the minimax rate of the problem in  $\rho B_p^{mT}$  is

$$\sigma^2 \rho^p \left( \frac{m+T}{N} \right)^{1-\frac{p}{2}}$$

in some specific range of  $N, \sigma$  and  $\rho$ . Our result recovers this rate only for  $p = 1$ .

**Example: Max-norm regularization.** The max-norm of a matrix is defined by

$$\|A\|_{max} = \inf_{A=UV^\top} \|U\|_{2 \rightarrow \infty} \|V\|_{2 \rightarrow \infty},$$

with the infimum taken with respect to all pairs of matrices  $U, V$  for which  $A = UV^\top$ .

Constrained empirical risk minimization procedures that are based on the max-norm have been used in [Cai and Zhou \(2016\)](#); [Nathan and Adi \(2005\)](#) for bounded and Lipschitz loss functions, and in [Lecué and Mendelson](#) for the squared loss and for a subgaussian and isotropic design vector  $X$  and a subgaussian noise  $\xi$  independent of  $X$ . Consider the matrix regression model  $Y_i = \langle X_i, A^* \rangle + W_i, i = 1, \dots, N$  where  $X_1, \dots, X_N$  are independent isotropic and subgaussian matrices,  $W_1, \dots, W_N$  are independent centered gaussian variables with variance  $\sigma^2$  that are independent of the  $X_i$ 's and  $A^*$  belongs to the max-norm ball of radius  $\rho$ . One may show that the minimax rate in this case is

$$\max \left\{ \sigma \rho \sqrt{\frac{(mT)(m+T)}{N}}, \frac{\rho^2(mt)(m+T)}{N} \right\}, \quad (5.19)$$

at least in some specific regime of  $\rho, \sigma$  and  $N$  (cf. [Lecué and Mendelson](#)).

To apply Theorem 5.3, let us estimate the Gaussian mean-width of the unit ball of the max-norm ball, that is, of  $\mathcal{B} = \{A \in \mathbb{R}^{m \times T} : \|A\|_{max} \leq 1\}$ .

**Lemma 5.15** *There exists an absolute constant  $c$  for which, for every  $m$  and  $T$ ,*

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

**Proof.** An application of Grothendieck's inequality (see, e.g., [Nathan and Adi \(2005\)](#)) shows that

$$\text{conv}(\mathcal{X}_\pm) \subset \mathcal{B} \subset K_G \text{conv}(\mathcal{X}_\pm)$$

where  $K_G$  is the Grothendieck constant and  $\mathcal{X}_\pm = \{uv^\top : u \in \{\pm 1\}^m, v \in \{\pm 1\}^T\}$ . If  $\mathfrak{G} = (g_{ij})_{1 \leq u \leq m, 1 \leq v \leq T}$  is a standard  $m \times T$  Gaussian matrix, it follows from the Gaussian maximal inequality (see, e.g., Chapter 3 in [Ledoux and Talagrand \(1991\)](#)) that

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

■

**Theorem 5.16** *Using the assumptions and notation of Theorem 5.3, and setting  $\Lambda(\rho) = \rho\sqrt{(mT)(m+T)/N}$ , if*

$$\hat{A} \in \operatorname{argmin}_{A \in \mathbb{R}^{m \times T}} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, A \rangle)^2 + c_2 \sigma_q L w u \Lambda(\|A\|_{\max}) \right),$$

then with probability larger than the one in (5.6)

$$\|\hat{A} - A^*\|_{S_2}^2 \lesssim_{L,q} \begin{cases} \sigma_q w u \Lambda(\|A^*\|_{\max}) & \text{if } N \gtrsim_L mT, \\ \max \left\{ \sigma_q w u \Lambda(\|A^*\|_{\max}), \Lambda^2(\|A^*\|_{\max}) \right\} & \text{otherwise.} \end{cases}$$

As a consequence, we recover the minimax rate of convergence in the matrix regression model with subgaussian design and gaussian noise in the class  $\|A^*\|_{\max} \mathcal{B}$  thanks to max-norm regularization and without knowing  $\|A^*\|_{\max}$  in advance.

**Example: Atomic-norm regularization.**

The atomic-norm has been used in [Chandrasekaran et al. \(2012\)](#) in the context of exact and robust recovery using few Gaussian linear measurements of a signal or of a matrix.

Given  $\mathcal{A} \subset \mathbb{R}^{m \times T}$ , the elements in  $\mathcal{A}$  are called *atoms*. Set  $\operatorname{conv}(\mathcal{A})$  to be the convex hull of  $\mathcal{A}$  and put

$$\|A\|_{\mathcal{A}} = \inf \{t > 0 : A \in t \cdot \operatorname{conv}(\mathcal{A})\}. \quad (5.20)$$

Even though  $\|\cdot\|_{\mathcal{A}}$  need not be a norm (because  $\operatorname{conv}(\mathcal{A})$  need not be centrally-symmetric), it is positive homogeneous and satisfies a triangle inequality: for every  $A, B \in \mathbb{R}^{m \times T}$  and  $\lambda \geq 0$ :

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

And, if we assume that  $\mathcal{A}$  is centrally-symmetric, then  $\|\cdot\|_{\mathcal{A}}$  is a norm, (5.9) is satisfied and Theorem 5.3 applies.

Set  $\mathcal{B}$  to be the unit ball with respect to  $\|\cdot\|_{\mathcal{A}}$  and note that  $\ell^*(\mathcal{B}) = \ell^*(\mathcal{A})$ . For example, assume that  $m = T$  and put  $\mathcal{A}$  to be the set of all orthogonal matrices. Since the unit ball of the spectral norm is the convex hull of the set of orthogonal matrices, one has  $\|\cdot\|_{\mathcal{A}} = \|\cdot\|_{S_2}$  and

$$\ell^*(\mathcal{B}) = \mathbb{E} \|\mathfrak{G}\|_{S_2} \leq \sqrt{m} \mathbb{E} \|\mathfrak{G}\|_{S_\infty} \lesssim m.$$

**Theorem 5.17** *Using the assumptions and notation of Theorem 5.3, let  $\mathcal{A} \subset \mathbb{R}^{m \times T}$  be a centrally-symmetric set of atoms and set  $\Lambda(\rho) \geq \rho \ell^*(\mathcal{A})/\sqrt{N}$ . If*

$$\hat{A} \in \operatorname{argmin}_{A \in \mathbb{R}^{m \times T}} \left( \frac{1}{N} \sum_{i=1}^N (Y_i - \langle X_i, A \rangle)^2 + c_2 \sigma_q L w u \Lambda(\|A\|_{\mathcal{A}}) \right)$$

then with probability larger than the one in (5.6)

$$\|\hat{A} - A^*\|_{S_2}^2 \lesssim_{L,q} \begin{cases} \sigma_q w u \Lambda(\|A^*\|_{\mathcal{A}}) & \text{if } N \gtrsim_L mT, \\ \max \left\{ \sigma_q w u \Lambda(\|A^*\|_{\mathcal{A}}), \Lambda^2(\|A^*\|_{\mathcal{A}}) \right\} & \text{otherwise.} \end{cases}$$

## 6. Conclusions

We have presented a general result (Theorem 1.9) describing statistical properties of a constrained regularized procedure in the learning theoretical framework. This result highlights the role played by the quadratic and multiplier processes in calibrating the regularization parameter  $\lambda$  as well as their effect on the estimation error rate. It appears that:

1. The rates of convergence depend on  $\Psi(f^*)$  and we recover the minimax rate in the true model  $\{f \in F : \Psi(f) \leq \Psi(f^*)\}$  – up to a logarithmic factor – in many well-studied examples .
2. No statistical model is needed to study RERM; all the analysis has been carried out in the general learning theory setup, and thus without assuming any statistical model. Theorem 1.9 and all its corollaries and applications are true regardless of any relation between the target  $Y$  and the input  $X$ . For instance, when predicting  $Y$  using linear functionals there is no need to assume that  $Y = \langle t_0, X \rangle + W$  for an independent noise  $W$ ; our results hold even if  $Y$  were, for instance, a noisy version of a quadratic function of a linear functional of  $X$  (e.g. phase retrieval).

Our analysis shows that despite considering this general framework, the error rate and the regularization parameter used to construct RERM almost match the ones that would have been obtained with more information – namely, a given statistical model. In the examples we considered, statistical models are superfluous to the analysis of RERM and as a consequence, they may actually hide what really determines the error rate and the right choice of a regularization parameter:

- calibration of the regularization parameter depends only on the multiplier process – which measures the empirical correlation between the noise  $Y - f^*(X)$  and the class  $F$ . When this correlation is small or even null (in the free-noise case) the regularization parameter will also be small.
- the key parameters are the structure of the “unit ball” of the regularization function (measured here using the Gaussian mean width) and the noise level, which is measured through the  $L_q$  norm of  $Y - f^*(X)$ .

## 7. Proof of Theorem 3.6

Following Mendelson (2016), the proof of Theorem 3.6 is based on properties of the following norm:

**Definition 7.1** For a random variable  $Z$  and  $p \geq 1$ , set

$$\|Z\|_{(p)} = \sup_{1 \leq q \leq p} \frac{\|Z\|_{L_q}}{\sqrt{q}}.$$

The  $\|\cdot\|_{(p)}$  norm is a ‘local’ version of the  $\psi_2$  norm. While

$$\|Z\|_{\psi_2} \sim \sup_{q \geq 1} \frac{\|Z\|_{L_q}}{\sqrt{q}},$$

$\|Z\|_{(p)}$  captures the subgaussian behavior of  $Z$  up to the  $p$ -th moment.

Under Assumption 3.1, a high probability bound on (3.7) can be derived using the next result.

**Proposition 7.2 (Lemma 2.8 in Lecué and Mendelson (2017))** *There exists an absolute constant  $c_0$  for which the following holds. Let  $Z$  be a mean-zero real-valued random variable and let  $Z_1, \dots, Z_N$  be independent copies of  $Z$ . Let  $p_1 \geq 1$  and assume that  $\|Z\|_{(p_1)} \leq L$ , then*

$$\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N Z_i \right\|_{(p_1)} \leq c_0 L.$$

Setting  $U_j = N^{-1/2} \sum_{i=1}^N \varepsilon_i X_i(j)$  and  $p_1 = \log d$  (recalling that  $t \geq 1$  and  $d \geq N$  in Assumption 3.1), it follows from Proposition 7.2 that

$$\|U_j\|_{L_{p_1}} \leq c_0 L \sqrt{p_1} \|x_j\|_{L_2}.$$

Therefore,

$$\begin{aligned} Pr \left( \max_{1 \leq j \leq d} |U_j| \geq u \right) &\leq \sum_{j=1}^d Pr (|U_j| \geq u) \leq \sum_{j=1}^d \left( \frac{\|U_j\|_{L_{p_1}}}{u} \right)^{p_1} \\ &\leq d \left( \frac{c_0 L \sqrt{p_1} \max_{1 \leq j \leq d} \|x_j\|_{L_2}}{u} \right)^{p_1} = d \left( \frac{c_0 L \sqrt{\log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}}{u} \right)^{\log d}. \end{aligned}$$

Let  $w \geq e$  and set  $u = c_0 L w \sqrt{\log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}$ ; therefore,

$$Pr \left( \max_{1 \leq j \leq d} |U_j| \geq c_1 L w \sqrt{\log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2} \right) \leq \left( \frac{e}{w} \right)^{\log d}, \quad (7.1)$$

which is a high probability estimate on (3.7) under a limited moment assumption. Integrating the tail,

$$\mathbb{E} \max_{1 \leq j \leq d} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i(j) \right| \lesssim L \sqrt{\log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}$$

proving (3.9).

Next, we obtain high probability bounds on (3.8),—estimates that require some preparation.

Let  $j \in \{1, \dots, d\}$  and set  $Z_i = X_i(j)$ . Consider the Bernoulli sums

$$Q_j = \sum_{i=1}^N \varepsilon_i \xi_i X_i(j) = \sum_{i=1}^N \varepsilon_i \xi_i Z_i.$$

Denote by  $(a_i^*)_{i=1}^N$  the non-increasing rearrangement of  $(|a_i|)_{i=1}^N$ . A straightforward application of Höfdding's inequality shows that conditioned on  $(\xi_i)_{i=1}^N$  and  $(Z_i)_{i=1}^N$ , for any  $v > 0$ ,



with probability at least  $1 - 2 \exp(-v^2/2)$  relative to  $(\varepsilon_i)_{i=1}^N$ ,

$$\begin{aligned} |Q_j| &\leq \sum_{i \leq m} \xi_i^* Z_i^* + v \left( \sum_{i \geq m} (\xi_i^* Z_i^*)^2 \right)^{1/2} \\ &\leq \left( \sum_{i \leq m} (\xi_i^*)^2 \right)^{1/2} \left( \sum_{i \leq m} (Z_i^*)^2 \right)^{1/2} + v \left( \sum_{i \geq m} (\xi_i^*)^{2r} \right)^{1/2r} \left( \sum_{i \geq m} (Z_i^*)^{2r'} \right)^{1/2r'}, \end{aligned} \quad (7.2)$$

where  $r$  and  $r'$  are conjugate indices.

As a consequence, high probability bounds on the rearrangements  $(\xi_i^*)$  and  $(Z_i^*)$  can be used to obtain high probability bounds on  $|Q_j|$  (and therefore, on  $\max_{1 \leq j \leq d} |Q_j|$  as well, using the union bound).

The next two observations, whose proofs may be found in [Mendelson \(2016\)](#), give information on the structure of a typical  $(Z_i)_{i=1}^N$  when  $Z$  has at least  $\sim \log d$  subgaussian moments. It turns out that one may decompose  $(Z_i)_{i=1}^N$  to a sum of two vectors, supported on disjoint sets: one consists of the largest  $m$  coordinates of  $(|Z_i|)_{i=1}^N$ , and its  $\ell_2^N$  norm is determined by relatively high moments of  $Z$ ; the other consists of the  $N - m$  smaller coordinates of  $(|Z_i|)_{i=1}^N$ , and if  $Z \in L_{q_1}$ , its  $\ell_r^N$  norm is well-behaved for  $r < q_1$ .

The ‘level’  $m$  depends on the desired probability estimate and on the moments of  $Z$ : if one wishes to obtain a probability estimate of  $1 - 2 \exp(-p)$  for  $p \geq \log N$  (as we will), then  $Z$  should have roughly  $p$  moments and one should select  $m \sim p / \log(eN/p)$ .

First, let us consider the smaller coordinates:

**Lemma 7.3 (Lemma 3.2 in [Mendelson \(2016\)](#))** *There exist absolute constants  $a_0$  and  $c_1$  for which the following holds. Let  $1 \leq r_1 < q_1$ , set  $Z \in L_{q_1}$  and put  $Z_1, \dots, Z_N$  to be independent copies of  $Z$ . Fix  $1 \leq p \leq N$ , let  $u > 2$  and set*

$$m = \left\lceil \frac{a_0 p}{((q_1/r_1) - 1) \log(4 + eN/p)} \right\rceil.$$

If  $m > 1$ , then, with probability at least  $1 - 2u^{-mq_1} \exp(-p)$ ,

$$\left( \sum_{i=m}^N (Z_i^*)^{r_1} \right)^{1/r_1} \leq c_1 \left( \frac{q_1}{q_1 - r_1} \right)^{1/r_1} u N^{1/r_1} \|Z\|_{L_{q_1}}$$

and, if  $m = 1$  and  $0 < \beta < q_1/r_1 - 1$  then with probability at least  $1 - c_2 u^{-q_1} N^{-\beta}$ ,

$$\left( \sum_{i=1}^N |Z_i|^{r_1} \right)^{1/r_1} \leq c_1 \left( \frac{q_1}{q_1 - (\beta + 1)r_1} \right) u \|Z\|_{L_{q_1}} N^{1/r_1}.$$

Next, we consider the larger coordinates:

**Lemma 7.4 (Lemma 3.4 in [Mendelson \(2016\)](#))** *There exists absolute constants  $a_1$  and  $c_0$  for which the following holds. Let  $Z_1, \dots, Z_N$  be independent copies of a random variable*

$Z$ , set  $p \geq \log N$  and put  $1 \leq m \leq N/2e$  that satisfy  $m \leq a_1 p / \log(eN/p)$ . Then, for every  $u > 1$ , with probability at least  $1 - u^{-2p} \exp(-p)$ , one has

$$\left( \sum_{i=1}^m (Z_i^*)^2 \right)^{1/2} \leq c_0 u \sqrt{p} \|Z\|_{(2p)}.$$

In particular, under Assumption 3.1, we apply Lemma 7.3 and Lemma 7.4 to  $p, q_1$  and  $r_1$  defined by

$$2p = t \log d, r_1 = 2r' \text{ and } q_1 = r_1 \max \left\{ 2, 1 + \frac{a_0}{a_1} \right\},$$

where  $a_0$  and  $a_1$  are the absolute constants from Lemma 7.3 and 7.4. We also set

$$m = \left\lceil \frac{a_0 p}{\log(4 + eN/p)} \right\rceil \tag{7.3}$$

and observe that if

$$m_0 = \left\lceil \frac{a_0 p}{((q_1/r_1) - 1) \log(4 + eN/p)} \right\rceil \text{ then } m_0 \leq m \leq \frac{a_1 p}{\log(eN/p)}$$

and  $m_0 q_1 \sim m$ . Moreover, if  $\kappa_0$  is a large enough absolute constant and  $t \geq \kappa_0$ , then  $m_0 > 1$ . Recalling that  $p \geq 2 \log d$  and setting  $Z_i = X_i(j)$  for  $i = 1, \dots, N$ , it follows that for any  $u > 2$ , with probability at least  $1 - u^{-2p} \exp(-p/2) - 2u^{-c_0 m} \exp(-p/2)$ , for every  $1 \leq j \leq d$

$$\left( \sum_{i=1}^m (Z_i^*)^2 \right)^{1/2} \lesssim u \sqrt{p} \|Z\|_{(2p)} \lesssim u L \sqrt{t \log d} \|x_j\|_{L_2} \tag{7.4}$$

and

$$\left( \sum_{i=m}^N (Z_i^*)^{2r'} \right)^{1/2r'} \lesssim u \|Z\|_{L_{q_1}} N^{1/2r'} \lesssim u L \sqrt{r'} \|x_j\|_{L_2} N^{1/2r'}. \tag{7.5}$$

Let  $\xi_1, \dots, \xi_N$  be independent copies of  $\xi$  and recall that  $(\xi_i^*)_{i=1}^N$  is the monotone non-increasing rearrangement of  $(|\xi_i|)_{i=1}^N$ . We apply Lemma 7.3 for  $q_1 = q, r_1 = 2r$  and set

$$m_1 = \left\lceil \frac{a_0 p}{((q_1/r_1) - 1) \log(4 + eN/p)} \right\rceil.$$

Thus,  $m_1 > 1$  when  $t \geq \kappa_0$  for a large enough constant  $\kappa_0$ , and if  $m$  is as in (7.3) one has  $m \geq m_1$  and  $m_1 q_1 \sim m$ . Hence, for  $p = (t/2) \log d$ , with probability larger than  $1 - 2u^{-c_0 m} \exp(-p)$ ,

$$\left( \sum_{i=m}^N (\xi_i^*)^{2r} \right)^{1/2r} \leq c(q) u \|\xi\|_{L_q} N^{1/2r}. \tag{7.6}$$

This provides a high probability bound on the smaller coordinates of  $(|\xi_i|)_{i=1}^N$ . Next, let us turn to estimating the larger ones.

**Lemma 7.5 (Lemma 4.3 in Mendelson (2016))** *Let  $q > 2$  and assume that  $\xi \in L_q$ . If  $\xi_1, \dots, \xi_N$  are independent copies of  $\xi$  then for every  $w > 1$  with probability larger than  $1 - c_0 w^{-q} N^{-((q/2)-1) \log^q N}$ ,*

$$\left( \sum_{i=1}^m \xi_i^2 \right)^{1/2} \leq \left( \sum_{i=1}^N \xi_i^2 \right)^{1/2} \leq c_1 w \|\xi\|_{L_q} \sqrt{N}.$$

Setting  $Z_i = X_i(j)$  for  $i = 1, \dots, N$  and applying (7.4), (7.5), (7.6) and Lemma 7.5, we obtain, with probability larger than

$$1 - \frac{\exp(-p/2)}{u^{2p}} - \frac{4 \exp(-p/2)}{u^{c_0 m}} - \frac{c_0 \log^q N}{w^q N^{q/2-1}}$$

that for every  $j = 1, \dots, d$

$$\left( \sum_{i \leq m} (\xi_i^*)^2 \right)^{1/2} \left( \sum_{i \leq m} (Z_i^*)^2 \right)^{1/2} \lesssim u w L \sqrt{t \log d} \|x_j\|_{L_2} \|\xi\|_{L_q} \sqrt{N}$$

and

$$\left( \sum_{i \geq m} (\xi_i^*)^{2r} \right)^{1/2r} \left( \sum_{i \geq m} (Z_i^*)^{2r'} \right)^{1/2r'} \leq c(q) u^2 L \|\xi\|_{L_q} \sqrt{N} \|x_j\|_{L_2}.$$

Then, by plugging those inequalities in (7.2), it is evident that under Assumption 3.1, for  $u > 2, v > 0, w \geq 2, 2p = t \log d$  and  $m \sim p / \log(eN/p)$ , with probability at least

$$1 - \frac{\exp(-p/2)}{u^{2p}} - \frac{4 \exp(-p/2)}{u^{c_0 m}} - \frac{c_1 \log^q N}{w^q N^{q/2-1}} - 2 \exp(-v^2 t \log d),$$

$$\max_{1 \leq j \leq d} \left| \sum_{i=1}^N \varepsilon_i \xi_i X_i(j) \right| = \max_{1 \leq j \leq d} |Q_j| \lesssim_q (u w + u^2 v) L \|\xi\|_{L_q} \sqrt{N} \sqrt{t \log d} \max_{1 \leq j \leq d} \|x_j\|_{L_2}.$$

## Acknowledgments

We would like to acknowledge support for this project from Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and the Israel Science Foundation, grant 707/14

## References

Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000074. URL <http://dx.doi.org/10.1214/009053606000000074>.

- Peter L. Bartlett, Shahar Mendelson, and Joseph Neeman.  $\ell_1$ -regularized linear regression: persistence and oracle inequalities. *Probab. Theory Related Fields*, 154(1-2):193–224, 2012. ISSN 0178-8051. doi: 10.1007/s00440-011-0367-2. URL <http://dx.doi.org/10.1007/s00440-011-0367-2>.
- Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS620. URL <http://dx.doi.org/10.1214/08-AOS620>.
- Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140, 2015. ISSN 1932-6157. doi: 10.1214/15-AOAS842. URL <http://dx.doi.org/10.1214/15-AOAS842>.
- Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2. doi: 10.1007/978-3-642-20192-9. URL <http://dx.doi.org/10.1007/978-3-642-20192-9>. Methods, theory and applications.
- T. Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electron. J. Stat.*, 10(1):1493–1525, 2016. ISSN 1935-7524. doi: 10.1214/16-EJS1147. URL <http://dx.doi.org/10.1214/16-EJS1147>.
- 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. ISSN 0018-9448. doi: 10.1109/TIT.2011.2111771. URL <http://dx.doi.org/10.1109/TIT.2011.2111771>.
- Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009. ISSN 1615-3375. doi: 10.1007/s10208-009-9045-5. URL <http://dx.doi.org/10.1007/s10208-009-9045-5>.
- Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2044061. URL <http://dx.doi.org/10.1109/TIT.2010.2044061>.
- Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006. ISSN 0018-9448. doi: 10.1109/TIT.2005.862083. URL <http://dx.doi.org/10.1109/TIT.2005.862083>.
- Djalil Chafaï, Olivier Guédon, Guillaume Lécué, 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. ISBN 978-2-85629-370-6.
- 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.

- ISSN 1615-3375. doi: 10.1007/s10208-012-9135-7. URL <http://dx.doi.org/10.1007/s10208-012-9135-7>.
- Víctor H. de la Peña and Evarist Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. ISBN 0-387-98616-2. doi: 10.1007/978-1-4612-0537-1. URL <http://dx.doi.org/10.1007/978-1-4612-0537-1>. From dependence to independence, Randomly stopped processes.  $U$ -statistics and processes. Martingales and beyond.
- Sjoerd Dirksen, Guillaume Lecué, and Holger Rauhut. On the gap between rip-properties and sparse recovery conditions. Technical report, 2014. To appear in IEEE Transactions on Information Theory.
- 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. ISSN 0178-8051. doi: 10.1007/BF01199026. URL <http://dx.doi.org/10.1007/BF01199026>.
- D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996. ISBN 0-521-56036-5. doi: 10.1017/CBO9780511662201. URL <http://dx.doi.org/10.1017/CBO9780511662201>.
- Stéphane Gaïffas and Guillaume Lecué. Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Trans. Inform. Theory*, 57(10):6942–6957, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2136318. URL <http://dx.doi.org/10.1109/TIT.2011.2136318>.
- Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. ISSN 0091-1798. With discussion.
- Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015. ISBN 978-1-4822-3794-8.
- F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7:219–269, 1995.
- Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. ISSN 0040-1706. doi: 10.2307/1268518. URL <http://dx.doi.org/10.2307/1268518>.
- 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. ISSN 0021-9045. doi: 10.1016/j.jat.2007.04.007. URL <http://dx.doi.org/10.1016/j.jat.2007.04.007>.
- David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2104999. URL <http://dx.doi.org/10.1109/TIT.2011.2104999>.

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL <http://dx.doi.org/10.1007/978-0-387-84858-7>. Data mining, inference, and prediction.
- Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014. ISSN 1350-7265. doi: 10.3150/12-BEJ486. URL <http://dx.doi.org/10.3150/12-BEJ486>.
- Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. ISBN 978-3-642-22146-0. doi: 10.1007/978-3-642-22147-7. URL <http://dx.doi.org/10.1007/978-3-642-22147-7>. 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].
- Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN*, (23):12991–13008, 2015. ISSN 1073-7928. doi: 10.1093/imrn/rnv096. URL <http://dx.doi.org/10.1093/imrn/rnv096>.
- 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. ISSN 0090-5364. doi: 10.1214/11-AOS894. URL <http://dx.doi.org/10.1214/11-AOS894>.
- Guillaume Lécué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. In *Topics in Learning Theory*. Société Mathématique de France, (S. Boucheron and N. Vayatis Eds.). To appear.
- Guillaume Lécué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016. ISSN 1350-7265. doi: 10.3150/15-BEJ701. URL <http://dx.doi.org/10.3150/15-BEJ701>.
- Guillaume Lécué and Shahar Mendelson. Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc. (JEMS)*, 19(3):881–904, 2017. ISSN 1435-9855. URL <https://doi.org/10.4171/JEMS/682>.
- Guillaume Lécué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. *Ann. Statist.*, to appear.
- 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. ISBN 3-540-52013-9. Isoperimetry and processes.
- Pascal Massart and Caroline Meynet. The Lasso as an  $\ell_1$ -ball model selection procedure. *Electron. J. Stat.*, 5:669–687, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS623. URL <http://dx.doi.org/10.1214/11-EJS623>.

- Andreas Maurer and Massimiliano Pontil. Structured sparsity and generalization. *J. Mach. Learn. Res.*, 13:671–690, 2012. ISSN 1532-4435.
- Andreas Maurer, Charles Micchelli, and Massimiliano Pontil. A family of penalty functions for structured sparsity. *NIPS*, 2010.
- Shahar Mendelson. A remark on the diameter of random sections of convex bodies. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 395–404. Springer, Cham, 2014. doi: 10.1007/978-3-319-09477-9\_25. URL [http://dx.doi.org/10.1007/978-3-319-09477-9\\_25](http://dx.doi.org/10.1007/978-3-319-09477-9_25).
- Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015. ISSN 0004-5411. doi: 10.1145/2699439. URL <http://dx.doi.org/10.1145/2699439>.
- Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.*, 126(12):3652–3680, 2016.
- Shahar Mendelson. Learning without concentration for a general loss function. *Probab. Theory Related Fields*, To appear.
- Srebro Nathan and Shraibman Adi. Rank, trace-norm and max-norm. *18th Annual Conference on Learning Theory (COLT)*, 2005.
- 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. ISSN 1532-4435.
- 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. ISBN 0-8247-8478-2.
- Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2165799. URL <http://dx.doi.org/10.1109/TIT.2011.2165799>.
- Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011. ISSN 0090-5364. doi: 10.1214/10-AOS854. URL <http://dx.doi.org/10.1214/10-AOS854>.
- Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011. ISSN 0090-5364. doi: 10.1214/10-AOS860. URL <http://dx.doi.org/10.1214/10-AOS860>.
- Mark Rudelson and Roman Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *Int. Math. Res. Not. IMRN*, (19):9594–9617, 2015. ISSN 1073-7928. doi: 10.1093/imrn/rnu243. URL <http://dx.doi.org/10.1093/imrn/rnu243>.
- Weijie Su and Emmanuel J. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068, 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1397. URL <http://dx.doi.org/10.1214/15-AOS1397>.



- Michel Talagrand. The supremum of some canonical processes. *Amer. J. Math.*, 116(2): 283–325, 1994. ISSN 0002-9327. doi: 10.2307/2374931. URL <http://dx.doi.org/10.2307/2374931>.
- Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. ISBN 3-540-24518-9. Upper and lower bounds of stochastic processes.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246. URL [http://links.jstor.org/sici?sici=0035-9246\(1996\)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN](http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN).
- A. N. Tikhonov. On the stability of inverse problems. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 39:176–179, 1943.
- Sara A. van de Geer. The deterministic lasso. 2007. In JSM proceedings. American Statistical Association.
- Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. With applications to statistics.
- Nicolas Verzelen. Minimax risks for sparse regressions: ultra-high dimensional phenomena. *Electron. J. Stat.*, 6:38–90, 2012. ISSN 1935-7524. doi: 10.1214/12-EJS666. URL <http://dx.doi.org/10.1214/12-EJS666>.
- Zhan Wang, Sandra Paterlini, Fuchang Gao, and Yuhong Yang. Adaptive minimax regression estimation over sparse  $\ell_q$ -hulls. *J. Mach. Learn. Res.*, 15:1675–1711, 2014. ISSN 1532-4435.
- 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. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00532.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x>.