High-dimensional Statistical Methods

Skoltech

Q. Paris[*]

National Research University HSE
Faculty of Computer Science
Moscow, Russia

Chapter 1. High-dimensional regression

Lecture 3

Penalized least squares and optimality

[*] email:qparis@hse.ru, teaching material: http://www.qparis-math.com/teaching.

The material presented below is borrowed or inspired from Sun and Zhang (2012); Rigollet (2015); Giraud (2015) and Dalalyan et al. (2017).

1 Bayes information Criterion

For any $\lambda > 0$, the BIC estimator of $\boldsymbol{\beta}^\star$ is defined as any

$$\hat{\boldsymbol{\beta}}^{\text{bic}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbf{R}^p} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda^2 \|\boldsymbol{\beta}\|_0 \right\}.$$

Theorem 1.1. *Suppose that the noise vector $\boldsymbol{\xi} \in \mathbf{R}^n$ is sub-gaussian with variance proxy $\sigma^2 > 0$. Suppose the parameter λ of the BIC estimator is chosen as*

$$\lambda^2 = 16 \{\log(6) + 2\log(ep)\} \frac{\sigma^2}{n}.$$

Then, for all $n \geq 1$ and all $\delta \in (0, 1)$,

$$\mathcal{E}(\hat{\mu}^{\text{bic}}) \leq C \frac{\sigma^2 \|\boldsymbol{\beta}^\star\|_0}{n} \log\left(\frac{ep}{\delta}\right),$$

with probability at least $1 - \delta$, where $C \leq 201$.

Proof of Theorem 1.1. See the lectures notes from Rigollet (2015). □

2 Lasso

For any $\lambda > 0$, the Lasso estimator of $\boldsymbol{\beta}^\star$ is defined as any

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbf{R}^p} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + 2\lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

2.1 Slow rates

We first investigate the performance of the Lasso estimator when making no assumption on the design matrix $\mathbf{X}$.

Theorem 2.1. *Suppose that the noise vector* $\boldsymbol{\xi} \in \mathbf{R}^n$ *is sub-gaussian with variance proxy* $\sigma^2 > 0$. *Denote* $\varkappa = \max\{\|\mathbf{x}^j\|_2 : 1 \le j \le p\}$. *Fix* $\delta \in (0,1)$ *and suppose the parameter* λ *of the Lasso estimator is chosen such that*

$$\lambda \ge \frac{\sigma\varkappa}{n}\sqrt{2\log\left(\frac{2p}{\delta}\right)}.$$

Then, for all $n \ge 1$,

$$\mathcal{E}(\hat{\mu}^{\text{lasso}}) \le 4\lambda\|\boldsymbol{\beta}^\star\|_1,$$

with probability at least $1-\delta$.

Proof of Theorem 2.1. Seen in class. □

When the columns $\mathbf{x}^j$ of $\mathbf{X}$ are renormalized in such a way that $\varkappa \le \sqrt{n}$, it follows from the previous result that the choice of

$$\lambda = \sigma\sqrt{\frac{2}{n}\log\left(\frac{2p}{\delta}\right)}$$

yields the upper bound

$$\mathcal{E}(\hat{\mu}^{\text{lasso}}) \le 4\sigma\|\boldsymbol{\beta}^\star\|_1\sqrt{\frac{2}{n}\log\left(\frac{2p}{\delta}\right)},$$

holding with probability at least $1-\delta$. This upper bound is usually referred to as a slow rate of convergence, by comparison to the rate obtained in the context of the BIC. In addition, a few other specificities are to be pointed out here. In particular, contrary to the case of the BIC, the smoothing parameter λ has here to be adapted to the confidence level required δ. Finally, note that in this result, the Lasso estimator adapts to the unknown ℓ_1 norm of $\boldsymbol{\beta}^\star$ and not its sparsity level. The result presented in the next paragraph addresses in particular this issue.

2.2 Fast rates

In this paragraph, we present an improved performance bound for the Lasso estimator obtained at the price of some assumptions on the structure of the design matrix $\mathbf{X}$. First, we introduce some notation. For any vector $\boldsymbol{\beta} \in \mathbf{R}^p$ and any $J \subset \{1,\dots,p\} = [p]$, we denote $J^c = [p] \setminus J$ and $\boldsymbol{\beta}_J \in \mathbf{R}^p$ the vector whose coordinates of index $j \in J^c$ are set to 0. Then, for any $c > 0$ and any $J \subset [p]$, we define the restricted eigenvalue constant $\text{RE}(c,J)$ with parameters c and J

$$\text{RE}(c,J) := \inf\left\{\frac{\|\mathbf{X}\boldsymbol{\beta}\|_2^2}{n\|\boldsymbol{\beta}_J\|_2^2} : \beta \in \mathbf{R}^p,\ \|\boldsymbol{\beta}_{J^c}\|_1 < c\|\boldsymbol{\beta}_J\|_1\right\}.$$

We are now in position to state a first important result.

Theorem 2.2. *Fix $\lambda > 0$. Then, on the event*

$$\left\{ \frac{1}{n} \|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \leq \frac{\lambda}{2} \right\}, \tag{2.1}$$

the Lasso estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{lasso}}$, satisfies, for all $n \geq 1$,

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \leq \inf_{\boldsymbol{\beta} \in \mathbf{R}^p, J \subset [p]} \left\{ \frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2 + 4\lambda \|\boldsymbol{\beta}_{J^c}\|_1 + \frac{9}{4} \frac{\lambda^2 |J|}{\text{RE}(3, J)} \right\}.$$

Proof of Theorem 2.2. In the proof, we denote $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{lasso}}$ for brevity. First, let us denote $\mathcal{L} : \mathbf{R}^p \to \mathbf{R}$ the Lasso criterion

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + 2\lambda\|\boldsymbol{\beta}\|_1.$$

The function $\mathcal{L}$ is convex and we know from the properties of convex functions that $\hat{\beta}$ minimises $\mathcal{L}$ over $\mathbf{R}^p$ if and only if the null vector $\mathbf{0}$ belongs to its sub-gradient $\partial\mathcal{L}(\hat{\beta})$ at point $\hat{\beta}$ (see, for instance, Section 4.1 or Lemma D.4 in the book by Giraud, 2015). For all $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbf{R}^p$, we introduce the notation

$$\text{sign}(\boldsymbol{\beta}) = \begin{pmatrix} \text{sign}(\beta_1) \\ \vdots \\ \text{sign}(\beta_p) \end{pmatrix} \quad \text{where} \quad \forall j : \ \text{sign}(\beta_j) = \begin{cases} \{1\} & \text{if } \beta_j > 0, \\ [-1, 1] & \text{if } \beta_j = 0, \\ \{-1\} & \text{if } \beta_j < 0. \end{cases}$$

It has been proven in the seminar sessions that, for any $\boldsymbol{\beta} \in \mathbf{R}^p$, $\partial\|\boldsymbol{\beta}\|_1 = \text{sign}(\boldsymbol{\beta})$. Hence, it follows that, for all $\boldsymbol{\beta} \in \mathbf{R}^p$,

$$\partial\mathcal{L}(\boldsymbol{\beta}) = \frac{2}{n}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}) + 2\lambda\,\text{sign}(\boldsymbol{\beta}). \tag{2.2}$$

Therefore, the condition $\mathbf{0} \in \partial\mathcal{L}(\hat{\boldsymbol{\beta}})$ and equation (2.2) yield

$$\frac{1}{n}\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \in \lambda\,\text{sign}(\hat{\boldsymbol{\beta}}). \tag{2.3}$$

Considering the inner product with $\hat{\boldsymbol{\beta}}$ in equation (2.3), we obtain that

$$\frac{1}{n}\hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \lambda\|\hat{\boldsymbol{\beta}}\|_1. \tag{2.4}$$

Also, for all $\boldsymbol{\beta} \in \mathbf{R}^p$, it follows from (2.3) that

$$\frac{1}{n}\boldsymbol{\beta}^\top\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \in [-\lambda\|\boldsymbol{\beta}\|_1, \lambda\|\boldsymbol{\beta}\|_1]. \tag{2.5}$$

As a result, for any $\boldsymbol{\beta} \in \mathbf{R}^p$, substracting (2.4) to (2.5) we get that,

$$\frac{1}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \le \lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1). \tag{2.6}$$

Now by expanding $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi}$ on the left hand side of (2.6), we obtain that for any $\boldsymbol{\beta} \in \mathbf{R}^p$,

$$\begin{aligned}\frac{1}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) &\le \frac{1}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \boldsymbol{\xi} + \lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1)\\ &\le \frac{1}{n}\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + \lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1). \qquad (2.7)\end{aligned}$$

Lets now work on the event $\{\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty \le n\lambda/2\}$. On this event, we get that

$$\frac{1}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) \le \frac{\lambda}{2}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\hat{\boldsymbol{\beta}}\|_1\right). \tag{2.8}$$

Now, using the notation introduced before the Theorem, observe that for all $J \subset [p]$:

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\hat{\boldsymbol{\beta}}\|_1 &= \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1 + \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1 + 2\|\boldsymbol{\beta}_J\|_1\\ &\quad + 2\|\boldsymbol{\beta}_{J^c}\|_1 - 2\|\hat{\boldsymbol{\beta}}_J\|_1 - 2\|\hat{\boldsymbol{\beta}}_{J^c}\|_1. \qquad (2.9)\end{aligned}$$

From the inequalities $\|\boldsymbol{\beta}_J\|_1 - \|\hat{\boldsymbol{\beta}}_J\|_1 \le \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1$ and $\|\hat{\boldsymbol{\beta}}_{J^c}\|_1 \ge \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1 - \|\boldsymbol{\beta}_{J^c}\|_1$, we deduce from equation (2.9) that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 + 2\|\boldsymbol{\beta}\|_1 - 2\|\hat{\boldsymbol{\beta}}\|_1 \le 3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1 + 4\|\boldsymbol{\beta}_{J^c}\|_1. \tag{2.10}$$

Combining inequalities (2.8) and (2.10) yields that, on the event $\{\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty \le n\lambda/2\}$, and for any $J \subset [p]$ and any $\boldsymbol{\beta} \in \mathbf{R}^p$,

$$\begin{aligned}&\frac{1}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})\\ &\le \frac{\lambda}{2}\left\{3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1\right\} + 2\lambda\|\boldsymbol{\beta}_{J^c}\|_1. \qquad (2.11)\end{aligned}$$

Then, by definition of the restricted eigenvalue constant $\text{RE}(3, J)$, we deduce that, if $3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1 > \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1$,

$$\begin{aligned}3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1 &\le 3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1\\ &\le 3\sqrt{|J|}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_2\\ &\le 3\sqrt{\frac{|J|}{n\text{RE}(3, J)}}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2.\end{aligned}$$

Notice next that if $3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_J\|_1 \le \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{J^c}\|_1$ then the last inequality trivially holds. As a result, it follows from (2.11) that, for all $J \subset [p]$ and all

$\boldsymbol{\beta} \in \mathbf{R}^p$,

$$\frac{1}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) \leq \frac{3\lambda}{2}\sqrt{\frac{|J|}{n\mathrm{RE}(3, J)}}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 + 2\lambda\|\boldsymbol{\beta}_{J^c}\|_1. \tag{2.12}$$

To complete the proof, a few more tricks are in order. First, using in (2.12) the identity $2u^\top v = \|u\|_2^2 + \|v\|_2^2 - \|u-v\|_2^2$ with $u = \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ and $v = \mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})$, yields

$$\begin{aligned}\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \leq \frac{1}{2n}&\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2 + 2\lambda\|\boldsymbol{\beta}_{J^c}\|_1 \\ &+ \frac{3\lambda}{2}\sqrt{\frac{|J|}{n\mathrm{RE}(3, J)}}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 - \frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2. \end{aligned} \tag{2.13}$$

Finally, using the inequality $ax - bx^2 \leq a^2/4b$, we may upper bound the sum of the last two terms in (2.13) by

$$\frac{9}{8}\frac{|J|\lambda^2}{\mathrm{RE}(3, J)},$$

and the proof is complete. □

For the last result to be of more practical interest, it remain to upper bound the probability of the event defined in (2.1). In the context where the noise vector is sub-gaussian, this will follow from standard arguments.

Theorem 2.3. *Suppose that the noise vector $\boldsymbol{\xi} \in \mathbf{R}^n$ is sub-gaussian with variance proxy $\sigma^2 > 0$. Denote $\varkappa = \max\{\|\mathbf{x}^j\|_2 : 1 \leq j \leq p\}$ and fix $\delta \in (0, 1)$. Then, the following statements hold.*

(1) *For any*

$$\lambda \geq 2\frac{\sigma\varkappa}{n}\sqrt{2\log\left(\frac{2p}{\delta}\right)},$$

the event (2.1) has probability at least $1 - \delta$.

(2) *Suppose that $\varkappa \leq \sqrt{n}$ and that*

$$\lambda \geq 2\sigma\sqrt{\frac{2}{n}\log\left(\frac{2p}{\delta}\right)}.$$

Then, choosing $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ and J equal to the support $J^\star$ of $\boldsymbol{\beta}^\star$ in the inequality of the previous Theorem, we deduce that the Lasso estimator $\hat{\boldsymbol{\beta}}$ with parameter λ satisfies, for all $n \geq 1$,

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \leq \frac{18}{\mathrm{RE}(3, J^\star)}\frac{\sigma^2\|\boldsymbol{\beta}^\star\|_0}{n}\log\left(\frac{2p}{\delta}\right),$$

with probability at least $1 - \delta$.

Proof of Theorem 2.3. Seen in class. □

The reader may notice that, up to the factor $18\textsc{re}(3, J^\star)^{-1}$, the previous upper bound is of exact same order as the BIC estimator. In particular, provided the constant $\textsc{re}(3, J^\star)$ is bounded from below independently of n, the Lasso possesses the exact same performance as the idealized BIC. The next issue will be to understand when can the restricted eigenvalue constant be lower bounded independently of n.

2.3 On the restricted eigenvalue constant

This paragraph exposes a simple sufficient condition under which the restricted eigenvalue condition can be lower bounded independently of n.

Definition 2.1. *For $\varepsilon > 0$, we'll say that the design matrix $\mathbf{X}$ is ε-incoherent if*

$$\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_p\|_\infty \leq \varepsilon,$$

where, for any matrix $A = (a_{i,j})$, $\|A\|_\infty$ denotes $\max_{i,j} |a_{i,j}|$.

Theorem 2.4. *If the design matrix $\mathbf{X}$ is ε-incoherent, then for all $c > 0$ and all $J \subset \{1, \ldots, p\}$,*

$$\textsc{re}(c, J) \geq (1 - (2c+1)\varepsilon|J|)_+ ,$$

where $x_+ = \max\{x, 0\}$.

Proof of Theorem 2.4. Seen in class. □

3 Optimality

In this section, we assess the optimality of the perfomance bounds obtained so far through the so-called minimax point of view. The main result presented here builds upon three lemmas. Below, for any two probability measures P and Q, we denote

$$\mathbb{K}(P, Q) = \begin{cases} \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P & \text{if } P \ll Q \\ +\infty & \text{otherwise,} \end{cases}$$

the Kullback-Leibler divergence between P and Q.

Lemma 3.1 (Birgé's inequality). *Let $(S, \mathcal{S})$ be a measurable space and $A_1, \ldots, A_N$ be disjoints measurable sets in S. Then, for any collection $P_1, \ldots, P_N$ of probability measures on S, we have*

$$\min_{1 \le i \le N} P_i(A_i) \le \max \left\{ \frac{2e}{2e+1} \, ; \, \frac{\max_{i \ne j} \mathbb{K}(P_i, P_j)}{\log N} \right\}.$$

Proof of Lemma 3.1. Seen in class. □

Now, we present an application of Birgé's inequality. We fix $\sigma^2 > 0$ and an integer $n \geq 1$. For $f \in \mathbf{R}^n$, we denote P_f the gaussian distribution $\mathcal{N}(f, \sigma^2 \mathbf{I}_n)$. For any set $\mathcal{V} \subset \mathbf{R}^n$, we denote $\mathcal{R}_n(\mathcal{V})$ the minimax risk associated to $\mathcal{V}$ defined by

$$\mathcal{R}_n(\mathcal{V}) := \inf_t \max_{f \in \mathcal{V}} \frac{1}{n} \mathbf{E}_f \|t(\mathbf{Y}) - f\|_2^2,$$

where the infimum is taken over all (measurable) functions $t : \mathbf{R}^n \to \mathbf{R}^n$ and where, for any t and f, the notation

$$\mathbf{E}_f \|t(\mathbf{Y}) - f\|_2^2,$$

denotes the expectation of $\|t(\mathbf{Y}) - f\|^2$ when the random variable $\mathbf{Y}$ has distribution P_f. The next result provides a general lower bound for $\mathcal{R}_n(\mathcal{V})$ when $\mathcal{V}$ is finite.

Lemma 3.2. *For any finite set $\mathcal{V} \subset \mathbf{R}^n$, we have*

$$\mathcal{R}_n(\mathcal{V}) \ge \frac{1}{4n} \left(1 - \max \left\{ \frac{2e}{2e+1} \, ; \, \max_{f \ne f'} \frac{\|f - f'\|^2}{2\sigma^2 \log |\mathcal{V}|} \right\} \right) \min_{f \ne f'} \|f - f'\|^2,$$

where it is understood that the min *and the second* max *are taken over the set $\{(f, f') : f, f' \in \mathcal{V}, f \neq f'\}$.*

Proof of Lemma 3.2. Seen in class. □

The next and last lemma provides a version of the Varshamov-Gilbert in the sparse framework.

Lemma 3.3 (Sparse Varshamov-Gilbert). *For any integer $s \le p/5$, there exists a set $\mathcal{B} \subset \{\boldsymbol{\beta} \in \{0,1\}^p : \|\boldsymbol{\beta}\|_0 = s\}$, satisfying the following conditions:*

(1) $\forall \boldsymbol{\beta} \neq \boldsymbol{\beta}' \in \mathcal{B} : \ \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_0 > s$.

(2) *The cardinality $|\mathcal{B}|$ of $\mathcal{B}$ satisfies*

$$\log |\mathcal{B}| \ge \frac{s}{2} \log \left(\frac{p}{5s} \right).$$

Proof of Lemma 3.3. We refer the interested reader to Exercise 2.9.6 in Giraud (2015) or Lemma 5.14 in Rigollet (2015) for the proof. □

We are now in position to state the main result of this section. Denote

$$c_-(\mathbf{X}) := \inf_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0 \le 2s} \frac{\|\mathbf{X}\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2} \quad \text{and} \quad c_+(\mathbf{X}) := \sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0 \le 2s} \frac{\|\mathbf{X}\boldsymbol{\beta}\|_2}{\|\boldsymbol{\beta}\|_2}.$$

Then, letting

$$\mathcal{F}(s) = \{\mathbf{X}\beta : \|\boldsymbol{\beta}\|_0 \le s\},$$

we deduce from the two previous lemmas the following result.

Theorem 3.1. *For any integer $s \le p/5$ and any $n \ge 1$,*

$$\mathcal{R}_n(\mathcal{F}(s)) \ge \frac{e}{8(2e+1)^2} \left(\frac{c_-(\mathbf{X})}{c_+(\mathbf{X})}\right)^2 \frac{\sigma^2 s}{n} \log\left(\frac{p}{5s}\right).$$

Proof of Lemma 3.1. Seen in class □

References

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

C. Giraud. *Introduction to High-dimensional Statistics*. CRC Press, 2015.

P. Rigollet. High-dimensional statistics. MIT lecture notes, 2015.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4): 879–898, 2012.