Towards the study of least squares estimators with convex penalty

Pierre C. Bellec[2], Guillaume Lecué[1], Alexandre B. Tsybakov[1]

[1] CREST-ENSAE, CNRS UMR9194
[2] Rutgers University

Abstract

Penalized least squares estimation is a popular technique in high-dimensional statistics. It includes such methods as the LASSO, the group LASSO, and the nuclear norm penalized least squares. The existing theory of these methods is not fully satisfying since it allows one to prove oracle inequalities with fixed high probability only for the estimators depending on this probability. Furthermore, the control of compatibility factors appearing in the oracle bounds is often not explicit. Some very recent developments suggest that the theory of oracle inequalities can be revised in an improved way. In this paper, we provide an overview of ideas and tools leading to such an improved theory. We show that, along with overcoming the disadvantages mentioned above, the methodology extends to the hilbertian framework and it applies to a large class of convex penalties. This paper is partly expository. In particular, we provide adapted proofs of some results from other recent work.

1 Introduction

Penalized least squares (LS) estimators play an important role in statistics. One of the classical examples is ridge regression estimator, for which the penalty is defined as the squared Euclidean norm. More recently, a great deal of attention has been focused on high-dimensional statistical models. In this context, some new penalized LS estimators have been proposed and extensively studied. The most famous examples are the LASSO (i.e., the ℓ_1 norm penalized estimator) and its generalizations such as the group LASSO or the nuclear norm penalized least squares for matrix estimation. A common feature of these and related estimators is the fact that the penalty is a norm satisfying some specific decomposability conditions. Starting from [4], there has been a considerable interest in developing a general approach to the analysis of such methods. For a detailed account, we refer the reader to [12, 11, 28] where further references can be found. As shown in [4], the two main ingredients of the analysis are geometric considerations based on the restricted eigenvalue (compatibility) property, and the empirical process bounds on the stochastic error. With this approach, several techniques have been proposed for a unified treatment of LS estimators with decomposable penalties, see the overviews in [22, 27, 28].

However, the existing theory is not fully satisfying in the following aspects.

(i) The results are obtained in the form of oracle inequalities depending on the restricted eigenvalue (compatibility) parameters that are, in general, not specified. An exception is the standard LASSO, for which the values of these parameters are evaluated in some situations.

(ii) The penalties (and thus, the estimators) considered in that theory depend on the confidence level (the probability), with which the oracle inequality holds. In other words, there is no means, in that framework, to provide oracle bounds for one given penalized LS estimator with any given confidence level. As a

consequence, oracle inequalities for the mean squared risk or upper bounds on any other moments of the risk are not derivable from these results.

Very recent developments show that, in some cases, the problems (i) and (ii) can be resolved. For (i), a relatively general solution can be obtained by the small ball method of [14, 20]. It has been already successfully implemented for such procedures as LASSO and SLOPE [9, 15, 2].

Techniques to overcome the disadvantage (ii) have been recently proposed in [2, 3]. The argument in [2] is based on a refined bound for the stochastic error, and the results focus on the LASSO and SLOPE estimators. Thanks to these techniques, an improvement of the classical rates is obtained for the prediction and estimation errors. In [3], a different argument is used to resolve the problem described in (ii). The results are valid only for the prediction error but extend to other penalized LS estimators than LASSO and SLOPE. The proof is based on a Lipschitz property of the solutions and the Gaussian concentration theorem.

In view of these developments, the theory of oracle inequalities for penalized LS estimators can be revised in an improved way. In this paper, we provide an overview of ideas and tools leading to such an improved theory. Along with overcoming the disadvantages mentioned in (i) and (ii), the method extends to the hilbertian framework and applies to a large class of convex penalties. The approach is based on a refinement of the argument in [2]. This paper is partly expository. In particular, we provide adapted proofs of some results from the previous work.

2 Statement of the problem

Assume that we observe the vector

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\xi},$$

where $\mathbf{f} \in \mathbb{R}^n$ is an unknown deterministic mean and $\boldsymbol{\xi} \in \mathbb{R}^n$ is a random noise vector. Let $\sigma > 0$. We assume that $\boldsymbol{\xi}$ has normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$, where $I_{n\times n}$ denotes the $n \times n$ identity matrix.

For all $\boldsymbol{u} = (u_1, \dots, u_n) \in \mathbb{R}^n$, define the empirical norm of $\boldsymbol{u}$ by

$$\|\boldsymbol{u}\|_n^2 = \frac{1}{n}\sum_{i=1}^n u_i^2.$$

Let H be a Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and the corresponding norm $\|\cdot\|_H$. Let $\mathbb{B}$ a convex subset of H such that $\mathbb{B}$ is a closed set with respect to $\|\cdot\|_H$. We study the performance of the estimator $\hat{\boldsymbol{\beta}}$ defined as a solution of the following minimization problem:

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{B}} \Big(\|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2 + F(\boldsymbol{\beta}) \Big) \tag{2.1}$$

where $\mathbb{X} : H \to \mathbb{R}^n$ is a linear operator and $F : H \to \mathbb{R}$ is a convex function called a penalty. Our main results will be given for the case when $F(\boldsymbol{\beta})$ is some norm of $\boldsymbol{\beta}$. The value $\mathbb{X}\hat{\boldsymbol{\beta}}$ is used as a prediction for $\mathbf{f}$. If the model is well-specified, that is $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{B}$, then $\hat{\boldsymbol{\beta}}$ is used as an estimator of $\boldsymbol{\beta}^*$.

3 Basic tools

In this section, we provide two basic facts that are used in the subsequent argument. The first of them is the following proposition generalizing [2, Proposition E.3] that plays a role of a "reduction lemma" for the stochastic error term. It is crucial to overcome the disadvantage (ii) mentioned in the Introduction.

A mapping $h : H \to [0, \infty)$ will be called positive homogeneous if $h(a\boldsymbol{u}) = ah(\boldsymbol{u})$ for all $a \geq 0, \boldsymbol{u} \in H$ and $h(\boldsymbol{u}) > 0$ for $\boldsymbol{u} \neq \mathbf{0}$. Denote by $\Phi(\cdot)$ the cumulative distribution function of the standard Gaussian law.

Proposition 3.1. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$. Let $h : H \to [0, +\infty)$ be a positive homogeneous mapping and let $\tau > 0$. Assume that the event*

$$\Omega \triangleq \left\{ \sup_{\boldsymbol{v}\in H: h(\boldsymbol{v})\leq 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{v} \leq \tau \right\}$$

satisfies $\mathbb{P}(\Omega) \geq 1/2$. Then for all $\delta \in (0, 1)$ we have

$$\mathbb{P}\left(\forall \boldsymbol{u} \in H : \quad \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u} \leq (\tau + 1) \max\left(h(\boldsymbol{u}), \sigma \frac{\Phi^{-1}(1-\delta)}{\sqrt{n}} \|\mathbb{X}\boldsymbol{u}\|_n\right)\right) \geq 1 - \delta.$$

Proof. By homogeneity, it is enough to consider only $\boldsymbol{u} \in H$ such that

$$\max\left(h(\boldsymbol{u}), \|\mathbb{X}\boldsymbol{u}\|_n / L\right) = 1$$

where $L \triangleq \sqrt{n}/(\sigma\Phi^{-1}(1-\delta))$. Define $T \subset H$ and $f : \mathbb{R}^n \to \mathbb{R}$ by

$$T \triangleq \left\{ \boldsymbol{u} \in H : \max\left(h(\boldsymbol{u}), \frac{1}{L}\|\mathbb{X}\boldsymbol{u}\|_n\right) \leq 1 \right\}, \qquad f(\boldsymbol{v}) \triangleq \sup_{\boldsymbol{u}\in T} \frac{1}{n}(\sigma\boldsymbol{v})^T \mathbb{X}\boldsymbol{u} \tag{3.1}$$

for all $\boldsymbol{v} \in \mathbb{R}^n$. Then, for every $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$, $|f(\boldsymbol{v}_1) - f(\boldsymbol{v}_2)| \leq ((\sigma L)/\sqrt{n})\,|\boldsymbol{v}_1 - \boldsymbol{v}_2|_2$ where $|\cdot|_2$ denotes the Euclidean norm onto $\mathbb{R}^n$. Therefore, f is a Lipschitz function with Lipschitz constant $\sigma L/\sqrt{n}$ and by the Gaussian concentration inequality, cf., e.g., [17, Theorem 6.2], we have that with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{\boldsymbol{u}\in T} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u} &\leq \mathrm{Med}\left[\sup_{\boldsymbol{u}\in T} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u}\right] + \sigma L \frac{\Phi^{-1}(1-\delta)}{\sqrt{n}} \\
&\leq \mathrm{Med}\left[\sup_{\boldsymbol{u}\in\mathbb{R}^p: h(\boldsymbol{u})\leq 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{u}\right] + \sigma L \frac{\Phi^{-1}(1-\delta)}{\sqrt{n}} \\
&\leq \tau + \sigma L \frac{\Phi^{-1}(1-\delta)}{\sqrt{n}} = \tau + 1,
\end{aligned}$$

where $\mathrm{Med}[\zeta]$ denotes the median of random variable ζ and we have used the fact that $\mathbb{P}(\Omega) \geq 1/2$ to bound the median from above. ∎

The next proposition is a simple fact based on convexity argument. In different versions, it was used as an element of the proof of oracle inequalities with leading constant 1 starting from [13]. Some special cases of it are explicitly stated in [1, Lemma 1] and [2, Lemma A.2].

Proposition 3.2. *Let $F : H \to \mathbb{R}$ be a convex function, and let $\mathbb{X} : H \to \mathbb{R}^n$ be a linear operator. If $\hat{\boldsymbol{\beta}}$ is a solution of the minimization problem (2.1), then $\hat{\boldsymbol{\beta}}$ satisfies, for all $\boldsymbol{\beta} \in \mathbb{B}$ and all $\mathbf{f} \in \mathbb{R}^n$,*

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \leq \frac{2}{n}\boldsymbol{\xi}^T(\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) + F(\boldsymbol{\beta}) - F(\hat{\boldsymbol{\beta}}) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2. \tag{3.2}$$

Proof. Define the functions f and g by the relations $g(\boldsymbol{\beta}) = \|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2$, and $f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + F(\boldsymbol{\beta})$ for all $\boldsymbol{\beta} \in H$. Since f is convex and $\hat{\boldsymbol{\beta}}$ is a minimizer of f on $\mathbb{B}$, it follows that for some $\boldsymbol{w}$ in the sub-differential of f at $\hat{\boldsymbol{\beta}}$, we have $\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{w}\rangle \leq 0$ for all $\boldsymbol{\beta} \in \mathbb{B}$, cf., e.g., [23]. Using the Moreau-Rockafellar theorem, we obtain that there exists $\boldsymbol{v}$ in the sub-differential of F at $\hat{\boldsymbol{\beta}}$ such that $\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{w}\rangle = \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \frac{2}{n}\mathbb{X}^*(\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{y}) + \boldsymbol{v}\rangle$ for all $\boldsymbol{\beta} \in \mathbb{B}$ where $\mathbb{X}^*$ is the adjoint operator of $\mathbb{X}$. Thus,

$$\frac{2}{n}(\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^T(\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{y}) \leq \langle \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \boldsymbol{v}\rangle.$$

Note also that by simple algebra,

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 = \frac{2}{n}(\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^T(\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2.$$

Combining the last two displays we obtain

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \le \frac{2}{n}\boldsymbol{\xi}^T(\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) - \|\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_n^2 + \langle \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \boldsymbol{v}\rangle.$$

To complete the proof, notice that by definition of the subdifferential of F at $\hat{\boldsymbol{\beta}}$, we have $\langle \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}, \boldsymbol{v}\rangle \le F(\boldsymbol{\beta}) - F(\hat{\boldsymbol{\beta}})$. ∎

4 Oracle inequalities

In this section, we consider a Hilbert space H and a linear operator $\mathbb{X} : H \to \mathbb{R}^n$ defined by the relation

$$\mathbb{X}\boldsymbol{\beta} = (\langle \boldsymbol{\beta}, X_1\rangle, \dots, \langle \boldsymbol{\beta}, X_n\rangle)^\top, \quad \forall \boldsymbol{\beta} \in H,$$

where $X_1, \dots, X_n$ are deterministic elements of H.

We will also assume that $F(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|$ where $\|\cdot\|$ is a norm on H (called the regularization norm) and $\lambda > 0$ is a tuning constant. Thus, the minimization problem (2.1) takes the form

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{B}} \Big(\|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2 + \lambda\|\boldsymbol{\beta}\| \Big) \tag{4.1}$$

where $\mathbb{B}$ is a closed convex subset of H.

To each matrix $A \in H$, we associate a linear operator $\mathcal{P}_A : H \to H$. Examples of $\mathcal{P}_A$ that are interesting in the context of high-dimensional statistics will be given later. Set $\mathcal{P}_A^\perp = I - \mathcal{P}_A$ where I is the identity operator on H. The following assumption will be crucial for the subsequent argument.

Assumption 4.1. *There exists a subset $\mathbb{A}$ of $\mathbb{B}$ such that*

$$\mathcal{P}_A A = A, \qquad \forall\ A \in \mathbb{A},$$

$$\|A\| - \|B\| \le \|\mathcal{P}_A(A - B)\| - \|\mathcal{P}_A^\perp B\|, \qquad \forall\ A \in \mathbb{A},\ \forall\ B \in H. \tag{4.2}$$

Note that since $\mathcal{P}_A A = A$, inequality (4.2) can be rewritten as

$$\|A\| + \|\mathcal{P}_A^\perp B\| \le \|A - \mathcal{P}_A(B)\| + \|B\|, \qquad \forall\ A \in \mathbb{A},\ \forall\ B \in H. \tag{4.3}$$

Looking at (4.3), it is easy to check that Assumption 4.1 is satisfied if the following decomposability property holds.

Assumption 4.2 (Decomposability assumption)**.** *There exists a subset $\mathbb{A}$ of $\mathbb{B}$ such that*

$$\mathcal{P}_A A = A, \qquad \forall\ A \in \mathbb{A},$$

$$\|A\| + \|\mathcal{P}_A^\perp B\| = \|A + \mathcal{P}_A^\perp B\|, \qquad \forall\ A \in \mathbb{A},\ \forall\ B \in H. \tag{4.4}$$

This decomposability assumption is satisfied, with suitable definitions of $\mathcal{P}_A$, for the three regularization norms $\|\cdot\|$ playing the central role in high-dimensional statistics: the ℓ_1-norm, the group LASSO norm, and the nuclear norm. They are analyzed in Section 6. Beyond the decomposable case, one may turn to other assumptions stated in terms of the "size" of sub-differential of the regularization norm, cf. [15].

To state the result, we will need some notation. For any $A \in H$ and any constant $c_0 > 0$, define the following cone in $\mathbb{B}$:

$$\mathbb{C}_{A,c_0} \triangleq \Big\{ B \in \mathbb{B} : \|\mathcal{P}_A^\perp B\| \le c_0\|\mathcal{P}_A B\| \Big\},$$

and introduce the associated quantity that we will call the *compatibility factor* :

$$\mu_{c_0}(A) \triangleq \inf\Big\{\mu' > 0 : \|\mathcal{P}_A B\| \le \mu' \|\mathbb{X}B\|_n,\ \forall B \in \mathbb{C}_{A,c_0}\Big\}. \tag{4.5}$$

Note that $\mu_{c_0}(A)$ is a nondecreasing function of c_0.

Theorem 4.3. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$, and that Assumption 4.1 holds. Let $\tau' > 0$ be such that the event*

$$\Omega = \left\{ \sup_{\boldsymbol{v}\in H:\|\boldsymbol{v}\|\le 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{v} \le \tau' \right\}$$

satisfies $\mathbb{P}(\Omega) \ge 1/2$. Let $\lambda \ge 10\tau'$ and $\delta \in (0,1)$. Then, the estimator $\hat{\boldsymbol{\beta}}$ defined in (4.1) satisfies, with probability at least $1-\delta$,

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \inf_{\boldsymbol{\beta}\in\mathbb{A}} \left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2 \mu_4^2(\boldsymbol{\beta}) \right] + \frac{16\sigma^2(\Phi^{-1}(1-\delta))^2}{n} \tag{4.6}$$

where, in particular, $(\Phi^{-1}(1-\delta))^2 \le 2\log(1/\delta)$. If, in addition, $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{A}$, then with probability at least $1-\delta$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \le 4\lambda\mu_4^2(\boldsymbol{\beta}^*) + \frac{20\sigma^2(\Phi^{-1}(1-\delta))^2}{n\lambda}. \tag{4.7}$$

Proof. Note that

$$\Omega = \left\{ \sup_{\boldsymbol{v}\in H:\lambda\|\boldsymbol{v}\|/5\le 1} \frac{1}{n}\boldsymbol{\xi}^T \mathbb{X}\boldsymbol{v} \le 5\tau'/\lambda \right\}.$$

By Proposition 3.1 with $h(\boldsymbol{v}) = \lambda\|\boldsymbol{v}\|/5$ and $\tau = 5\tau'/\lambda$ we obtain that, on an event Ω' of probability at least $1-\delta$,

$$\forall \boldsymbol{u} \in H: \quad \frac{1}{n}\boldsymbol{\xi}^T\mathbb{X}\boldsymbol{u} \le (5\tau'/\lambda + 1)\max\left(\lambda\|\boldsymbol{u}\|/5, \nu\|\mathbb{X}\boldsymbol{u}\|_n\right)$$

where

$$\nu = \frac{\sigma\Phi^{-1}(1-\delta)}{\sqrt{n}}.$$

In the rest of the proof, we will place ourselves on the event Ω'. Using Proposition 3.2 and the last display we find that on Ω', for all $\boldsymbol{\beta} \in \mathbb{B}$,

$$\begin{aligned} \|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 &\le 2(5\tau'/\lambda + 1)\max\left(\lambda\|\boldsymbol{u}\|/5, \nu\|\mathbb{X}\boldsymbol{u}\|_n\right) \\ &\quad + \lambda\|\boldsymbol{\beta}\| - \lambda\|\hat{\boldsymbol{\beta}}\| - \|\mathbb{X}\boldsymbol{u}\|_n^2 \\ &\le 3\max\left(\lambda\|\boldsymbol{u}\|/5, \nu\|\mathbb{X}\boldsymbol{u}\|_n\right) \\ &\quad + \lambda\|\boldsymbol{\beta}\| - \lambda\|\hat{\boldsymbol{\beta}}\| - \|\mathbb{X}\boldsymbol{u}\|_n^2 \end{aligned} \tag{4.8}$$

where $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. We now consider separately three cases.

Case 1: Matrix $\boldsymbol{\beta} \in \mathbb{A}$ is such that $\lambda\|\boldsymbol{u}\|/5 \le \nu\|\mathbb{X}\boldsymbol{u}\|_n$. Then,

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \le 8\nu\|\mathbb{X}\boldsymbol{u}\|_n - \|\mathbb{X}\boldsymbol{u}\|_n^2 \le 16\nu^2. \tag{4.9}$$

Thus, for such $\boldsymbol{\beta}$ inequality (4.6) is satisfied.

The next two cases correspond to $\boldsymbol{\beta} \in \mathbb{A}$ such that $\lambda\|\boldsymbol{u}\|/5 > \nu\|\mathbb{X}\boldsymbol{u}\|_n$. If this inequality holds, then (4.8) implies

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \le \lambda(3\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/5 + \|\boldsymbol{\beta}\| - \|\hat{\boldsymbol{\beta}}\|) - \|\mathbb{X}\boldsymbol{u}\|_n^2. \tag{4.10}$$

Assumption 4.1 with $A = \boldsymbol{\beta}$ and $B = \hat{\boldsymbol{\beta}}$ grants that

$$\|\boldsymbol{\beta}\| - \|\hat{\boldsymbol{\beta}}\| \le \|\mathcal{P}_{\boldsymbol{\beta}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\| - \|\mathcal{P}_{\hat{\boldsymbol{\beta}}}^{\perp}\hat{\boldsymbol{\beta}}\|$$

while, by the triangle inequality,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \le \|\mathcal{P}_{\boldsymbol{\beta}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\| + \|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\| = \|\mathcal{P}_{\boldsymbol{\beta}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\| + \|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}\hat{\boldsymbol{\beta}}\|.$$

Combining the last two inequalities we obtain

$$3\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|/5 + \|\boldsymbol{\beta}\| - \|\hat{\boldsymbol{\beta}}\| \le 8\|\mathcal{P}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|/5 - 2\|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}\hat{\boldsymbol{\beta}}\|/5.$$

This inequality and (4.10) imply

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \le (2\lambda/5)\left(4\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{u}\| - \|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}\boldsymbol{u}\|\right) - \|\mathbb{X}\boldsymbol{u}\|_n^2. \tag{4.11}$$

Case 2: Matrix $\boldsymbol{\beta} \in \mathbb{A}$ is such that $\lambda\|\boldsymbol{u}\|/5 > \nu\|\mathbb{X}\boldsymbol{u}\|_n$ and $4\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{u}\| < \|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}\boldsymbol{u}\|$. Then, in view of (4.11), inequality (4.6) holds trivially.

Case 3: Matrix $\boldsymbol{\beta} \in \mathbb{A}$ is such that $\lambda\|\boldsymbol{u}\|/5 > \nu\|\mathbb{X}\boldsymbol{u}\|_n$ and $4\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{u}\| \ge \|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}\boldsymbol{u}\|$. Then $\boldsymbol{u}$ belongs to the cone $\mathbb{C}_{\boldsymbol{\beta},4}$, so that $\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{u}\| \le \mu_4(\boldsymbol{\beta})\|\mathbb{X}\boldsymbol{u}\|_n$. This and (4.11) yield

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 - \|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 \le \frac{8\lambda\mu_4(\boldsymbol{\beta})}{5}\|\mathbb{X}\boldsymbol{u}\|_n - \|\mathbb{X}\boldsymbol{u}\|_n^2 \le \frac{16}{25}\lambda^2\mu_4^2(\boldsymbol{\beta}),$$

and hence inequality (4.6).

Consider now the well-specified case: $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{A}$. Set $\boldsymbol{u} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Again, we proceed in cases.

Case 1: Matrix $\boldsymbol{\beta}^* \in \mathbb{A}$ is such that $\lambda\|\boldsymbol{u}\|/5 \le \nu\|\mathbb{X}\boldsymbol{u}\|_n$. Then, inequality (4.9) with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ implies $\|\mathbb{X}\boldsymbol{u}\|_n \le 4\nu$, so that $\|\boldsymbol{u}\| \le 20\nu^2/\lambda$. The bound (4.7) follows.

Case 2: Matrix $\boldsymbol{\beta}^* \in \mathbb{A}$ is such that $\lambda\|\boldsymbol{u}\|/5 > \nu\|\mathbb{X}\boldsymbol{u}\|_n$. Then, from (4.11) with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ we obtain that $4\|\mathcal{P}_{\boldsymbol{\beta}^*}\boldsymbol{u}\| \ge \|\mathcal{P}_{\boldsymbol{\beta}^*}^{\perp}\boldsymbol{u}\|$, and consequently, $\|\mathcal{P}_{\boldsymbol{\beta}^*}\boldsymbol{u}\| \le \mu_4(\boldsymbol{\beta}^*)\|\mathbb{X}\boldsymbol{u}\|_n$. On the other hand, (4.11) also implies that

$$\|\mathbb{X}\boldsymbol{u}\|_n^2 \le 4\lambda\|\mathcal{P}_{\boldsymbol{\beta}^*}\boldsymbol{u}\|/5.$$

In conclusion, $\|\mathcal{P}_{\boldsymbol{\beta}^*}\boldsymbol{u}\| \le 4\lambda\mu_4^2(\boldsymbol{\beta}^*)/5$. Finally, $\|\boldsymbol{u}\| = \|\mathcal{P}_{\boldsymbol{\beta}^*}\boldsymbol{u}\| + \|\mathcal{P}_{\boldsymbol{\beta}^*}^{\perp}\boldsymbol{u}\| \le 5\|\mathcal{P}_{\boldsymbol{\beta}^*}\boldsymbol{u}\| \le 4\lambda\mu_4^2(\boldsymbol{\beta}^*)$. The bound (4.7) follows.

■

By integration over δ, we can readily derive from Theorem 4.3 bounds for any moments of $\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$. In particular, the following corollary is immediate.

Corollary 4.4. *Under the assumptions of Theorem 4.3, the estimator $\hat{\boldsymbol{\beta}}$ defined in (4.1) satisfies*

$$\mathbb{E}\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \le \min_{\boldsymbol{\beta}\in\mathbb{A}}\left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(\boldsymbol{\beta})\right] + \frac{16\sigma^2}{n}. \tag{4.12}$$

If, in addition, $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{A}$, then

$$\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \le 8\lambda\mu_4^2(\boldsymbol{\beta}^*) + \frac{20\sigma}{\lambda n}. \tag{4.13}$$

Note that the regularization parameter λ does not depend on the parameter δ that defines the confidence level. This is a key to get results in expectation as in Corollary 4.4.

5 Control of the compatibility factor

As follows from Theorem 4.3 and Corollary 4.4, the performance of penalized LS estimators is driven by the compatibility factor $\mu_{c_0}(A)$ defined in (4.5). The aim of this section is to provide a control of this quantity uniformly over all $A \in \mathbb{A}$ with high probability when $X_1, \dots, X_n$ are n independent and identically distributed (i.i.d.) realizations of an H-valued random variable X. We will consider X satisfying the following assumption, cf. [14, 20].

Assumption 5.1 (Small ball assumption)**.** *There exist constants $\beta_0 > 0$ and $\kappa_0 \in (0,1)$ such that for all $B \in \mathbb{B}$,*

$$\mathbb{P}\left[|\langle X, B\rangle| \geq \beta_0 \|B\|_H\right] \geq \kappa_0.$$

This assumption is rather mild. We refer the reader to [14, 19, 20] for some examples. A simple sufficient condition for the small ball assumption is given in the next lemma.

Lemma 5.2. *Assume that X is isotropic in the sense that*

$$\forall B \in H, \qquad \mathbb{E}\langle X, B\rangle^2 = \|B\|_H^2. \tag{5.1}$$

Furthermore, assume that there exists a constant $L > 0$ such that for any $B \in H$,

$$\mathbb{E}\left[\langle X, B\rangle^4\right]^{1/4} \leq 2L\mathbb{E}\left[\langle X, B\rangle^2\right]^{1/2}. \tag{5.2}$$

Then X satisfies the small ball assumption with parameters

$$\beta_0 = 1/\sqrt{2} \qquad \text{and} \qquad \kappa_0 = 1/(64L^4). \tag{5.3}$$

Proof. It follows from the Paley-Zygmund inequality (cf., for instance, Proposition 3.3.1 in [8])) that

$$\begin{aligned}\mathbb{P}\left(|\langle X, B\rangle| \geq \beta_0\|B\|_H\right) &= \mathbb{P}\left(|\langle X, B\rangle|^2 \geq \beta_0^2\mathbb{E}\left[\langle X, B\rangle^2\right]\right), \\ &\geq (1-\beta_0^2)^2\mathbb{E}\left[\langle X, B\rangle^2\right]^2\mathbb{E}\left[\langle X, B\rangle^4\right]^{-1} \geq (1-\beta_0^2)^2\left(\frac{1}{2L}\right)^4.\end{aligned}$$

Hence, X satisfies the small ball assumption with parameters β_0, κ_0 defined in (5.3). ∎

Lemma (5.2) shows that the small ball assumption is satisfied under weak moment conditions. Indeed, the existence of moments $\mathbb{E}\langle X, B\rangle^p$ for $p > 4$ is not required.

The small ball assumption is helpful in situations where one needs to bound from below an empirical process with nonnegative terms. Note that $\|\mathbb{X}B\|_n^2 = (1/n)\sum_{i=1}^n \langle X_i, B\rangle^2$ is an empirical process with nonnegative terms considered as a function of $B \in \mathbb{C}_{A,c_0}$. If we obtain a uniform lower bound on it, an upper bound on the compatibility factor $\mu_{c_0}(A)$ follows. The next theorem, cf. [14], provides such a lower bound on $\|\mathbb{X}B\|_n^2$ based on the small ball argument. For the sake of completeness, we recall here its proof.

Theorem 5.3 (cf. Theorem 2.1 in [14])**.** *Let X be an H-valued random variable satisfying Assumption 5.1 with parameters $\beta_0 > 0$ and $\kappa_0 \in (0,1)$. Let $X_1, \ldots, X_n$ be n i.i.d. realizations of X. Assume that*

$$\mathbb{E}\sup_{B \in S_2 \cap \left(\cup_{A\in\mathbb{A}}\mathbb{C}_{A,c_0}\right)}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle X_i, B\rangle\right| \leq \frac{\kappa_0\beta_0}{16} \tag{5.4}$$

where S_2 is the unit sphere in H and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random variables uniformly distributed on $\{-1, 1\}$ and independent of $X_1, \ldots, X_n$. Then, with probability greater than $1 - \exp\left(-n\kappa_0^2/32\right)$, for all $B \in \cup_{A\in\mathbb{A}}\mathbb{C}_{A,c_0}$ we have

$$\|\mathbb{X}B\|_n \geq \|B\|_H\sqrt{\frac{\beta_0^2\kappa_0}{8}}.$$

Proof. By homogeneity, it is enough to prove the result for all $B \in \mathcal{B}$ where $\mathcal{B} \triangleq S_2 \cap (\cup_{A\in\mathbb{A}}\mathbb{C}_{A,c_0})$. Denote by P_n the empirical measure associated to $X_1,\dots,X_n$. Let $B \in S_2$. We have

$$\begin{aligned}\|\mathbb{X}B\|_n^2 &= \frac{1}{n}\sum_{i=1}^n \langle X_i, B\rangle^2 \triangleq P_n\langle\cdot, B\rangle^2 \geq \frac{\beta_0^2}{4} P_n\left[|\langle\cdot, B\rangle| \geq (\beta_0/2)\right] \\ &= \frac{\beta_0^2}{4}\left\{\mathbb{P}\left[|\langle X, B\rangle| \geq \beta_0\right] + P_n\left[|\langle\cdot, B\rangle| \geq (\beta_0/2)\right] - \mathbb{P}\left[|\langle X, B\rangle| \geq \beta_0\right]\right\} \\ &\geq \frac{\beta_0^2}{4}\left\{\kappa_0 + (P_n - P)\phi\left(|\langle\cdot, B\rangle|\right)\right\} \qquad (5.5)\end{aligned}$$

where in the last inequality we used the small ball assumption and the fact that $P_n\left[|\langle\cdot, B\rangle| \geq (\beta_0/2)\right] \geq P_n\phi\left(|\langle\cdot, B\rangle|\right)$ and $\mathbb{P}\left[|\langle X, B\rangle| \geq \beta_0\right] \leq P\phi\left(|\langle\cdot, B\rangle|\right)$ where ϕ is defined by

$$\phi(t) = \begin{cases} 1 & \text{if } t \geq \beta_0 \\ 2t/\beta_0 - 1 & \text{if } \beta_0/2 \leq t \leq \beta_0 \\ 0 & \text{otherwise.}\end{cases}$$

Set now

$$f(X_1,\dots,X_n) = \sup_{B\in\mathcal{B}}(P - P_n)\phi\left(|\langle\cdot, B\rangle|\right).$$

It follows from the bounded difference inequality (cf., for instance, Theorem 6.2 in [5]) that for all $x > 0$, with probability greater than $1 - \exp(-x)$,

$$f(X_1,\dots,X_n) \leq \mathbb{E}f(X_1,\dots,X_n) + \sqrt{\frac{2x}{n}}.$$

This and the Giné-Zinn symmetrization inequality (cf., for instance, Chapter 2.3 in [29]) yields that for all $x > 0$, with probability greater than $1 - \exp(-x)$,

$$f(X_1,\dots,X_n) \leq 2\mathbb{E}\sup_{B\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi\left(|\langle X_i, B\rangle|\right) + \sqrt{\frac{2x}{n}}. \qquad (5.6)$$

Note that ϕ is a Lipschitz function with Lipschitz constant $2/\beta_0$ and $\phi(0) = 0$. Thus, it follows from the contraction inequality (cf. equation (4.20) in [16]) that

$$\mathbb{E}\sup_{B\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi\left(|\langle X_i, B\rangle|\right) \leq \frac{2}{\beta_0}\mathbb{E}\sup_{B\in\mathcal{B}}\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle X_i, B\rangle \leq \frac{\kappa_0}{8}$$

where the last inequality is due to (5.4). Combining this bound with (5.6) and choosing $x = n\kappa_0^2/32$ we obtain that $f(X_1,\dots,X_n) \leq \kappa_0/2$ with probability greater than $1-\exp\left(-n\kappa_0^2/32\right)$. Therefore, with the same probability, $(P_n-P)\phi\left(|\langle\cdot, B\rangle|\right) \geq -\kappa_0/2$ for all $B \in \mathcal{B}$. This and (5.5) prove the theorem. ∎

It follows from Theorem 5.3 that if X satisfies the small ball assumption and n is large enough so that (5.4) holds then, with probability greater than $1-\exp(-n\kappa_0^2/32)$, for all $A \in \mathbb{A}$,

$$\mu_{c_0}(A) \leq \left(\frac{8}{\beta_0^2\kappa_0}\right)^{1/2}\sup_{B\in\mathbb{C}_{A,c_0}}\frac{\|\mathcal{P}_A B\|}{\|B\|_H}. \qquad (5.7)$$

Thus, we have reduced the control of $\mu_{c_0}(A)$ to the bound (5.4) on the expectation of the empirical process. Under certain assumptions, this expectation can be controlled in terms of the *Gaussian mean width* of the set $S_2\cap(\cup_{A\in\mathbb{A}}\mathbb{C}_{A,c_0})$ as explained below. Then, we can derive an estimate on a sufficient number n of observations for (5.4) to hold. The argument can be carried out using the main result from [21]. To state this result, we first introduce the definition of the Gaussian mean width of a subset of a Hilbert space and the definition of a K-unconditional norm.

Let $\mathbb{C}$ be a subset of the Hilbert space H. We denote by $(G_B)_{B\in\mathbb{C}}$ the centered gaussian process indexed by $\mathbb{C}$ having the same covariance structure as X, that is $\mathbb{E}G_B = 0$ and $\mathbb{E}G_{B_1}G_{B_2} = \mathbb{E}\langle B_1, X\rangle\langle X, B_2\rangle$ for all $B, B_1, B_2 \in \mathbb{C}$ (we refer the reader to [17] or to Chapter 12 in [10] for more details on Gaussian processes in Hilbert spaces). The *Gaussian mean width of* $\mathbb{C}$ is defined as

$$\ell^*(\mathbb{C}) = \sup\left\{\mathbb{E}\max_{B\in\mathbb{C}'} G_B : \mathbb{C}' \subset \mathbb{C} \text{ is finite}\right\}. \tag{5.8}$$

This supremum is called the lattice supremum (see Chapter 2.2 in [16] for more details).

In the following, we consider a finite dimensional Hilbert space H and we denote by d its dimension. The two examples analyzed in Section 6 are $H = \mathbb{R}^p$ and $H = \mathbb{R}^{k\times m}$. In this case, for all $\mathbb{C} \subset H$ we have

$$\ell^*(\mathbb{C}) = \mathbb{E}\sup_{B\in\mathbb{C}}\langle G, B\rangle$$

where G is a H-valued random variable with i.i.d. $\mathcal{N}(0,1)$ components. We will also need the following definition, cf. [21].

Definition 5.4. *Let H be a finite dimensional Hilbert space, let $(e_j)_{j=1,\dots,d}$ be a basis in H, and $K > 0$. A norm $\|\cdot\|$ on H is called K-unconditional norm with respect to the basis $(e_j)_{j=1,\dots,d}$ if the following two properties hold.*

- *For any $B \in H$ and any permutation π of $\{1,\dots,d\}$,*

$$\Big\|\sum_{j=1}^d \langle B, e_j\rangle e_j\Big\| \le K\Big\|\sum_{j=1}^d \langle B, e_{\pi(j)}\rangle e_j\Big\|.$$

- *If $A \in H$ is such that $\langle A, e_j\rangle^\sharp \le \langle B, e_j\rangle^\sharp$ for all $j = 1,\dots,d$, then*

$$\Big\|\sum_{j=1}^d \langle A, e_j\rangle e_j\Big\| \le K\Big\|\sum_{j=1}^d \langle B, e_j\rangle e_j\Big\|$$

  where $(\langle B, e_j\rangle^\sharp)_j$ is the nonincreasing rearrangement of $(|\langle B, e_j\rangle|)_j$.

The class of K-unconditional norms is rather rich. It includes, in particular, the ℓ_p-norms. For more details see [21].

A bound on the expectation of the empirical process in (5.4) can be obtained from the following result.

Theorem 5.5. [21, Theorem 1.6] *There exists an absolute constant $c_1 > 0$ such that the following holds. Let H be a finite dimensional Hilbert space, let X be a random vector with values in H and let $\mathbb{C} \subset H$. Denote by $(e_j)_{j=1,\dots,d}$ a basis in H. Let $L \ge 1$, and assume that:*

(i) The set $\mathbb{C}$ is such that $\|\cdot\|_{\mathbb{C}^\circ} \triangleq \sup_{v\in\mathbb{C}}\langle v, \cdot\rangle$ is a K-unconditional norm.

(ii) The distribution of X is isotropic, i.e., satisfies (5.1), and for any $j = 1,\dots,d$, and any positive integer k smaller than $c_1 \log d$ we have

$$(\mathbb{E}|\langle X, e_j\rangle|^k)^{1/k} \le L\sqrt{k}.$$

Let $X_1,\dots,X_n$ be i.i.d. realizations of X and let $\epsilon_1,\dots,\epsilon_n$ be i.i.d. random variables uniformly distributed on $\{-1,1\}$ and independent of $X_1,\dots,X_n$. Then

$$\mathbb{E}\sup_{B\in\mathbb{C}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\langle X_i, B\rangle\right| \le C(L,K)\ell^*(\mathbb{C}), \tag{5.9}$$

where $C(L,K)$ is a constant that depends only on K and L.

If condition (i) of this theorem does not hold, i.e., if $\|\cdot\|_{\mathbb{C}^\circ}$ is not an unconditional norm, one may derive a similar result under a more constraining assumption, namely that the random variable $\langle X, B\rangle$ is subgaussian for any $B \in H$. The next proposition follows from the majorizing measure theorem, cf. [26, Chapter 1] or [30, Chapter 6].

Proposition 5.6. *Let $L \geq 1$ and let H be a finite dimensional Hilbert space. Assume that X is isotropic, i.e., it satisfies (5.1). Assume also that X is L-subgaussian in the sense that for all $B \in H$ such that $\|B\|_H = 1$ we have $\mathbb{E}\exp(t\langle X, B\rangle) \leq \exp(t^2L^2/2)$ for all $t > 0$. Then X satisfies the small ball assumption with parameters β_0, κ_0 defined in (5.3). Furthermore, there exists an absolute constant $c_2 > 0$ such that (5.9) holds with $C(L, K) = c_2L$ for any $\mathbb{C} \subset H$.*

Proof. Let $Z = \langle X, B\rangle$ and assume w.l.o.g. that $\|B\|_H = 1$. The random variable Z is L-subgaussian and, by isotropy, $\mathbb{E}Z^2 = 1$. Thus by [5, (2.3) from Theorem 2.1] we have $\mathbb{E}Z^4 \leq 16L^4$, or equivalently $\mathbb{E}[Z^4]^{1/4} \leq 2L\mathbb{E}[Z^2]^{1/2}$. By Lemma 5.2, this implies that X satisfies the small ball assumption with parameters β_0, κ_0 defined in (5.3).

To prove (5.9), note that $\epsilon_i X_i$ is L-subgaussian. Thus, (5.9) with $C(L, K) = c_2L$ follows from the majorizing measure theorem for subgaussian processes, cf. [30, Corollary 6.26]. ∎

6 Examples

In what follows, we denote by $|\cdot|_q$ the ℓ_q norm of a finite dimensional vector, $1 \leq q \leq \infty$. We denote by $\|\cdot\|_{Fr}$ and by $\|\cdot\|_{sp}$ the Frobenius norm and the spectral norm of a matrix, respectively. Let S_2^{p-1} and B_q^p denote the unit Euclidean sphere in $\mathbb{R}^p$ and the unit ℓ_q-ball in $\mathbb{R}^p$, respectively. The canonical basis of $\mathbb{R}^p$ is denoted by $(e_j)_{j=1,\dots,p}$. For a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ and a subset $S \subseteq \{1, \dots, p\}$, we denote by $\mathrm{supp}(\boldsymbol{\beta})$ the support of $\boldsymbol{\beta}$, by $\boldsymbol{\beta}_S$ the orthogonal projection of $\boldsymbol{\beta}$ onto the linear span of $\{e_j : j \in S\}$, and by $|S|$ the cardinality of S. We will write $a \lesssim b$ if there is an absolute constant $C > 0$ such that $a \leq Cb$.

6.1 LASSO

We consider here $H = \mathbb{B} = \mathbb{R}^p$ equipped with the Euclidean norm $\|\cdot\|_H = |\cdot|_2$ and we define the regularization norm $\|\cdot\|$ as the ℓ_1 norm. Then the estimator $\hat{\boldsymbol{\beta}}$ is the LASSO estimator

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_n^2 + \lambda|\boldsymbol{\beta}|_1 \right) \tag{6.1}$$

where $\lambda > 0$ is a tuning parameter.

Given $\boldsymbol{\beta} \in \mathbb{R}^p$ it is straightforward to verify that Assumption 4.1 is satisfied when $\mathcal{P}_{\boldsymbol{\beta}}$ is the orthogonal projection operator onto the linear span of $\{e_j : j \in \mathrm{supp}(\boldsymbol{\beta})\}$ where $(e_j)_{j=1,\dots,p}$ is the canonical basis of $\mathbb{R}^p$.

The operator $\mathbb{X}$ is a matrix in $\mathbb{R}^{n\times p}$ while the event Ω in Theorem 4.3 can be written in the form

$$\Omega = \left\{ \frac{1}{n}|\mathbb{X}^T\boldsymbol{\xi}|_\infty \leq \tau' \right\}. \tag{6.2}$$

In order to apply Theorem 4.3, we need to find τ' such that $\mathbb{P}(\Omega) \geq 1/2$. Assume first that $\mathbb{X}$ is deterministic. The following lemma is a direct consequence of the normal tail probability bounds and the union bound, cf. [3].

Lemma 6.1. *Let $(e_j)_{j=1,\dots,p}$ be the canonical basis of $\mathbb{R}^p$ and let $\mathbb{X}$ be deterministic. If*

$$\tau' \geq \sigma \max_{1\leq j\leq p} \|\mathbb{X}e_j\|_n \sqrt{\frac{2\log p}{n}}, \tag{6.3}$$

then the event (6.2) has probability at least 1/2.

In view of this lemma, oracle inequalities for the LASSO estimator with tuning parameter λ satisfying

$$\lambda \geq 10\sigma \max_{1\leq j\leq p} \|\mathbb{X}e_j\|_n \sqrt{\frac{2\log p}{n}} \tag{6.4}$$

follow from Theorem 4.3 and Corollary 4.4. They have the following form.

Theorem 6.2. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that $\mathbb{X}$ is deterministic. Let $\delta \in (0,1)$. The LASSO estimator $\hat{\boldsymbol{\beta}}$ with tuning parameter satisfying (6.4) is such that, with probability at least $1-\delta$,*

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(\boldsymbol{\beta})\right] + \frac{16\sigma^2(\Phi^{-1}(1-\delta))^2}{n}$$

and

$$\mathbb{E}\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(\boldsymbol{\beta})\right] + \frac{16\sigma^2}{n}.$$

If, in addition, $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$, then with probability at least $1-\delta$,

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 4\lambda\mu_4^2(\boldsymbol{\beta}^*) + \frac{20\sigma^2(\Phi^{-1}(1-\delta))^2}{n\lambda}$$

and

$$\mathbb{E}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \leq 8\lambda\mu_4^2(\boldsymbol{\beta}^*) + \frac{20\sigma}{\lambda n}.$$

To make these inequalities more explicit, we need to control the compatibility factor $\mu_{c_0}(\boldsymbol{\beta})$. First note that one may use the Restricted Eigenvalue constant [4] to bound $\mu_{c_0}(\boldsymbol{\beta})$ from above. For any $S \subset \{1,\dots,p\}$ and $c_0 > 0$, we define the Restricted Eigenvalue constant $\kappa(S, c_0) \geq 0$ by the formula

$$\kappa^2(S, c_0) \triangleq \min_{\boldsymbol{\delta}\in\mathbb{R}^p\setminus\{0\}:|\boldsymbol{\delta}_{S^c}|_1\leq c_0|\boldsymbol{\delta}_S|_1} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n^2}{|\boldsymbol{\delta}|_2^2}. \tag{6.5}$$

Therefore, for all $\boldsymbol{\beta}$ such that $\kappa^2(\mathrm{supp}(\boldsymbol{\beta}), c_0) \neq 0$ we obtain the bound

$$\mu_{c_0}^2(\boldsymbol{\beta}) \leq \frac{|\mathrm{supp}(\boldsymbol{\beta})|}{\kappa^2(\mathrm{supp}(\boldsymbol{\beta}), c_0)}.$$

When $\mathbb{X}$ is deterministic and $\boldsymbol{\beta}$ is s-sparse (i.e., $|\mathrm{supp}(\boldsymbol{\beta})| \leq s$), there exist various sufficient conditions on $\mathbb{X}$ allowing one to bound $\kappa^2(\mathrm{supp}(\boldsymbol{\beta}), c_0)$ from below by a universal constant, cf., e.g., [4]. This leads to the bound $\mu_{c_0}^2(\boldsymbol{\beta}) \lesssim s$ for all s-sparse vectors $\boldsymbol{\beta}$.

Consider now the case of random $\mathbb{X}$. Specifically, assume that $X_1,\dots,X_n$ are i.i.d. realizations of a random vector X with values in $\mathbb{R}^p$. Then, it turns out that the bound $\mu_{c_0}^2(\boldsymbol{\beta}) \lesssim s$ for s-sparse vectors $\boldsymbol{\beta}$ can be guaranteed with high probability (with respect to the distribution of $X_1,\dots,X_n$) provided that $n \gtrsim s\log(ep/s)$. Indeed, combining Theorems 5.3 and 5.5 we obtain the following result.

Proposition 6.3. *Let $L \geq 1$ and let β_0, κ_0 be positive constants. There exist a constant $C(L) > 0$ depending only on L and an absolute constant $c_1 > 0$ such that the following holds.*

Let $X_1,\dots,X_n$ be i.i.d. realizations of a random vector X with values in $\mathbb{R}^p$ such that

- *(i) X satisfies the small ball assumption (Assumption 5.1) with parameters β_0, κ_0,*
- *(ii) X is isotropic (i.e., $\mathbb{E}XX^\top = I_{p\times p}$) and for all $j = 1,\dots,p$, and all positive integers k smaller than $c_1\log p$ we have $(\mathbb{E}|\langle X, e_j\rangle|^k)^{1/k} \leq L\sqrt{k}$.*

Let $s \in \{1, \dots, p\}$ and $c_0 > 0$. If

$$n \geq C(L)[(1+c_0)/(\kappa_0 \beta_0)]^2 s \log(ep/s), \tag{6.6}$$

then with probability greater than $1 - \exp(-n\kappa_0^2/32)$, for every $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\mathrm{supp}(\boldsymbol{\beta})| \leq s$ we have

$$\mu_{c_0}(\boldsymbol{\beta}) \leq \sqrt{\frac{8|\mathrm{supp}(\boldsymbol{\beta})|}{\beta_0^2 \kappa_0}}.$$

Proof. Denote by $B_0(s)$ the set of all s-sparse vectors in $\mathbb{R}^p$:

$$B_0(s) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \ |\mathrm{supp}(\boldsymbol{\beta})| \leq s\}.$$

Let $\boldsymbol{\beta} \in B_0(s)$ and recall that

$$\mathbb{C}_{\boldsymbol{\beta}, c_0} = \left\{\boldsymbol{\beta}' \in \mathbb{R}^p : |\mathcal{P}_{\boldsymbol{\beta}}^{\perp} \boldsymbol{\beta}'|_1 \leq c_0 |\mathcal{P}_{\boldsymbol{\beta}} \boldsymbol{\beta}'|_1\right\}$$

where $\mathcal{P}_{\boldsymbol{\beta}}$ is the projection operator onto the linear span of $\{e_j : j \in \mathrm{supp}(\boldsymbol{\beta})\}$. It follows from Theorem 5.3 and (5.7) that, if (5.4) with $\mathbb{A} = B_0(s)$ holds, then with probability greater than $1 - \exp(-n\kappa_0^2/32)$, for all $\boldsymbol{\beta} \in B_0(s)$ we have

$$\mu_{c_0}(\boldsymbol{\beta}) \leq \left(\frac{8}{\beta_0^2 \kappa_0}\right)^{1/2} \sup_{\boldsymbol{\beta}' \in \mathbb{C}_{\boldsymbol{\beta}, c_0}} \frac{|\mathcal{P}_{\boldsymbol{\beta}} \boldsymbol{\beta}'|_1}{|\boldsymbol{\beta}'|_2} \leq \left(\frac{8}{\beta_0^2 \kappa_0}\right)^{1/2} \sqrt{|\mathrm{supp}(\boldsymbol{\beta})|}$$

where we have used that $|\mathcal{P}_{\boldsymbol{\beta}} \boldsymbol{\beta}'|_1 \leq \sqrt{|\mathrm{supp}(\boldsymbol{\beta})|} |\boldsymbol{\beta}'|_2$ for all $\boldsymbol{\beta}' \in \mathbb{R}^p$.

Therefore, it only remains to prove that (6.6) implies (5.4) with $\mathbb{A} = B_0(s)$. First note that $S_2^{p-1} \cap \left(\cup_{\boldsymbol{\beta} \in B_0(s)} \mathbb{C}_{\boldsymbol{\beta}, c_0}\right) \subset \mathbb{C}$ where $\mathbb{C} = ((1+c_0)\sqrt{s} B_1^p) \cap B_2^p$. Since the ℓ_2 and ℓ_1 norms are 1-unconditional, it is straightforward to check that $\|\cdot\|_{\mathbb{C}^\circ} = \sup_{v \in \mathbb{C}} \langle v, \cdot \rangle$ is a 1-unconditional norm. Therefore, we can apply Theorem 5.5, which gives

$$\begin{aligned}
\mathbb{E} \sup_{\boldsymbol{\beta} \in S_2^{p-1} \cap \left(\cup_{\boldsymbol{\beta} \in B_0(s)} \mathbb{C}_{\boldsymbol{\beta}, c_0}\right)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle X_i, \boldsymbol{\beta} \rangle \right| &\leq \mathbb{E} \sup_{\boldsymbol{\beta} \in \mathbb{C}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle X_i, \boldsymbol{\beta} \rangle \right|, \\
&\leq \frac{c_2(L) \ell^*(\mathbb{C})}{\sqrt{n}} \leq \frac{c_3(L)(1+c_0)\sqrt{s \log(ep/s)}}{\sqrt{n}},
\end{aligned}$$

where $c_2(L)$ and $c_3(L)$ are positive constants depending only on L and where we used that $\ell^*(\mathbb{C}) \leq (1+c_0)\ell^*(\sqrt{s} B_1^p \cap B_2^p) \leq c_4(1+c_0)\sqrt{s \log(ep/s)}$ for some absolute constant c_4 (cf., for instance, Lemma 5.3 in [15]). If (6.6) holds with large enough constant $C(L) > 0$ depending only on L, then the right hand side of the last display is bounded from above by $\beta_0 \kappa_0 / 16$ and (5.4) is satisfied. ∎

Combining Theorem 6.2 and Proposition 6.3 we can obtain oracle inequalities for the LASSO estimator when $X_1, \dots, X_n$ are i.i.d. random vectors independent of the noise vector $\boldsymbol{\xi}$. To illustrate it, consider the following result for the basic example where all entries of matrix $\mathbb{X}$ are i.i.d. standard Gaussian.

Theorem 6.4. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n \times n})$ and that all entries of matrix $\mathbb{X}$ are i.i.d. standard Gaussian random variables independent of the noise vector $\boldsymbol{\xi}$. Let $\delta \in (0,1)$, and*

$$\lambda = a\sigma \sqrt{\frac{2 \log p}{n}} \tag{6.7}$$

where $a \geq 20$. There exist an absolute constant $C_1 > 0$ and a constant $C_2 > 0$ depending only on a such that the following holds. If $n \geq C_1 s \log(ep/s)$, then for the LASSO estimator $\hat{\boldsymbol{\beta}}$ with tuning parameter (6.7) we have that, with probability at least $1 - \delta - (p+1)e^{-n/C_1}$,

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \min_{\boldsymbol{\beta} \in B_0(s)} \left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + C_2 \sigma^2 \frac{|\mathrm{supp}(\boldsymbol{\beta})| \log p}{n}\right] + \frac{16\sigma^2 (\Phi^{-1}(1-\delta))^2}{n}.$$

If, in addition, $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ *for some* $\boldsymbol{\beta}^* \in B_0(s)$*, then with probability at least* $1 - \delta - (p+1)e^{-n/C_1}$*,*

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_1 \le C_2\sigma\Big(s\sqrt{\frac{\log p}{n}} + \frac{(\Phi^{-1}(1-\delta))^2}{\sqrt{n\log p}}\Big).$$

Proof. We first plug the bound on μ_4 given by Proposition 6.3 into the oracle inequalities in deviation of Theorem 6.2. Then, we obtain resulting oracle inequalities that hold with probability $1 - \delta - e^{-n/C_1}$ for all $\boldsymbol{\beta} \in B_0(s)$. To finish the proof, it suffices to compare the definitions of λ in (6.4) and in (6.7), and notice that

$$\mathbb{P}(\max_{1\le j\le p} \|\mathbb{X}e_j\|_n \le 2) \ge 1 - pe^{-n/2}. \tag{6.8}$$

Indeed, the random variable $\zeta_j = \|\mathbb{X}e_j\|_n$ is a $1/\sqrt{n}$-Lipschitz function of the standard Gaussian vector in $\mathbb{R}^n$. Thus, by the Gaussian concentration inequality, cf., e.g., [5, Theorem 5.6], we get $\mathbb{P}(\zeta_j > 2) \le \mathbb{P}(\zeta_j - \mathbb{E}(\zeta_j) > 1) \le e^{-n/2}$, where we have used that $\mathbb{E}(\zeta_j) \le (\mathbb{E}(\zeta_j^2))^{1/2} = 1$. This and the union bound yield (6.8). ∎

6.2 Group LASSO

We consider here $H = \mathbb{B} = \mathbb{R}^p$ equipped with the Euclidean norm $\|\cdot\|_H = |\cdot|_2$ and define the regularization norm $\|\cdot\|$ as follows. Let $G_1, \dots, G_M$ be a partition of $\{1, \dots, p\}$. For any $\boldsymbol{\beta} \in \mathbb{R}^p$ we set

$$\|\boldsymbol{\beta}\| = |\boldsymbol{\beta}|_{2,1} \triangleq \sum_{k=1}^{M} |\boldsymbol{\beta}_{G_k}|_2. \tag{6.9}$$

The group LASSO estimator is a solution of the convex minimization problem

$$\hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\|\mathbf{y} - \mathbb{X}\boldsymbol{\beta}\|_n^2 + \lambda \sum_{k=1}^{M} |\boldsymbol{\beta}_{G_k}|_2 \right), \tag{6.10}$$

where $\lambda > 0$ is a tuning parameter. In the following, we assume that the groups G_k have the same cardinality $|G_k| = T = p/M$, $k = 1, \dots, M$.

To any $\boldsymbol{\beta} \in \mathbb{R}^p$, we associate the set

$$\mathcal{K}(\boldsymbol{\beta}) = \{k \in \{1, ..., M\} : \boldsymbol{\beta}_{G_k} \neq \mathbf{0}\},$$

which plays the role of "support by block" of vector $\boldsymbol{\beta}$. Given $\boldsymbol{\beta} \in \mathbb{R}^p$, it is straightforward to check that Assumption 4.1 is satisfied when $\mathcal{P}_{\boldsymbol{\beta}}$ is the orthogonal projection operator onto the linear span of $\{e_j : j \in \cup_{k\in\mathcal{K}(\boldsymbol{\beta})} G_k\}$.

The operator $\mathbb{X}$ is a matrix in $\mathbb{R}^{n\times p}$ while the event Ω in Theorem 4.3 takes now the form

$$\Omega = \left\{ \max_{k=1,\dots,M} \frac{1}{n} |\mathbb{X}_{G_k}^T \boldsymbol{\xi}|_2 \le \tau' \right\} \tag{6.11}$$

where $\mathbb{X}_{G_k}$ is the $n \times |G_k|$ submatrix of $\mathbb{X}$ composed from all the columns of $\mathbb{X}$ with indices in G_k.

In order to apply Theorem 4.3, we need to find τ' such that $\mathbb{P}(\Omega) \ge 1/2$. Denote by $\|\mathbb{X}_{G_k}\|_{sp} \triangleq \sup_{|\boldsymbol{x}|_2\le 1} |\mathbb{X}_{G_k}\boldsymbol{x}|_2$ the spectral norm of matrix $\mathbb{X}_{G_k}$ and by $\|\mathbb{X}_{G_k}\|_{Fr}$ its Frobenius norm. Then, set $\psi_{sp}^* = \max_{k=1,\dots,M} \|\mathbb{X}_{G_k}\|_{sp}/\sqrt{n}$ and $\psi_{Fr}^* = \max_{k=1,\dots,M} \|\mathbb{X}_{G_k}\|_{Fr}/\sqrt{n}$. The constant τ' is determined by the following straightforward modification of Lemma 2 in [3].

Lemma 6.5. *Let* $\mathbb{X}$ *be deterministic. If*

$$\tau' \ge \frac{\sigma}{\sqrt{n}} \left(\psi_{Fr}^* + \psi_{sp}^* \sqrt{2\log(2M)} \right), \tag{6.12}$$

then the event (6.11) *has probability at least 1/2.*

Using this lemma, oracle inequalities for the group LASSO estimator with tuning parameter λ satisfying

$$\lambda \geq \frac{10\sigma}{\sqrt{n}}\left(\psi^*_{Fr} + \psi^*_{sp}\sqrt{2\log(2M)}\right) \tag{6.13}$$

can be deduced from Theorem 4.3 and Corollary 4.4. They have the following form.

Theorem 6.6. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that $\mathbb{X}$ is deterministic. Let $\delta \in (0,1)$. The group LASSO estimator $\hat{\boldsymbol{\beta}}$ with tuning parameter satisfying (6.13) is such that, with probability at least $1-\delta$,*

$$\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(\boldsymbol{\beta})\right] + \frac{16\sigma^2(\Phi^{-1}(1-\delta))^2}{n}$$

and

$$\mathbb{E}\|\mathbb{X}\hat{\boldsymbol{\beta}} - \mathbf{f}\|_n^2 \leq \min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[\|\mathbb{X}\boldsymbol{\beta} - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(\boldsymbol{\beta})\right] + \frac{16\sigma^2}{n}.$$

If, in addition, $\mathbf{f} = \mathbb{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$, then with probability at least $1-\delta$,

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_{2,1} \leq 4\lambda\mu_4^2(\boldsymbol{\beta}^*) + \frac{20\sigma^2(\Phi^{-1}(1-\delta))^2}{n\lambda}$$

and

$$\mathbb{E}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|_{2,1} \leq 8\lambda\mu_4^2(\boldsymbol{\beta}^*) + \frac{20\sigma}{\lambda n}.$$

Consider now a control of parameter $\mu_{c_0}(\boldsymbol{\beta})$ for vectors $\boldsymbol{\beta}$ with a "sparse by block" structure. To that end, one can use the "group" analog of the Restricted Eigenvalue constant, cf. [18]. For any $S \subset \{1, ..., M\}$ and $c_0 > 0$, we define the group Restricted Eigenvalue constant $\kappa_G(S, c_0) \geq 0$ by the formula

$$\kappa_G^2(S, c_0) \triangleq \min_{\boldsymbol{\delta}\in\mathcal{C}(S,c_0)} \frac{\|\mathbb{X}\boldsymbol{\delta}\|_n^2}{|\boldsymbol{\delta}|_2^2}, \tag{6.14}$$

where $\mathcal{C}(S, c_0)$ is the cone

$$\mathcal{C}(S, c_0) \triangleq \left\{\boldsymbol{\delta} \in \mathbb{R}^p \setminus \{0\} : \sum_{k\in S^c} |\boldsymbol{\delta}_{G_k}|_2 \leq c_0 \sum_{k\in S} |\boldsymbol{\delta}_{G_k}|_2\right\}.$$

In particular, for all $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\kappa_G(\mathcal{K}(\boldsymbol{\beta}), c_0) \neq 0$ we have

$$\mu_{c_0}^2(\boldsymbol{\beta}) \leq \frac{|\mathcal{K}(\boldsymbol{\beta})|}{\kappa_G^2(\mathcal{K}(\boldsymbol{\beta}), c_0)}.$$

When $\mathbb{X}$ is deterministic and $\boldsymbol{\beta}$ is such that $|\mathcal{K}(\boldsymbol{\beta})| \leq s$ sufficient conditions on $\mathbb{X}$ allowing one to bound $\kappa_G^2(\mathrm{supp}(\boldsymbol{\beta}), c_0)$ from below by a universal constant can be found in [18]. This leads to the bound $\mu_{c_0}^2(\boldsymbol{\beta}) \lesssim s$ for all vectors $\boldsymbol{\beta}$ such that $|\mathcal{K}(\boldsymbol{\beta})| \leq s$ (i.e., all *s-sparse by block vectors*).

Finally, we give an upper bound on $\mu_{c_0}(\boldsymbol{\beta})$ in the case of random $\mathbb{X}$. Let $X_1, \ldots, X_n$ be i.i.d. realizations of a random vector X with values in $\mathbb{R}^p$. The following proposition shows that, with high probability (with respect to the distribution of $X_1, \ldots, X_n$), we have $\mu_{c_0}^2(\boldsymbol{\beta}) \lesssim |\mathcal{K}(\boldsymbol{\beta})|$ for all s-sparse by block vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ provided that $n \gtrsim s(T + \log(M/s))$.

Proposition 6.7. *Let $L \geq 1$. Let $X_1, \ldots, X_n$ be i.i.d. realizations of a random vector X with values in $\mathbb{R}^p$ such that*

(i) X is isotropic (i.e., $\mathbb{E}XX^\top = I_{p\times p}$),

(ii) X is L-subgaussian: $\mathbb{E}\exp\left(t\langle X, \boldsymbol{\beta}\rangle\right) \leq \exp(L^2t^2/2)$ for all $t > 0$ and all $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\boldsymbol{\beta}|_2 = 1$.

Let $s \in \{1, \dots, M\}$ and $c_0 > 0$. There exist positive constants $C(L)$ and $C'(L)$ depending only on L such that the following holds. If

$$n \geq C(L)(1+c_0)^2 s\,(T + \log(M/s)) \tag{6.15}$$

then with probability greater than $1 - \exp(-C'(L)n)$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $|\mathcal{K}(\boldsymbol{\beta})| \leq s$ we have

$$\mu_{c_0}(\boldsymbol{\beta}) \leq 32L^2\sqrt{|\mathcal{K}(\boldsymbol{\beta})|}.$$

Proof. Since X is L-subgaussian and isotropic, it follows from Proposition 5.6 that X satisfies the small ball assumption with parameters β_0, κ_0 defined in (5.3).

The definition of $\|\cdot\|$ in (6.9) and the fact that $\mathcal{P}_{\boldsymbol{\beta}}$ is the projection operator onto the linear span of $\{e_j : j \in \cup_{k\in\mathcal{K}(\boldsymbol{\beta})} G_k\}$ imply

$$\begin{aligned}\mathbb{C}_{\boldsymbol{\beta},c_0} &= \left\{\boldsymbol{\beta}' \in \mathbb{R}^p : \left\|\mathcal{P}_{\boldsymbol{\beta}}^{\perp}\boldsymbol{\beta}'\right\| \leq c_0 \left\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{\beta}'\right\|\right\} \\ &= \left\{\boldsymbol{\beta}' \in \mathbb{R}^p : \sum_{k\in\mathcal{K}(\boldsymbol{\beta})^c} |\boldsymbol{\beta}'_{G_k}|_2 \leq c_0 \sum_{k\in\mathcal{K}(\boldsymbol{\beta})} |\boldsymbol{\beta}'_{G_k}|_2\right\}.\end{aligned}$$

Denote by Σ_s the set of all vectors $\boldsymbol{\beta}$ in $\mathbb{R}^p$ such that $|\mathcal{K}(\boldsymbol{\beta})| \leq s$. It follows from Theorem 5.3 and (5.7) that, if (5.4) holds with $\mathbb{A} = \Sigma_s$, then with probability at least $1 - \exp(-n\kappa_0^2/32)$, for all $\boldsymbol{\beta} \in \Sigma_s$ we have

$$\mu_{c_0}(\boldsymbol{\beta}) \leq \left(\frac{8}{\beta_0^2\kappa_0}\right)^{1/2} \sup_{\boldsymbol{\beta}'\in\mathbb{C}_{\boldsymbol{\beta},c_0}} \frac{\left\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{\beta}'\right\|}{|\boldsymbol{\beta}'|_2} \leq \sqrt{\frac{8|\mathcal{K}(\boldsymbol{\beta})|}{\beta_0^2\kappa_0}} = 32L^2\sqrt{|\mathcal{K}(\boldsymbol{\beta})|}$$

since $\left\|\mathcal{P}_{\boldsymbol{\beta}}\boldsymbol{\beta}'\right\| \leq \sqrt{|\mathcal{K}(\boldsymbol{\beta})|}|\boldsymbol{\beta}'|_2$ for all $\boldsymbol{\beta}' \in \mathbb{R}^p$.

Therefore, it only remains to prove that (6.15) implies (5.4) with $\mathbb{A} = \Sigma_s$. Denote by B the unit ball with respect to the group LASSO norm $\|\cdot\|$ in $\mathbb{R}^p$. It is straightforward to check that $S_2^{p-1} \cap (\cup_{\boldsymbol{\beta}\in\Sigma_s}\mathbb{C}_{\boldsymbol{\beta},c_0}) \subset \mathbb{C}$ where $\mathbb{C} = ((1+c_0)\sqrt{s}B) \cap B_2^p$. By Proposition 5.6, we have, for an absolute constant $c_2 > 0$,

$$\mathbb{E}\sup_{\boldsymbol{\beta}\in S_2^{p-1}\cap\left(\cup_{\boldsymbol{\beta}\in\Sigma_s}\mathbb{C}_{\boldsymbol{\beta},c_0}\right)} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle X_i, \boldsymbol{\beta}\rangle\right| \leq \mathbb{E}\sup_{\boldsymbol{\beta}\in\mathbb{C}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle X_i, \boldsymbol{\beta}\rangle\right| \leq \frac{c_2 L\ell^*(\mathbb{C})}{\sqrt{n}}.$$

We now bound $\ell^*(\mathbb{C})$ from above. First note that $\ell^*(\mathbb{C}) \leq (1+c_0)\ell^*(\sqrt{s}B \cap B_2^p)$. Denote by $\boldsymbol{\eta} = (\eta_j)_{j=1}^p$ a standard Gaussian vector in $\mathbb{R}^p$ and by $(\zeta_k^\sharp)_{k=1}^M$ a non-increasing rearrangement of $(|\boldsymbol{\eta}_{G_k}|_2)_{k=1}^M$. We have

$$\begin{aligned}\ell^*(\sqrt{s}B \cap B_2^p) &= \mathbb{E}\sup\left(\sum_{k=1}^M t_k\zeta_k : \sum_{k=1}^M |t_k| \leq \sqrt{s}, \sum_{k=1}^M t_k^2 \leq 1\right) \\ &\leq \mathbb{E}\sup\left(\sum_{k=1}^M t_k^\sharp\zeta_k^\sharp : \sum_{k=1}^M |t_k| \leq \sqrt{s}, \sum_{k=1}^M t_k^2 \leq 1\right) \\ &\leq \mathbb{E}\sup\left(\sum_{k=1}^s t_k^\sharp\zeta_k^\sharp : \sum_{k=1}^M t_k^2 \leq 1\right) + \mathbb{E}\sup\left(\sum_{k=s+1}^M t_k^\sharp\zeta_k^\sharp : \sum_{k=1}^M |t_k| \leq \sqrt{s}\right) \\ &= \mathbb{E}\sqrt{\sum_{k=1}^s (\zeta_k^\sharp)^2} + \sqrt{s}\mathbb{E}\max_{k=s+1,\dots,M}\zeta_k^\sharp \leq 2\mathbb{E}\sqrt{\sum_{k=1}^s (\zeta_k^\sharp)^2} \\ &\leq \frac{2\sqrt{8s}}{\sqrt{3}}\left[\mathbb{E}\left(\frac{1}{s}\sum_{k=1}^s \frac{3(\zeta_k^\sharp)^2}{8}\right)\right]^{1/2}.\end{aligned}$$

Then, using Jensen's inequality we obtain

$$\mathbb{E}\left(\frac{1}{s}\sum_{k=1}^{s}\frac{3(\zeta_k^\sharp)^2}{8}\right) \le \log\mathbb{E}\exp\left(\frac{1}{s}\sum_{k=1}^{s}\frac{3(\zeta_k^\sharp)^2}{8}\right)$$
$$\le \log\mathbb{E}\exp\left(\frac{1}{s}\sum_{k=1}^{M}\frac{3|\boldsymbol{\eta}_{G_k}|_2^2}{8}\right) \le \log\left(\frac{1}{s}\sum_{k=1}^{M}\mathbb{E}\exp\left(\frac{3|\boldsymbol{\eta}_{G_k}|_2^2}{8}\right)\right)$$
$$= \log\left(\frac{1}{s}\sum_{k=1}^{M}\prod_{j\in G_k}\mathbb{E}\exp\left(\frac{3\eta_j^2}{8}\right)\right) = \log\left(\frac{2^T M}{s}\right).$$

Therefore, there exists an absolute constant $c' > 0$ such that

$$\frac{c_2 L\ell^*(\mathbb{C})}{\sqrt{n}} \le \frac{c' L(1+c_0)\sqrt{sT + s\log(M/s)}}{\sqrt{n}}.$$

For β_0 and κ_0 given in (5.3), the expression in the last display can be rendered smaller than $\kappa_0\beta_0/16$ provided that (6.15) holds with large enough constant $C(L) > 0$ depending only on L. Thus, (5.4) follows. ∎

6.3 Nuclear norm penalty

We consider here $H = \mathbb{B} = \mathbb{R}^{k\times m}$ equipped with the Frobenius norm $\|\cdot\|_H = \|\cdot\|_{Fr}$ and we define the regularization norm $\|\cdot\|$ as the nuclear norm $\|\cdot\|_*$ (i.e., the sum of the singular values). The corresponding penalized LS estimator $\hat{A}$ is a solution of the minimization problem

$$\hat{A} \in \operatorname*{argmin}_{A\in\mathbb{R}^{k\times m}}\left(\|\mathbb{X}A - \mathbf{y}\|_n^2 + \lambda\|A\|_*\right) \tag{6.16}$$

where $\lambda > 0$ is a tuning parameter. Penalized LS estimators with nuclear norm penalty were considered in several papers starting from [25, 6, 13]. For more references, see [12, 11, 28].

For $A \in \mathbb{R}^{k\times m}$, let $r = \operatorname{rank}(A)$ denote its rank. By the singular value decomposition, $A = \sum_{j=1}^{r}\sigma_j(A)u_j v_j^\top$ with orthonormal vectors $u_1,\dots,u_r \in \mathbb{R}^k$, orthonormal vectors $v_1,\dots,v_r \in \mathbb{R}^m$ and singular values $\sigma_1(A) \ge \cdots \ge \sigma_r(A) > 0$. The pair of linear vector spaces (S_1, S_2) where S_1 is the linear span of $\{u_1,\dots,u_r\}$ and S_2 is the linear span of $\{v_1,\dots,v_r\}$ will be called the support of A. We will denote by $S_j^\perp$ the orthogonal complement of S_j, $j = 1, 2$, and by P_S the orthogonal projector on the linear vector space S. Given $A \in \mathbb{R}^{k\times m}$ with support (S_1, S_2), and $B \in \mathbb{R}^{k\times m}$, we set

$$\mathcal{P}_A(B) \triangleq B - P_{S_1^\perp}BP_{S_2^\perp} \text{ and } \mathcal{P}_A^\perp(B) \triangleq P_{S_1^\perp}BP_{S_2^\perp}. \tag{6.17}$$

For the norm $\|\cdot\| = \|\cdot\|_*$, Assumption 4.1 is satisfied with the operator $\mathcal{P}_A$ defined in (6.17). Indeed, it is clear that $\mathcal{P}_A(A) = A$. Furthermore, by definition of $\mathcal{P}_A^\perp$, the columns of A are orthogonal to the columns of $\mathcal{P}_A^\perp(B)$ and the rows of A are orthogonal to the rows of $\mathcal{P}_A^\perp(B)$. Thus

$$\|A\|_* + \|\mathcal{P}_A^\perp(B)\|_* = \|A + \mathcal{P}_A^\perp(B)\|_*,$$

which means that the nuclear norm satisfies Assumption 4.2 (the decomposability assumption), and a fortiori Assumption 4.1.

Oracle inequalities for the estimator (6.16) follow from Theorem 4.3 and Corollary 4.4. In order to apply those results, one has to find τ' such that $\mathbb{P}(\Omega) \ge 1/2$ where

$$\Omega = \left\{\sup_{\|B\|_*\le 1}\frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X_i, B\rangle \le \tau'\right\} = \left\{\frac{\|\Gamma\|_{sp}}{\sqrt{n}} \le \tau'\right\}, \tag{6.18}$$

$\Gamma = n^{-1/2}\sum_{i=1}^n \xi_i X_i$, $\|\Gamma\|_{sp}$ is the spectral norm (i.e., the largest singular value of Γ), and ξ_i are i.i.d. random variables with distribution $\mathcal{N}(0,\sigma^2)$. The next result from [25] provides a control of the spectral norm of Γ. Define

$$\phi_{max} \triangleq \sup_{\substack{A\in\mathbb{R}^{k\times m}:\|A\|_{Fr}=1 \\ \text{and } \operatorname{rank}(A)=1}} \left(\frac{1}{n}\sum_{i=1}^n \langle X_i, A\rangle^2\right)^{1/2}.$$

The quantity ϕ_{max} is the maximal rank-1 restricted eigenvalue of the operator $\mathbb{X}$.

Lemma 6.8 (Lemma 2 in [25] with $D=2$)**.** *Let $k \geq 2$ and $m \geq 2$. Let $X_1, \ldots, X_n$ be deterministic matrices in $\mathbb{R}^{k\times m}$ and let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables with distribution $\mathcal{N}(0,\sigma^2)$. If*

$$\tau' \geq 8\sigma\phi_{max}\sqrt{\frac{k+m}{n}}$$

then for the event Ω in (6.18) we have $\mathbb{P}(\Omega) \geq 1 - 2\exp(-(2-\log 5)(m+k)) \geq 1/2$.

In view of this lemma, oracle inequalities for the nuclear norm regularization procedure (6.16) with tuning parameter λ satisfying

$$\lambda \geq 80\sigma\phi_{max}\sqrt{\frac{k+m}{n}} \tag{6.19}$$

can be deduced from Theorem 4.3 and Corollary 4.4. We have the following result.

Theorem 6.9. *Let $k \geq 2$ and $m \geq 2$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that $X_1, \ldots, X_n$ are deterministic matrices. Let $\delta \in (0,1)$. The estimator $\hat{A}$ defined in (6.16) with tuning parameter satisfying (6.19) is such that, with probability at least $1-\delta$,*

$$\|\mathbb{X}\hat{A} - \mathbf{f}\|_n^2 \leq \min_{A\in\mathbb{R}^{k\times m}} \left[\|\mathbb{X}A - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(A)\right] + \frac{16\sigma^2(\Phi^{-1}(1-\delta))^2}{n}$$

and

$$\mathbb{E}\|\mathbb{X}\hat{A} - \mathbf{f}\|_n^2 \leq \min_{A\in\mathbb{R}^{k\times m}} \left[\|\mathbb{X}A - \mathbf{f}\|_n^2 + \frac{16}{25}\lambda^2\mu_4^2(A)\right] + \frac{16\sigma^2}{n}.$$

If, in addition, $\mathbf{f} = \mathbb{X}A^*$ for some $A^* \in \mathbb{R}^{k\times m}$, then with probability at least $1-\delta$,

$$\|\hat{A} - A^*\|_* \leq 4\lambda\mu_4^2(A^*) + \frac{20\sigma^2(\Phi^{-1}(1-\delta))^2}{n\lambda}$$

and

$$\mathbb{E}\|\hat{A} - A^*\|_* \leq 8\lambda\mu_4^2(A^*) + \frac{20\sigma}{\lambda n}.$$

Finally, we give a bound on the compatibility factor $\mu_{c_0}(A)$ for low rank matrices A in the case where $X_1, \ldots, X_n$ are i.i.d. random matrices. Using Theorem 5.3 we obtain the following result.

Proposition 6.10. *Let $L \geq 1$. Let $X_1, \ldots, X_n$ be i.i.d. realizations of a random matrix X with values in $\mathbb{R}^{k\times m}$ such that*

(i) X is isotropic: $\mathbb{E}\langle X, A\rangle^2 = \|A\|_{Fr}^2$ for all $A \in \mathbb{R}^{k\times m}$,

(ii) X is L-subgaussian: $\mathbb{E}\exp\left(t\langle X, A\rangle\right) \leq \exp(L^2t^2/2)$ for all $t > 0$ and all $A \in \mathbb{R}^{k\times m}$ such that $\|A\|_{Fr} = 1$.

Let $s \in \{1, \ldots, \min(k,m)\}$ and $c_0 > 0$. There exist positive constants $c(L)$ and $c'(L)$ depending only on L such that the following holds. If

$$n \geq c(L)(1+c_0)^2 s\max(k,m), \tag{6.20}$$

then with probability greater than $1-\exp(-c'(L)n)$, for any $A \in \mathbb{R}^{k\times m}$ such that $\operatorname{rank}(A) \leq s$ we have

$$\mu_{c_0}(A) \leq 32L^2\sqrt{\operatorname{rank}(A)}.$$

Proof. Since X is L-subgaussian and isotropic, it follows from Proposition 5.6 that X satisfies the small ball assumption with parameters β_0, κ_0 defined in (5.3).

Denote by M_s the set of all matrices in $\mathbb{R}^{k\times m}$ with rank at most s. For any $A \in \mathbb{R}^{k\times m}$ we have

$$\mathbb{C}_{A,c_0} = \left\{A' \in \mathbb{R}^{k\times m} : \left\|\mathcal{P}_A^\perp A'\right\|_* \le c_0 \left\|\mathcal{P}_A A'\right\|_*\right\}$$

where $\mathcal{P}_A$ is the operator defined in (6.17). It follows from Theorem 5.3 and (5.7) that, if (5.4) holds with $\mathbb{A} = M_s$, then with probability at least $1 - \exp(-n\kappa_0^2/32)$, for all $A \in M_s$ we have

$$\mu_{c_0}(A) \le \left(\frac{8}{\beta_0^2\kappa_0}\right)^{1/2} \sup_{A'\in\mathbb{C}_{A,c_0}} \frac{\|\mathcal{P}_A A'\|_*}{\|A'\|_{Fr}} \le \sqrt{\frac{8\mathrm{rank}(A)}{\beta_0^2\kappa_0}} = 32L^2\sqrt{\mathrm{rank}(A)}$$

since $\|\mathcal{P}_A A'\|_* \le \sqrt{\mathrm{rank}(A)}\|A'\|_{Fr}$ for all $A' \in \mathbb{R}^{k\times m}$.

Therefore, it only remains to prove that (6.20) implies (5.4) with $\mathbb{A} = M_s$. Denote by S_2^{km-1} and B_2^{km} the unit Euclidean sphere and the unit Euclidean ball in $\mathbb{R}^{k\times m}$, respectively, and by B_* the unit ball in $\mathbb{R}^{k\times m}$ with respect to the nuclear norm. It is straightforward to check that $S_2^{km-1} \cap (\cup_{A\in M_s}\mathbb{C}_{A,c_0}) \subset \mathbb{C}$ where $\mathbb{C} = ((1+c_0)\sqrt{s}B_*) \cap B_2^{km}$. Proposition 5.6 yields that

$$\begin{aligned}\mathbb{E}\sup_{A\in S_2^{km-1}\cap\left(\cup_{A\in M_s}\mathbb{C}_{A,c_0}\right)} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\langle X_i, A\rangle\right| &\le c_2 L\ell^*\left(S_2^{km-1}\cap(\cup_{A\in M_s}\mathbb{C}_{A,c_0})\right)\\ &\le c_2 L\ell^*(\mathbb{C}).\end{aligned}$$

Next, by inclusion we have $\ell^*(\mathbb{C}) \le (1+c_0)\ell^*(\sqrt{s}B_*\cap B_2^{km}) \le (1+c_0)\sqrt{s}\ell^*(B_*)$. By duality, $\ell^*(B_*) = \mathbb{E}\|G\|_{sp}$ where G is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. In addition, $\mathbb{E}\|G\|_{sp} \le \sqrt{k}+\sqrt{m}$, cf. [7]. Thus, for large enough constant $c(L) > 0$, condition (6.20) implies (5.4) with $\mathbb{A} = M_s$. ∎

Combining Theorem 6.9 and Proposition 6.10 we can obtain oracle inequalities for the estimator $\hat{A}$ defined in (6.16) when $X_1, \dots, X_n$ are i.i.d. random matrices independent of the noise vector $\boldsymbol{\xi}$. We illustrate it by the following result for the basic example where the entries of each of the matrices X_i are i.i.d. standard Gaussian.

Theorem 6.11. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that $X_1, \dots, X_n$ are i.i.d. realizations of a random matrix X whose entries are i.i.d. standard Gaussian random variables. We also assume that $X_1, \dots, X_n$ are independent of the noise vector $\boldsymbol{\xi}$. Let $\delta \in (0,1)$, $k \ge 2$, $m \ge 2$, and*

$$\lambda = a\sigma\sqrt{\frac{k+m}{n}} \tag{6.21}$$

with $a \ge 120$. There exist an absolute constant $C_3 > 0$ and a constant $C_4 > 0$ depending only on a such that the following holds. If $n \ge C_3 s\max(k,m)$, then for the estimator $\hat{A}$ defined in (6.16) with tuning parameter (6.21) we have that, with probability at least $1 - \delta - e^{-n/C_4}$,

$$\begin{aligned}\|\mathbb{X}\hat{A} - \mathbf{f}\|_n^2 \quad \le \quad &\min_{A\in\mathbb{R}^{k\times m}:\mathrm{rank}(A)\le s}\left(\|\mathbb{X}A - \mathbf{f}\|_n^2 + C_4\frac{\sigma^2\mathrm{rank}(A)(k+m)}{n}\right)\\ &+ \frac{16\sigma^2(\Phi^{-1}(1-\delta))^2}{n}.\end{aligned}$$

If, in addition, $\mathbf{f} = \mathbb{X}A^*$ for some $A^* \in \mathbb{R}^{k\times m}$ such that $\mathrm{rank}(A^*) \le s$, then with probability at least $1 - \delta - e^{-n/C_4}$,

$$\|\hat{A} - A^*\|_* \le C_4\sigma\left(s\sqrt{\frac{k+m}{n}} + \frac{(\Phi^{-1}(1-\delta))^2}{\sqrt{(k+m)n}}\right).$$

Proof. Under the assumptions of the theorem, $\mathbb{X}$ is a nearly isometric linear map, cf. [24]. Then, it follows from [24, Lemma 4.3] that there exists an absolute constant $C_5 > 0$ such that $\phi_{max} \leq 3/2$ with probability at least $1 - e^{-n/C_5}$. Therefore, we can use Theorem 6.9 with λ defined in (6.21). Plugging the bound on μ_4 from Proposition 6.10 in the oracle inequalities in deviation from Theorem 6.9 we obtain the result. ■

References

[1] P. C. Bellec, A. S. Dalalyan, E. Grappin, and Q. Paris. On the prediction loss of the Lasso in the partially labeled setting. *arXiv preprint arXiv:1606.06179*, 2016.

[2] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016.

[3] P. C. Bellec and A. B. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In V. Panov, editor, *Festschrift for Valentin Konakov*. Springer, 2017, *arXiv preprint arXiv:1609.06675*.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.

[6] E. Candes and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a mininal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57:2342–2359, 2011.

[7] K.R. Davidson and S.J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces*, volume I, pages 317–366. North-Holland, Amsterdam, 2001.

[8] V.H. de la Peña and E. Giné. *Decoupling*. Springer, New York, 1999.

[9] S. Dirksen, G. Lecué, and H. Rauhut. On the gap between RIP-properties and sparse recovery conditions. *To appear in IEEE Transactions on Information Theory*, 2015.

[10] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002.

[11] C. Giraud. *Introduction to High-dimensional Statistics*. CRC Press, 2014.

[12] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Math.* Springer, Heidelberg, 2011.

[13] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39:2302–2329, 2011.

[14] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN*, 23:12991–13008, 2015.

[15] G. Lecué and S. Mendelson. Regularization and the small ball method I: sparse recovery. Technical report, CNRS and Technion, 2016.

[16] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, Berlin, 2011.

[17] M. Lifshits. *Lectures on Gaussian Processes*. Springer, Heidelberg, 2012.

[18] K. Lounici, M. Pontil, A.B. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39:2164 – 2204, 2011.

[19] S. 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, 2014.

[20] S. Mendelson. Learning without concentration. *J. ACM*, 62:Art. 21, 25, 2015.

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

[22] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 4:538—557, 2012.

[23] J. Peypouquet. *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. Springer, New York, 2015.

[24] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, 2010.

[25] A. Rohde and A.B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39:887–930, 2011.

[26] M. Talagrand. *Upper and lower bounds for stochastic processes*. Springer, Heidelberg, 2014.

[27] J. Taylor. The geometry of least squares in the 21st century. *Bernoulli*, 19:1449–1464, 2013.

[28] S. van de Geer. *Estimation and Testing under Sparsity*. Springer, Heidelberg, 2016.

[29] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, New York, 1996.

[30] R. van Handel. Probability in high dimension. Technical report, Princeton University, 2014.