ON PREDICTION WITH THE LASSO WHEN THE DESIGN IS NOT INCOHERENT

STÉPHANE CHRÉTIEN

Abstract. The LASSO estimator is an ℓ_1-norm penalized least-squares estimator, which was introduced for variable selection in the linear model. When the design matrix satisfies, e.g. the Restricted Isometry Property, or has a small coherence index, the LASSO estimator has been proved to recover, with high probability, the support and sign pattern of sufficiently sparse regression vectors. Under similar assumptions, the LASSO satisfies adaptive prediction bounds in various norms. The present note provides a prediction bound based on a new index for measuring how favorable is a design matrix for the LASSO estimator. We study the behavior of our new index for matrices with independent random columns uniformly drawn on the unit sphere. Using the simple trick of appending such a random matrix (with the right number of columns) to a given design matrix, we show that a prediction bound similar to [6, Theorem 2.1] holds without any constraint on the design matrix, other than restricted non-singularity.

Keywords: LASSO; Coherence; Restricted Isometry Property; ℓ_1-penalization; High dimensional linear model.

1. Introduction

Given a linear model

$$y = X\beta + \varepsilon \tag{1.1}$$

where $X \in \mathbb{R}^{n\times p}$ and ε is a random vector with gaussian distribution $\mathcal{N}(0, \sigma^2 I)$ the LASSO estimator is given by

$$\hat{\beta} \quad = \quad \operatorname{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{1.2}$$

This estimator was first proposed in the paper of Tibshirani [18]. The LASSO estimator $\hat{\beta}$ is often used in the high dimensional setting where p is much larger than n. As can be expected, when $p \gg n$, estimation of β is hopeless in general unless some additional property of β is assumed. In many practical situations, it is considered relevant to assume that β is sparse, i.e. has only a few nonzero components, or at least compressible, i.e. the magnitude of the non zero coefficients decays with high rate. It is now well recognized that the ℓ_1 penalization of the likelihood often promotes sparsity under certain assumptions on the matrix X. We refer the reader to the book [4] and the references therein for a state of the art presentation of the LASSO and the tools involved in the theoretical analysis of its properties. One of the main interesting properties of the LASSO estimator is that it is a solution of a convex optimization problem and it can be computed in polynomial time, i.e. very quickly in the sense of computational complexity theory. This makes a big difference with other approaches based on variable selection criteria like AIC [1], BIC [17], Foster and George's Risk Inflation Criterion [12], etc, which are based on enumeration of the possible models, or even with the recent proposals of Dalalyan, Rigollet and Tsybakov [15], [9], although enumeration is replaced with a practically more efficient Monte Carlo Markov Chain algorithm.

In the problem of estimating $X\beta$, i.e. the prediction problem, it is often believed that the price to pay for reducing the variable selection approach to a convex optimization problem is a certain set of assumptions on the design matrix X [*]. One of the main contributions of [15] is that no particular

Laboratoire de Mathématiques, UMR 6623, Université de Franche-Comté, 16 route de Gray, 25030 Besancon, France. Email: stephane.chretien@univ-fcomte.fr.

[*]Conditions for model selection consistency are given in e.g. [22], and for exact support and sign pattern recovery with finite samples and $p \gg n$, in [6], [21].

assumption on X is required for the prediction problem, as opposed to the known results concerning the LASSO such that [3], [2], [6] and [20], and the many references cited in these works.

An impressive amount of work has been done in the recents years in order to understand the properties of $\hat{\beta}$ unded various assumptions on X. See the recent book by P. Buhlmann and S. Van de Geer [4] for the state of the art. Two well known assumptions on the design matrix are

- small Coherence $\mu(X)$
- small Restricted Isometry Constant $\delta(X)$

where the Coherence $\mu(X)$ is defined as

$$\mu(X) \quad = \quad \max_{j,j'} |X_j^t X_{j'}|,$$

and the Restricted Isometry Constant $\delta(X)$ is the smallest δ such that

$$(1 - \delta)\|\beta_T\|_2 \leq \|X_T \beta_T\|_2 \leq (1 + \delta)\|\beta_T\|_2 \tag{1.3}$$

for **all** subset T with cardinal s and all $\beta \in \mathbb{R}^p$. Other conditions are listed in [5]; see Figure 1 in that paper for a diagram summarizing all relationships between them. The Restricted Isometry property is very stringent and implies almost other conditions. Moreover, the Restricted Isometry Constant is NP-hard to compute for general matrices. On the other hand, the Coherence only requires of the order $np(p-1)$ elementary operations. However, it was proved in [6] that a small coherence, say of the order of $1/\log(p)$, is sufficient to prove a property very close to the Restricted Isometry Property: (1.3) holds for a large proportion of subsets $T \subset \{1, \ldots, p\}$, $|T| = s$ (of the order $1 - 1/p^\alpha$, $\alpha > 0$). This result was later refined in [7] with better constants using the recently discovered Non-Commutative deviation inequalities [19]. Less stringent properties are the restricted eigenvalue, the irrepresentable and the compatibility properties.

The goal of this short note is to show that, using a very simple trick, one can prove prediction bounds similar to [6, Theorem 2.1] without any assumption on the design matrix X at the low expense of appending to X a random matrix with independent columns uniformly distributed on the sphere.

For this purpose, we introduce a new index for design matrices, denoted by $\gamma_{s,\rho_-}(X)$ that allows to obtain novel adaptive bounds on the prediction error. This index is defined for any $s \leq n$ and $\rho_- \in (0, 1)$ as

$$\gamma_{s,\rho_-}(X) \quad = \quad \sup_{v \in B(0,1)} \inf_{I \subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\|_\infty, \tag{1.4}$$

where $\mathcal{S}_{s,\rho_-}(X)$ is the family of all S of $\{1, \ldots, p\}$ with cardinal $|S| = s$, such that $\sigma_{\min}(X_S) \geq \rho_-$. The meaning of the index γ_{s,ρ_-} is the following: for any $v \in \mathbb{R}^n$, we look for the "almost orthogonal" family inside the set of columns of X with cardinal s, which is the most orthogonal to v .

One major advantage of this new parameter is that imposing the condition that γ_{s,ρ_-} is small is much less stringent than previous criteria required in the litterature. In particular, many submatrices of X may be very badly conditioned or even singular without altering the smallness of γ_{s,ρ_-}. Computing the new index $\gamma_{s,\rho_-}(X)$ for random matrices with independent columns uniformly distributed on the sphere †, shows that a prediction bound involving $\gamma_{s,\rho_-}(X)$ can be obtained which is of the same order as the bound of [6, Theorem 2.1].

One very nice property of the index γ_{s,ρ_-} is that it decreases after the operation appending any matrix to a given one. As a very nice consequence of this observation, the results obtained for random matrices can be extended to any matrix X to which a random matrix is appended. This trick can be used to prove new prediction bounds for a modified version of the LASSO obtained by appending a random matrix to any given design matrix. This simple modification of the LASSO retains the fundamental property of being polynomial time solvable unlike the recent approaches based on non-convex criteria for which no computational complexity analysis is available.

The plan of the paper is as follows. In Section 2 we present the index γ_{s,ρ_-} for X and provide an upper bound on this index for random matrices with independent columns uniformly distributed on the sphere, holding with high probability. Then, we present our prediction bound in Theorem 2.4: we give a bound on the prediction squared error $\|X(\beta - \hat{\beta}\|_2^2$ which depends linearly on s. This result is similar in spirit to [6, Theorem 1.2]. The proofs of the above results are given in Section 3. In Section

†or equivalently, post-normalized Gaussian i.i.d. matrices with components following $\mathcal{N}(0, 1/n)$.

4, we show how these results can be applied in practice to any problem with a matrix for which γ_{s,ρ_-} is unknown by appending to X an $n \times p_0$ random matrix with i.i.d. columns uniformly distributed on the unit sphere of $\mathbb{R}^n$ and with only $p_0 = O(n^{\frac{3}{2}} \rho_+)$ columns. An appendix contains the proof of some intermediate results.

1.1. **Notations and preliminary assumptions.** A vector β in $\mathbb{R}$ is said to be s-sparse if exactly s of its components are different from zero. Let ρ_- be a positive real number. In the sequel, we will denote by $\mathcal{S}_{s,\rho_-}(X)$ the family of all index subsets S of $\{1, \dots, p\}$ with cardinal $|S| = s$, such that for all $S \in \mathcal{S}_{s,\rho_-}$, $\sigma_{\min}(X_S) \geq \rho_-$.

2. Main results

2.1. A new index for design matrices.

Definition 2.1. *The index $\gamma_{s,\rho_-}(X)$ associated with the matrix X in $\mathbb{R}^{n\times p}$ is defined by*

$$\gamma_{s,\rho_-}(X) \quad = \quad \sup_{v \in B(0,1)} \inf_{I \subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\|_\infty. \tag{2.5}$$

An important remark is that the function $X \mapsto \gamma_{s,\rho_-}(X)$ is nonincreasing in the sense that if we set $X'' = [X, X']$, where X' is a matrix in $\mathbb{R}^{n\times p'}$, then $\gamma_{s,\rho_-}(X) \geq \gamma_{s,\rho_-}(X')$.

Unlike the coherence $\mu(X)$, for fixed n and s, the quantity $\gamma_{s,\rho_-}(X)$ is very small for p sufficiently large, at least for random matrices such as normalized standard Gaussian matrices as shown in the following proposition.

Proposition 2.2. *Assume that X is random matrix in $\mathbb{R}^{n\times p}$ with i.i.d. columns with uniform distribution on the unit sphere of $\mathbb{R}^n$. Let ρ_- and $\varepsilon \in (0, 1)$, $C_\kappa \in (0, +\infty)$ and $p_0 \in \{\lceil e^{\frac{6}{\sqrt{2\pi}}} \rceil, \dots, p\}$. Set*

$$K_\varepsilon \quad = \quad \frac{\sqrt{2\pi}}{6} \left((1 + C_\kappa) \log \left(1 + \frac{2}{\varepsilon} \right) + C_\kappa + \log \left(\frac{C_\kappa}{4} \right) \right).$$

Assume that n, κ and s satisfy

$$n \geq 6, \tag{2.6}$$

$$\kappa = \max \left\{ 4e^{-2(\ln(2)-1)}, \frac{4e^3}{(1-\rho_-)^2} \left(\frac{(1+K_\varepsilon)(1+C_\kappa)}{c(1-\varepsilon)^4} \right)^2 \log^2(p_0) \log(C_\kappa n) \right\}, \tag{2.7}$$

$$\frac{\max\{\kappa s, 2 \times 36 \times 3 \times 3, \exp((1-\rho_-)/2)\}}{C_\kappa} \leq n \leq \min \left\{ \left(\frac{p_0}{\log(p_0)} \right)^2, \frac{\exp\left(\frac{1-\rho_-}{\sqrt{2}} p_0 \right)}{C_\kappa} \right\}. \tag{2.8}$$

Then, we have

$$\gamma_{s,\rho_-}(X) \quad \leq \quad 80 \, \frac{\log(p_0)}{p_0} \tag{2.9}$$

with probability at least $1 - 5 \, \frac{n}{p_0 \, \log(p_0)^{n-1}} - 9 \, p_0^{-n}$.

Remark 2.3. *Notice that the constraints (2.7) and (2.8) together imply the following constraint on s:*

$$s \leq \quad C_{sparsity} \, \frac{n}{\log^2(p_0) \log(C_\kappa n)}$$

with

$$C_{sparsity} \quad = \quad \frac{c^2(1-\rho_-)^2(1-\varepsilon)^8}{4e^3} \, \frac{C_\kappa}{(1+K_\varepsilon)^2(1+C_\kappa)^2}.$$

2.2. **A bound of $\|X(\beta-\hat{\beta})\|_2^2$ based on $\gamma_{s,\rho_-}(X)$.** In the remainder of this paper, we will assume that the columns of X are ℓ_2-normalized. The main result of this paper is the following theorem.

Theorem 2.4. *Let $\rho_- \in (0,1)$. Let ν be a positive real such that*

$$\nu\ \gamma_{\nu n,\rho_-}(X) \le \frac{\rho_-\ \sigma_{\min}(X_S)}{n\ \max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\le n}} \sigma_{\max}(X_T)}. \tag{2.10}$$

Assume that $s \le \nu n$. Assume that β has support S with cardinal s and that

$$\lambda \ge \tag{2.11}$$

$$\sigma\left(B_{X,\nu,\rho_-}\ \max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\le n}} \sigma_{\max}(X_T)\sqrt{2\alpha\ \log(p)+\log(2\nu n)} + \sqrt{(2\alpha+1)\ \log(p)+\log(2)}\right)$$

with

$$B_{X,\nu,\rho_-} \quad = \quad \frac{\nu n\ \gamma_{\nu n,\rho_-}(X)}{\rho_-\ \sigma_{\min}(X_S) - \nu n\ \gamma_{\nu n,\rho_-}(X)\ \max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\le n}} \sigma_{\max}(X_T)}. \tag{2.12}$$

Then, with probability greater than $1-p^{-\alpha}$, we have

$$\frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \quad \le \quad s\ C_{n,p,\rho_-,\alpha,\nu,\lambda} \tag{2.13}$$

with

$$C_{n,p,\rho_-,\alpha,\nu,\lambda} \quad = \quad \frac{\lambda+\sigma\sqrt{(2\alpha+1)\ \log(p)+\log(2)}}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2\nu n)}+\lambda\right) \tag{2.14}$$

2.3. **Comments.** Equation (2.10) in Theorem 2.4 requires that

$$\gamma_{\nu n,\rho_-}(X) \quad < \quad \rho_- \frac{\sigma_{\min}(X_S)}{\nu n\ \max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\le n}} \sigma_{\max}(X_T)}. \tag{2.15}$$

Proposition 2.2 proves that for random matrices with independent columns uniformly drawn on the unit sphere of $\mathbb{R}^n$ (i.e. normalized i.i.d. gaussian matrices),

$$\gamma_{s,\rho_-}(X) \quad \le \quad 80\ \frac{\log(p_0)}{p_0} \tag{2.16}$$

with high probability. The case of general design matrices can be treated using a simple trick. It will be studied in Section 4.

The main advantage of using the parameter $\gamma_{\nu n,\rho_-}(X)$ is that it allows X to contain extremely badly conditioned submatrices, a situation that may often occur in practice when certain covariates are very correlated. This is in contrast with the Restricted Isometry Property or the Incoherence condition, or other conditions often required in the litterature. On the other hand, the parameter $\gamma_{\nu n,\rho_-}(X)$ is not easily computable. We will see however in Section 4 how to circumvent this problem in practice by the simple trick consisting of appending a random matrix with p_0 columns to the matrix X in order to ensure that X satisfies (2.15) with high probability.

Finally, notice that unlike in [6, Theorem 2.1], we make no assumption on the sign pattern of β. In particular, we do not require the sign pattern of the nonzero components to be random. Moreover, the extreme singular values of X_S are not required to be independent of n nor p and the condition (2.15) is satisfied for a wide range of configurations of the various parameters involved in the problem.

3. Proofs

3.1. **Proof of Proposition 2.2.**

3.1.1. *Constructing an outer approximation for I in the definition of γ_{s,ρ_-}.* Take $v \in \mathbb{R}^n$. We construct an outer approximation $\tilde{I}$ of I into which we be able to extract the set I. We procede recursively as follows: until $|\tilde{I}| = \min\{\kappa s, p_0/2\}$, for some positive real number κ to be specified later, do

- Choose $j_1 = \operatorname{argmin}_{j=1,\ldots,p_0} |\langle X_j, v\rangle|$ and set $\tilde{I} = \{j_1\}$
- Choose $j_2 = \operatorname{argmin}_{j=1,\ldots,p_0,\ j\notin \tilde{I}} |\langle X_j, v\rangle|$ and set $\tilde{I} = \tilde{I} \cup \{j_2\}$
- $\cdots$
- Choose $j_k = \operatorname{argmin}_{j=1,\ldots,p_0,\ j\notin \tilde{I}} |\langle X_j, v\rangle|$ and set $\tilde{I} = \tilde{I} \cup \{j_k\}$.

3.1.2. *An upper bound on $\|X_{\tilde{I}}^t v\|_\infty$.* If we denote by Z_j the quantity $|\langle X_j, v\rangle|$ and by $Z_{(r)}$ the r^{th} order statistic, we get that

$$\|X_{\tilde{I}}^t v\|_\infty = Z_{(\kappa s)}.$$

Since the X_j's are assumed to be i.i.d. with uniform distribution on the unit sphere of $\mathbb{R}^n$, we obtain that the distribution of $Z_{(r)}$ is the distribution of the r^{th} order statistics of the sequence $|X_j^t v|$, $j = 1, \ldots, p_0$. By (5) p.147 [13], $|X_j^t v|$ has density g and CDF G given by

$$g(z) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \left(1 - z^2\right)^{\frac{n-3}{2}} \text{ and } G(z) = 2 \int_0^z g(\zeta)\, d\zeta.$$

Thus,

$$F_{Z_{(r)}}(z) = \mathbb{P}\left(B \geq r\right)$$

where B is a binomial variable $\mathcal{B}\left(p_0, G(z)\right)$. Our next goal is to find the smallest value z_0 of z which satisfies

$$F_{Z_{(\kappa s)}}(z_0) \geq 1 - p_0^{-n}. \tag{3.17}$$

We have the following standard concentration bound for B (e.g. [11]):

$$\mathbb{P}\left(B \leq (1-\varepsilon)\mathbb{E}[B]\right) \leq \exp\left(-\frac{1}{2}\, \varepsilon^2 \mathbb{E}[B]\right)$$

which gives

$$\mathbb{P}\left(B \geq (1-\varepsilon) p_0 G(z)\right) \geq 1 - \exp\left(-\frac{1}{2}\, \varepsilon^2 p_0 G(z)\right)$$

We thus have to look for a root (or at least an upper bound to a root) of the equation

$$G(z) = \frac{1}{\frac{1}{2}\,\varepsilon^2} \frac{n}{p_0} \log(p_0).$$

Notice that

$$\begin{aligned} G(z) &= 2\, \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \int_0^z \left(1 - \zeta^2\right)^{\frac{n-3}{2}} d\zeta, \\ &\geq \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\, z \end{aligned}$$

for $z \leq 1/\sqrt{2}$. By a straightforward application of Stirling's formula (see e.g. (1.4) in [14]), we obtain

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \geq \frac{e^{2\ln(2)}}{2} \frac{(n-3)^{3/2}}{(n-2)^{1/2}}.$$

Thus, any choice of z_0 satisfying

$$z_0 \geq \frac{2\sqrt{\pi}}{e^{2\ln(2)}} \frac{(n-2)^{1/2}}{(n-3)^{3/2}} \frac{1}{\frac{1}{2}\,\varepsilon^2} \frac{n}{p_0} \log(p_0) \tag{3.18}$$

is an upper bound to the quantile for $(1-\varepsilon)p_0 G(z_0)$-order statistics at level p_0^{-n}. We now want to enforce the constraint that

$$(1-\varepsilon) p_0 G(z_0) \leq \kappa s.$$

By again a straightforward application of Stirling's formula, we obtain

$$G(z) \leq \frac{1}{\sqrt{\pi}} \frac{e^2}{2} \frac{(n-3)^{3/2}}{(n-2)^{1/2}} z$$

for $n \geq 4$. Thus, we need to impose that

$$z_0 \leq \frac{2\sqrt{\pi}}{e^2} \frac{(n-2)^{1/2}}{(n-3)^{3/2}} \frac{\kappa s}{(1-\varepsilon)p_0}. \tag{3.19}$$

Notice that the constraints (3.18) and (3.19) are compatible if

$$\kappa \geq \frac{4}{e^{2(\ln(2)-1)}} \frac{1-\varepsilon}{\varepsilon^2} \frac{n}{s} \log(p_0).$$

Take $\varepsilon = 1 - \frac{1}{n/s\log(p_0)}$ and obtain

$$\mathbb{P}\left(\|X_{\tilde{I}}^t v\|_\infty \geq \frac{8\sqrt{\pi}}{e^{2\ln(2)}} \frac{(n-2)^{1/2}}{(n-3)^{3/2}} \frac{n}{p_0} \log(p_0)\right) \leq p_0^{-n}$$

for

$$\kappa = \frac{4}{e^{2(\ln(2)-1)}}$$

for any p_0 such that $n/s\log(p_0) \geq \sqrt{2}$, which is clearly the case as soon as $p_0 \geq e^{\frac{6}{\sqrt{2\pi}}}$ for $s \leq n$ as assumed in the proposition.

If $n \geq 6$, we can simplify (3.20) with

$$\mathbb{P}\left(\|X_{\tilde{I}}^t v\|_\infty \geq 80 \frac{\log(p_0)}{p_0}\right) \leq p_0^{-n} \tag{3.20}$$

3.1.3. *Extracting a well conditionned submatrix of* $X_{\tilde{I}}$. The method for extracting X_I from $X_{\tilde{I}}$ uses random column selection. For this purpose, we will need to control the coherence and the norm of $X_{\tilde{I}}$.

Step 1: The coherence of $X_{\tilde{I}}$. Let us define the spherical cap

$$\mathcal{C}(v,h) = \{w \in \mathbb{R}^n \mid \langle v, w\rangle \geq h\}.$$

The area of $\mathcal{C}(v,h)$ is given by

$$Area\left(\mathcal{C}(v,h)\right) = Area(\mathcal{S}(0,1)) \int_0^{2h-h^2} t^{\frac{n-1}{2}} (1-t)^{\frac{1}{2}} dt.$$

Thus, the probability that a random vector w with Haar measure on the unit sphere $\mathcal{S}(0,1)$ falls into the spherical cap $\mathcal{C}(v,h)$ is given by

$$\mathbb{P}\left(w \in \mathcal{C}(v,h)\right) = \frac{\mathcal{C}(v,h)}{\mathcal{S}(0,1)} = \frac{\int_0^{2h-h^2} t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt}{\int_0^1 t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt}.$$

The last term is the CDF of the Beta distribution. Using the fact that

$$\mathbb{P}\left(X_j \in \mathcal{C}(X_{j'},h)\right) = \mathbb{P}\left(X_{j'} \in \mathcal{C}(X_j,h)\right)$$

the union bound, and the independence of the X_j's, the probability that $X_j \in \mathcal{C}(X_{j'}, h)$ for some (j, j') in $\{1, \ldots, p_0\}^2$ can be bounded as follows

$$\begin{aligned}
\mathbb{P}\left(\cup_{j \neq j'=1}^{p_0} \left\{X_j \in \mathcal{C}(X_{j'}, h)\right\}\right) &= \mathbb{P}\left(\cup_{j < j'=1}^{p_0} \left\{X_j \in \mathcal{C}(X_{j'}, h)\right\}\right) \\
&\leq \sum_{j<j'=1}^{p_0} \mathbb{P}\left(\left\{X_j \in \mathcal{C}(X_{j'}, h)\right\}\right) \\
&= \sum_{j<j'=1}^{p_0} \mathbf{E}\ \left[\mathbb{P}\left(\left\{X_j \in \mathcal{C}(X_{j'}, h)\right\} \mid X_{j'}\right)\right] \\
&= \frac{p_0(p_0-1)}{2} \int_0^{2h-h^2} t^{\frac{n-1}{2}} (1-t)^{\frac{1}{2}} dt.
\end{aligned}$$

Our next task is to choose h so that

$$\frac{p_0(p_0-1)}{2} \int_0^{2h-h^2} t^{\frac{n-1}{2}} (1-t)^{\frac{1}{2}} dt \quad \leq \quad p_0^{-n}.$$

Let us make the following crude approximation

$$\frac{p_0(p_0-1)}{2} \int_0^{2h-h^2} t^{\frac{n-1}{2}} (1-t)^{\frac{1}{2}} dt \quad \leq \quad \frac{p_0^2}{2} (2h)^{\frac{n-1}{2}} (2h-0).$$

Thus, taking

$$h \quad \geq \quad \frac{1}{2} \exp\left(-2\left(\log(p_0) + \frac{\log(p_0) - \log(2))}{n+1}\right)\right)$$

will work. Moreover, since $p_0 \geq 2$, we deduce that

$$\mu(X_{\tilde{I}}) \quad \leq \quad \frac{1}{2}\, p_0^{-2} \tag{3.21}$$

with probability at least $1 - p_0^{-n}$.

Step 2: The norm of $X_{\tilde{I}}$. The norm of any submatrix X_S with n rows and κs columns of X has the following variational representation

$$\|X_S\| \quad = \quad \max_{\substack{v \in \mathbb{R}^n,\ \|v\|=1 \\ w \in \mathbb{R}^{\kappa s},\ \|w\|=1}} v^t X_S w.$$

We will use an easy ε-net argument to control this norm. For any $v \in \mathbb{R}^n$, $v^t X_j$, $j \in S$ is a sub-Gaussian random variable satisfying

$$\mathbb{P}\left(|v^t X_j| \geq u\right) \quad \leq \quad 2\exp\left(-cn\ u^2\right),$$

for some constant c. Therefore, using the fact that $\|w\| = 1$, we have that

$$\mathbb{P}\left(\left|\sum_{j \in S} v^t X_S w\right| \geq u\right) \quad \leq \quad 2\exp\left(-cn\ u^2\right).$$

Let us recall two useful results of Rudelson and Vershynin. The first one gives a bound on the covering number of spheres.

Proposition 3.1. ([16, Proposition 2.1]). *For any positive integer d, there exists an ε-net of the unit sphere of $\mathbb{R}^d$ of cardinality*

$$2d\left(1 + \frac{2}{\varepsilon}\right)^{d-1}.$$

The second controls the approximation of the norm based on an ε-net.

Proposition 3.2. ([16, Proposition 2.2]). *Let $\mathcal{N}$ be an ε-net of the unit sphere of $\mathbb{R}^d$ and let $\mathcal{N}'$ be an ε'-net of the unit sphere of $\mathbb{R}^{d'}$. Then for any linear operator $A : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$, we have*

$$\|A\| \leq \frac{1}{(1-\varepsilon)(1-\varepsilon')} \sup_{\substack{v \in \mathcal{N} \\ w \in \mathcal{N}'}} |v^t A w|.$$

Let $\mathcal{N}$ (resp. $\mathcal{N}'$) be an ε-net of the unit sphere of $\mathbb{R}^{\kappa s}$ (resp. of $\mathbb{R}^n$). On the other hand, we have that

$$\begin{aligned} \mathbb{P}\left(\sup_{\substack{v \in \mathcal{N} \\ w \in \mathcal{N}'}} |v^t A w| \geq u \right) &\leq 2|\mathcal{N}||\mathcal{N}'| \exp\left(-cn\ u^2\right), \\ &\leq 8\ n\kappa s \left(1 + \frac{2}{\varepsilon}\right)^{n+\kappa s - 2} \exp\left(-cn\ u^2\right), \end{aligned}$$

which gives

$$\mathbb{P}\left(\sup_{\substack{v \in \mathcal{N} \\ w \in \mathcal{N}'}} |v^t A w| \geq u \right) \leq 8\ \frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2} \exp\left(-\left(cn\ u^2 - (n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)\right)\right).$$

Using Proposition (3.2), we obtain that

$$\mathbb{P}\left(\|X_S\| \geq u\right) \leq \mathbb{P}\left(\frac{1}{(1-\varepsilon)^2} \sup_{\substack{v \in \mathcal{N} \\ w \in \mathcal{N}'}} |v^t A w| \geq u \right).$$

Thus, we obtain

$$\mathbb{P}\left(\|X_S\| \geq u\right) \leq 8\ \frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2} \exp\left(-\left(cn\ (1-\varepsilon)^4\ u^2 - (n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)\right)\right).$$

To conclude, let us note that

$$\begin{aligned} \mathbb{P}\left(\|X_{\tilde{I}}\| \geq u\right) &\leq \mathbb{P}\left(\max_{\substack{S \subset \{1,\ldots,p_0\} \\ |S| = \kappa s}} \|X_S\| \geq u \right) \\ &\leq \binom{p_0}{\kappa s}\ 8\ \frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2} \exp\left(-\left(cn\ (1-\varepsilon)^4\ u^2 - (n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)\right)\right). \end{aligned}$$

and using the fact that

$$\binom{p_0}{\kappa s} \leq \left(\frac{e\ p_0}{\kappa s}\right)^{\kappa s},$$

one finally obtains

$$\mathbb{P}\left(\|X_{\tilde{I}}\| \geq u\right) \leq 8\ \exp\left(-\left(cn\ (1-\varepsilon)^4\ u^2 - (n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right) - \kappa s \log\left(\frac{e\ p_0}{\kappa s}\right) - \log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right)\right)\right).$$

The right hand side term will be less than $8p_0^{-n}$ when

$$n\log(p_0) \leq cn\ (1-\varepsilon)^4\ u^2 - (n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right) - \kappa s \log\left(\frac{e\ p_0}{\kappa s}\right) - \log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right).$$

This happens if

$$u^2 \geq \frac{1}{c(1-\varepsilon)^4}\left(n\frac{\log(p_0)}{n} + \left(1+\frac{\kappa s}{n}\right)\log\left(1+\frac{2}{\varepsilon}\right) + \frac{\kappa s}{n}\log\left(\frac{e\ p_0}{\kappa s}\right) + \frac{1}{n}\log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right)\right).$$

Notice that

$$\left(1+\frac{\kappa s}{n}\right)\log\left(1+\frac{2}{\varepsilon}\right)+\frac{\kappa s}{n}\log\left(\frac{e}{\kappa s}\right)+\frac{1}{n}\log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right) \tag{3.22}$$
$$\leq (1+C_\kappa)\log\left(1+\frac{2}{\varepsilon}\right)+C_\kappa+\frac{1}{n}\log\left(\frac{C_\kappa n^2}{4}\right),$$
$$\leq K_\varepsilon\ \frac{6}{\sqrt{2\pi}},$$

since $n \geq 1$. Now, since

$$\frac{6}{\sqrt{2\pi}} \leq \log(p_0) \leq \frac{n+\kappa s}{n}\log(p_0),$$

we finally obtain

$$\mathbb{P}\left(\|X_{\tilde{I}}\| \geq \frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right) \leq \frac{8}{p_0^n}. \tag{3.23}$$

Step 3. We will use the following lemma on the distance to identity of randomly selected submatrices.

Lemma 3.3. *Let $r \in (0,1)$. Let n, κ and s satisfy conditions (2.8) and (2.7) assumed in Proposition 2.2. Let $\Sigma \subset \{1,\dots,\kappa s\}$ be a random support with uniform distribution on index sets with cardinal s. Then, with probability greater than or equal to $1-9\ p_0^{-n}$ on X, the following bound holds:*

$$\mathbb{P}\left(\|X_\Sigma^t X_\Sigma - \mathrm{Id}_s\| \geq r \mid X\right) < 1. \tag{3.24}$$

Proof. See Appendix. □

Taking $r = 1-\rho_-$, we conclude from Lemma 3.3 that, for any s satisfying (2.10), there exists a subset $\tilde{\tilde{I}}$ of $\tilde{I}$ with cardinal s such that

$$\sigma_{\min}\left(X_{\tilde{\tilde{I}}}\right) \geq \rho_-.$$

3.1.4. *The supremum over an ε-net.* Recalling Proposition 3.1, there exists an ε-net $\mathcal{N}$ covering the unit sphere in $\mathbb{R}^n$ with cardinal

$$|\mathcal{N}| \leq 2n\left(1+\frac{2}{\varepsilon}\right)^{n-1}.$$

Combining this with (3.20), we have that

$$\mathbb{P}\left(\sup_{v\in\mathcal{N}}\inf_{I\subset\mathcal{S}_{s,\rho_-}}\left\|X_I^t v\right\| \geq \tfrac{8\ \sqrt{\pi}}{e^{2\ln(2)}}\ \tfrac{n\ (n-2)^{1/2}}{(n-3)^{3/2}}\ \tfrac{\log(p_0)}{p_0}\right)$$
$$\leq 2n\left(1+\tfrac{2}{\varepsilon}\right)^{n-1}\ p_0^{-n}+9\ p^{-n}. \tag{3.25}$$

3.1.5. *From the ε-net to the whole sphere.* For any v', one can find $v \in \mathcal{N}$ with $\|v'-v\|_2 \leq \varepsilon$. Thus, we have

$$\begin{aligned}\|X_I^t v'\|_\infty &\leq \|X_I^t v\|_\infty + \|X_I^t(v'-v)\|_\infty\\ &\leq \|X_I^t v\|_\infty + \max_{j\in I}|\langle X_j,(v'-v)\rangle|\\ &\leq \|X_I^t v\|_\infty + \max_{j\in I}\|X_j\|_2\|v'-v\|_2\\ &\leq \|X_I^t v\|_\infty + \varepsilon.\end{aligned} \tag{3.26}$$

Taking

$$\varepsilon = 80\ \frac{\log(p_0)}{p_0},$$

we obtain from (3.26) and (3.25) that

$$\mathbb{P}\left(\sup_{\|v\|_2=1}\inf_{I\subset\mathcal{S}_{s,\rho_-}} \left\|X_I^t v\right\| \geq 80\ \tfrac{\log(p_0)}{p_0}\right)$$
$$\leq 20\ n\left(1+\tfrac{p_0}{80\log(p_0)}\right)^{n-1}\ p_0^{-n}+9\ p_0^{-n}$$

and thus,

$$\mathbb{P}\left(\sup_{\|v\|_2=1}\inf_{I\subset\mathcal{S}_{s,\rho_-}} \left\|X_I^t v\right\| \geq 80\ \tfrac{\log(p_0)}{p_0}\right)$$
$$\leq 5\ \tfrac{n}{p_0\ \log(p_0)^{n-1}}+9\ p_0^{-n},$$

for $p_0 \geq \exp(6/\sqrt{2\pi})$.

3.2. **Proof of Theorem 2.4.**

3.2.1. *Optimality conditions.* The optimality conditions for the LASSO are given by

$$-X^t(y-X\hat{\beta})+\lambda g=0 \tag{3.27}$$

for some $g\in\partial\left(\|\cdot\|_1\right)_{\hat{\beta}}$. Thus, we have

$$X^tX(\hat{\beta}-\beta)=X^t\varepsilon-\lambda g. \tag{3.28}$$

from which one obtains that, for any index set $\mathcal{I}\subset\{1,\ldots,p\}$ with cardinal s,

$$\left\|X_{\mathcal{I}}^t X(\beta-\hat{\beta})\right\|_\infty \ \leq\ \lambda+\left\|X_{\mathcal{I}}^t\varepsilon\right\|_\infty, \tag{3.29}$$

3.2.2. *The support of $\hat{\beta}$.* As is well known, even when the solution of the LASSO optimization problem is not unique, there always exists a vector $\hat{\beta}$ whose support has cardinal n.

3.2.3. *A bound on $\|X_{\mathcal{I}}^t X_S(\beta_S-\hat{\beta}_S)\|_\infty$.* The argument is divided into three steps.

First step. Equation (3.29) implies that

$$\left\|X_{\mathcal{I}}^t X_S(\beta_S-\hat{\beta}_S)\right\|_\infty \ \leq\ \lambda+\left\|X_{\mathcal{I}}^t\varepsilon\right\|_\infty+\left\|X_{\mathcal{I}}^t X_{\hat{S}\cap S^c}(\beta_{\hat{S}\cap S^c}-\hat{\beta}_{\hat{S}\cap S^c})\right\|_\infty. \tag{3.30}$$

Second step. We now choose $\mathcal{I}$ as a solution of the following problem

$$\vartheta\ =\ \min_{\substack{I\subset\{1,\ldots,p\}\\ |I|=s}}\ \max_{j\in I}\left|\langle X_j, X_{\hat{S}\cap S^c}(\beta_{\hat{S}\cap S^c}-\hat{\beta}_{\hat{S}\cap S^c})\rangle\right|$$

subject to

$$\sigma_{\min}(X_I)\ \geq\ \rho_-.$$

By Definition 2.1,

$$\vartheta\ \leq\ \gamma_{s,\rho_-}(X)\ \|X_{\hat{S}\cap S^c}(\beta_{\hat{S}\cap S^c}-\hat{\beta}_{\hat{S}\cap S^c})\|_2$$

and thus,

$$\begin{aligned}\vartheta\ &\leq\ \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{\hat{S}\cap S^c})\|\beta_{\hat{S}\cap S^c}-\hat{\beta}_{\hat{S}\cap S^c}\|_2\\ &\leq\ \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{\hat{S}\cap S^c})\|\beta_{\hat{S}\cap S^c}-\hat{\beta}_{\hat{S}\cap S^c}\|_1\end{aligned}$$

which gives

$$\vartheta\ \leq\ \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{\hat{S}\cap S^c})\left(\|\beta_{\hat{S}\cap S^c}\|_1+\|\hat{\beta}_{\hat{S}\cap S^c}\|_1\right). \tag{3.31}$$

Third step. Combining (3.30) and (3.31), we obtain

$$\left\|X_{\mathcal{I}}^t X_S(\beta_S-\hat{\beta}_S)\right\|_\infty\ \leq\ \lambda+\left\|X_{\mathcal{I}}^t\varepsilon\right\|_\infty+\gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{\hat{S}\cap S^c})\left(\|\beta_{\hat{S}\cap S^c}\|_1+\|\hat{\beta}_{\hat{S}\cap S^c}\|_1\right).$$

Using the fact that

$$\left\|X_{\mathcal{I}}^t\varepsilon\right\|_\infty\ \leq\ \left\|X^t\varepsilon\right\|_\infty \tag{3.32}$$

and since

$$\mathbb{P}\left(\left\|X^t\varepsilon\right\|_\infty\geq\sigma\sqrt{2\alpha\ \log(p)+\log(2p)}\right)\ \leq\ p^{-\alpha},$$

we obtain that

$$\left\| X_{\mathcal{I}}^t X_S(\beta_S - \hat{\beta}_S) \right\|_\infty \leq \lambda + \sigma \sqrt{(2\alpha+1)\ \log(p) + \log(2)}$$

$$+\gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{\hat{S}\cap S^c}) \left(\|\beta_{\hat{S}\cap S^c}\|_1 + \|\hat{\beta}_{\hat{S}\cap S^c}\|_1 \right) \tag{3.33}$$

with probability greater than $1 - p^\alpha$.

3.2.4. *A basic inequality.* The definition of $\hat{\beta}$ gives

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Therefore, we have that

$$\frac{1}{2}\|\varepsilon - X(\hat{\beta} - \beta)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\|\beta\|_1$$

which implies that

$$\frac{1}{2}\|X(\hat{\beta} - \beta)\|_2^2 \leq \langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle + \langle \varepsilon, X_{\hat{S}\cap S^c}(\hat{\beta}_{\hat{S}\cap S^c} - \beta_{\hat{S}\cap S^c})\rangle$$

$$+\lambda\left(\|\beta_S\|_1 - \|\hat{\beta}_S\|_1\right) - \lambda\|\hat{\beta}_{\hat{S}\cap S^c}\|_1 + \lambda\|\beta_{\hat{S}\cap S^c}\|_1.$$

This can be further written as

$$\frac{1}{2}\|X(\hat{\beta} - \beta)\|_2^2 \leq \langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle + \langle X_{\hat{S}\cap S^c}^t \varepsilon, \hat{\beta}_{\hat{S}\cap S^c} - \beta_{\hat{S}\cap S^c}\rangle \tag{3.34}$$

$$+\lambda\left(\|\beta_S\|_1 - \|\hat{\beta}_S\|_1\right) - \lambda\|\hat{\beta}_{\hat{S}\cap S^c}\|_1 + \lambda\|\beta_{\hat{S}\cap S^c}\|_1. \tag{3.35}$$

3.2.5. *Control of $\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle$.* The argument is divided into two steps.

First step. We have

$$\begin{aligned}\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle &= \langle X_S^t\varepsilon, \hat{\beta}_S - \beta_S\rangle \\ &\leq \left\|X_S^t\varepsilon\right\|_\infty \|\hat{\beta}_S - \beta_S\|_1 \\ &\leq \sqrt{s}\ \left\|X_S^t\varepsilon\right\|_\infty \|\hat{\beta}_S - \beta_S\|_2\end{aligned}$$

and, using the fact that $\sigma_{\min}(X_{\mathcal{I}}) \geq \rho_-$,

$$\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle \leq \frac{s}{\rho_-\ \sigma_{\min}(X_S)}\ \left\|X_S^t\varepsilon\right\|_\infty \|X_{\mathcal{I}}^t X_S(\hat{\beta}_S - \beta_S)\|_\infty.$$

Second step. Since the columns of X have unit ℓ_2-norm, we have

$$\mathbb{P}\left(\left\|X_S^t\varepsilon\right\|_\infty \geq \sigma\sqrt{2\alpha\ \log(p) + \log(2s)}\right) \leq p^{-\alpha},$$

which implies that

$$\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle \leq \frac{s\ \sigma\sqrt{2\alpha\ \log(p) + \log(2s)}}{\rho_-\ \sigma_{\min}(X_S)}\ \|X_{\mathcal{I}}^t X_S\left(\hat{\beta}_S - \beta_S\right)\|_\infty \tag{3.36}$$

with probability at least $1 - p^{-\alpha}$.

3.2.6. *Control of $\langle X_{\hat{S}\cap S^c}^t\varepsilon, \hat{\beta}_{\hat{S}\cap S^c} - \beta_{\hat{S}\cap S^c}\rangle$.* We have

$$\langle X_{\hat{S}\cap S^c}^t\varepsilon, \hat{\beta}_{\hat{S}\cap S^c} - \beta_{\hat{S}\cap S^c}\rangle \leq \left\|X_{\hat{S}\cap S^c}^t\varepsilon\right\|_\infty \|\hat{\beta}_{\hat{S}\cap S^c} - \beta_{\hat{S}\cap S^c}\|_1. \tag{3.37}$$

On the other hand, we have

$$\mathbb{P}\left(\left\|X_{\hat{S}\cap S^c}^t\varepsilon\right\|_\infty \geq \sigma\sqrt{2\alpha\ \log(p) + \log(2(p-s))}\right) \leq p^{-\alpha},$$

which, combined with (3.37), implies that

$$\langle X_{\hat{S}\cap S^c}^t\varepsilon, \hat{\beta}_{\hat{S}\cap S^c} - \beta_{\hat{S}\cap S^c}\rangle \leq \sigma\sqrt{2\alpha\ \log(p) + \log(2(p-s))}\left(\|\hat{\beta}_{\hat{S}\cap S^c}\|_1 + \|\beta_{\hat{S}\cap S^c}\|_1\right)$$

with probability at least $1 - p^{-\alpha}$.

3.2.7. *Control of* $\|\beta_S\|_1 - \|\hat{\beta}_S\|_1$. The subgradient inequality gives

$$\|\hat{\beta}_S\|_1 - \|\beta_S\|_1 \quad \geq \quad \langle \mathrm{sign}(\beta_S), \hat{\beta}_S - \beta_S \rangle.$$

We deduce that

$$\begin{aligned} \|\beta_S\|_1 - \|\hat{\beta}_S\|_1 \quad &\leq \quad \|-\mathrm{sign}(\beta_S)\|_\infty \, \|\hat{\beta}_S - \beta_S\|_1 \\ &\leq \quad \frac{\sqrt{s}}{\rho_- \ \sigma_{\min}(X_S)} \|X_{\mathcal{I}}^t X_S(\hat{\beta}_S - \beta_S)\|_2 \end{aligned}$$

which implies

$$\|\beta_S\|_1 - \|\hat{\beta}_S\|_1 \quad \leq \quad \frac{s}{\rho_- \ \sigma_{\min}(X_S)} \ \|X_{\mathcal{I}}^t X_S(\hat{\beta}_S - \beta_S)\|_\infty. \tag{3.38}$$

3.2.8. *Summing up.* Combining (3.34) with (3.36), (3.38) and (3.33), the union bound gives that, with probability $1 - 3p^{-\alpha}$,

$$\begin{aligned} \frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \quad \leq \quad & \frac{s}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2s)} + \lambda\right) \left(\lambda + \sigma\sqrt{(2\alpha+1) \ \log(p) + \log(2)}\right. \\ & \left. + \gamma_{s,\rho_-}(X) \ \sigma_{\max}(X_{\hat{S}\cap S^c}) \left(\|\beta_{\hat{S}\cap S^c}\|_1 + \|\hat{\beta}_{\hat{S}\cap S^c}\|_1\right)\right) \\ & + \sigma\sqrt{2\alpha \ \log(p) + \log(2(p-s))} \left(\|\beta_{\hat{S}\cap S^c}\|_1 + \|\hat{\beta}_{\hat{S}\cap S^c}\|_1\right) \\ & + \lambda\left(\|\beta_{\hat{S}\cap S^c}\|_1 - \|\hat{\beta}_{\hat{S}\cap S^c}\|_1\right) \end{aligned}$$

which gives,

$$\begin{aligned} \frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \quad \leq \quad & s \ \frac{\lambda + \sigma\sqrt{(2\alpha+1) \ \log(p) + \log(2)}}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2s)} + \lambda\right) \\ & + \left(\frac{s}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2s)} + \lambda\right) \gamma_{s,\rho_-}(X) \ \sigma_{\max}(X_{\hat{S}\cap S^c})\right. \\ & \left. + \sigma\sqrt{2\alpha \ \log(p) + \log(2(p-s))} - \lambda\right) \|\hat{\beta}_{\hat{S}\cap S^c}\|_1 \\ & + \left(\frac{s}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2s)} + \lambda\right) \gamma_{s,\rho_-}(X) \ \sigma_{\max}(X_{\hat{S}\cap S^c})\right. \\ & \left. + \sigma\sqrt{2\alpha \ \log(p) + \log(2(p-s))} + \lambda\right) \|\beta_{\hat{S}\cap S^c}\|_1. \end{aligned}$$

Using the assumption that $s \leq \nu n$, we obtain

$$\begin{aligned} \frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \quad \leq \quad & s \ \frac{\lambda + \sigma\sqrt{(2\alpha+1) \ \log(p) + \log(2)}}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2\nu n)} + \lambda\right) \\ & + \left(\frac{\nu n}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2\nu n)} + \lambda\right) \gamma_{s,\rho_-}(X) \ \sigma_{\max}(X_{\hat{S}\cap S^c})\right. \\ & \left. + \sigma\sqrt{(2\alpha+1) \ \log(p) + \log(2)} - \lambda\right) \|\hat{\beta}_{\hat{S}\cap S^c}\|_1 \\ & + \left(\frac{\nu n}{\rho_- \ \sigma_{\min}(X_S)} \left(\sigma\sqrt{2\alpha \ \log(p) + \log(2\nu n)} + \lambda\right) \gamma_{s,\rho_-}(X) \ \sigma_{\max}(X_{\hat{S}\cap S^c})\right. \\ & \left. + \sigma\sqrt{2\alpha \ \log(p) + \log(2(p))} + \lambda\right) \|\beta_{\hat{S}\cap S^c}\|_1. \end{aligned}$$

Since, as recalled in Section 3.2.2, the support of $\hat{\beta}$ has cardinal less than or equal to n, we have

$$\sigma_{\max}(X_{\hat{S}\cap S^c}) \quad \leq \quad \max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T),$$

and the proof is completed.

4. A simple trick when γ_{s,ρ_-} is unknown: appending a random matrix

We have computed the index γ_{s,ρ_-} for the random matrix with independent columns uniformly distributed on the unit sphere of $\mathbb{R}^n$ in Theorem 2.2. The goal of this section is to show that this result can be used in a simple trick in order to obtain prediction bounds similar to [6, Theorem 2.1] without conditions on the design matrix X.

This idea is of course to use Theorem 2.4 above. However, the values of $\sigma_{\min}(X_S)$ and $\sigma_{\max}(X_{\hat{S}\cap S^c})$ are of course usually not known ahead of time and we have to provide easy to compute bounds for these quantities. The coherence $\mu(X)$ can be used for this purpose. Indeed, for any positive integer $t \leq p$ and any $T \subset \{1, \ldots, p\}$ with $|T| = t$, we have

$$\begin{aligned} \mu(X) &= \|X^tX - I\|_{1,1} \\ &= \max_{\|w\|_\infty=1} \max_{\|w'\|_1=1} w^t(X^tX - I)w' \\ &\geq \frac{1}{\sqrt{t}} \max_{\substack{\|w\|_2=1\\ \|w\|_0=t}} \max_{\substack{\|w'\|_2=1\\ \|w'\|_0=t}} w^t(X^tX - I)w'. \end{aligned}$$

Thus, we obtain that

$$1 - \mu(X)\sqrt{t} \quad \leq \sigma_{\min}(X_T) \leq \sigma_{\max}(X_T) \leq \quad 1 + \mu(X)\sqrt{t}.$$

However, the lower bound on $\sigma_{\min}(X_S)$ obtained in this manner may not be accurate enough. More precise, polynomial time computable, bounds have been devised in the litterature. The interested reader can find a very useful Semidefinite relaxation of the problem of finding the worst possible value of $\sigma_{\min}(X_T)$ over all subsets T of $\{1, \ldots, p\}$ with a given cardinal (related to the Restricted Isometry Constant) in [10].

Assuming we have a polynomial time computable a priori bound $\sigma^*_{\min}$ on $\sigma_{\min}(X_T)$ (resp. $\sigma^*_{\max}$ on $\max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T)$), our main result for the case of general design matrices is the following theorem.

Theorem 4.1. *Let X be an matrix in $\mathbb{R}^{n\times p}$ with ℓ_2-normalized columns and let X_0 be a random matrix with independent columns uniformly distributed on the unit sphere of $\mathbb{R}^n$. Let $X_\sharp$ denote the matrix corresponding to the concatenation of X and X_0, i.e. $X_\sharp = [X, X_0]$. Let $\hat{\beta}_\sharp$ denote the LASSO estimator with X replaced with $X_\sharp$ in (1.2). Let $\rho_- \in (0, 1)$. Let ν be a positive real. Assume that p_0 is such that*

$$80\ \frac{\log(p_0)}{p_0} \quad < \quad L\ \rho_- \frac{\sigma^*_{\min}}{\nu n\ \sigma^*_{\max}} \tag{4.39}$$

for some $L \in (0, 1)$. Assume moreover that p_0 is sufficiently large so that the second inequality in (2.8) is satisfied. Assume that β has support S with cardinal s and that

$$\lambda \quad \geq \quad \sigma\left(B'_{X,\nu,\rho_-}\ \sigma^*_{\max}\sqrt{2\alpha\ \log(p + p_0) + \log(2\nu n)} + \sqrt{(2\alpha + 1)\ \log(p + p_0) + \log(2)}\right)$$

with

$$B'_{X,\nu,\rho_-} \quad = \quad \frac{\nu n\ \gamma_{\nu n,\rho_-}(X)}{\rho_-\ \sigma^*_{\min} - \nu n\ \gamma_{\nu n,\rho_-}(X)\ \sigma^*_{\max}}. \tag{4.40}$$

Assume that s satisfies the first inequality in (2.8) and that $s \leq \nu n$. Then, with probability greater than $1 - p^{-\alpha} - 9p_0^{-n} - 20\ \frac{n}{\log(p_0)^{n-1}} p_0^{-1}$, we have

$$\frac{1}{2}\|X(\hat{\beta}_\# - \beta)\|_2^2 \quad \leq \quad s\ C'_{n,p,\rho_-,\alpha,\nu,\lambda} \tag{4.41}$$

with

$$C'_{n,p,\rho_-,\alpha,\nu,\lambda} \quad = \quad \frac{\lambda+\sigma\sqrt{(2\alpha+1)\ \log(p+p_0)+\log(2)}}{\rho_-\ \sigma^*_{\min}} \left(\sigma\sqrt{2\alpha\ \log(p+p_0)+\log(2\nu n)}+\lambda\right)$$

Proof. Since the index $\gamma_{s,\rho-}$ does not increase after appending a matrix with ℓ_2- normalized columns, the matrix $X_\#$ has at most the same index as that of X_0. Then (4.39) ensures that the index $\gamma_{s,\rho-}(X_\#)$ is sufficiently small. The rest of the proof is identical to the proof of Theorem 2.4. □

Appendix A. Proof of Lemma 3.3

For any index set $S \subset \{1,\dots,\kappa s\}$ with cardinal s, define R_S as the diagonal matrix with

$$(R_S)_{i,i} \quad = \quad \begin{cases} 1 \text{ if } i \in S, \\ 0 \text{ otherwise.} \end{cases}$$

Notice that we have

$$\left\|X_S^t X_S - I\right\| \quad = \quad \|R_S H R_S\|$$

with $H = X^t X - I$. In what follows, R_δ simply denotes a diagonal matrix with i.i.d. diagonal components δ_j, $j=1,\dots,\kappa s$ with Bernoulli $B(1,1/\kappa)$ distribution. Let R' be an independent copy of R. Assume that S is drawn uniformly at random among index sets of $\{1,\dots,\kappa s\}$ with cardinal s. By an easy Poissonization argument, similar to [6, Claim (3.29) p.2173], we have that

$$\mathbb{P}\left(\|R_s H R_s\| \geq r\right) \;\leq\; 2\,\mathbb{P}\left(\|RHR\| \geq r\right), \tag{A.42}$$

and by Proposition 4.1 in [7], we have that

$$\mathbb{P}\left(\|RHR\| \geq r\right) \quad\leq\quad 36\ \mathbb{P}\left(\|RHR'\| \geq r/2\right). \tag{A.43}$$

In order to bound the right hand side term, we will use [7, Proposition 4.2]. Set $r' = r/2$. Assuming that $\kappa\frac{r'^2}{e} \geq u^2 \geq \frac{1}{\kappa}\|X\|^4$ and $v^2 \geq \frac{1}{\kappa}\|X\|^2$, the right hand side term can be bounded from above as follows:

$$\mathbb{P}\left(\|RHR'\| \geq r'\right) \quad\leq\quad 3\ \kappa s\ \mathcal{V}(s,[r',u,v]), \tag{A.44}$$

with

$$\mathcal{V}(s,[r',u,v]) \quad=\quad \left(e\frac{1}{\kappa}\frac{u^2}{r'^2}\right)^{\frac{r'^2}{v^2}} + \left(e\frac{1}{\kappa}\frac{\|M\|^4}{u^2}\right)^{u^2/\|M\|^2} + \left(e\frac{1}{\kappa}\frac{\|M\|^2}{v^2}\right)^{v^2/\mu(M)^2}.$$

Using (3.21) and (3.23), we deduce that with probability at least $1-8p_0^{-n}-p_0^{-n}$, we have

$$\mathcal{V}(s,[r',u,v]) \quad=\quad \left(e\frac{1}{\kappa}\frac{u^2}{r'^2}\right)^{\frac{r'^2}{v^2}} + \left(e\frac{1}{\kappa}\frac{\left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right)^4}{u^2}\right)^{\frac{u^2}{\left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right)^2}}$$
$$+\left(e\frac{1}{\kappa}\frac{\left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right)^2}{v^2}\right)^{\frac{v^2}{\frac{1}{2}\ p_0^{-2}}}.$$

Take κ, u and v such that

$$v^2 \quad=\quad r'^2\ \frac{1}{\log(C_\kappa\ n)}$$

$$u^2 \quad=\quad C_{\mathcal{V}}\ \left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right)^2,$$

$$\kappa \quad\geq\quad e^3\ \frac{C_{\mathcal{V}}}{r'^2}\ \left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right)^2$$

for some $C_{\mathcal{V}}$ possibly depending on s. Since $\kappa s \leq C_\kappa n$, this implies in particular that

$$\kappa \;\geq\; e^3 \, \frac{C_{\mathcal{V}}}{r'^2} \left(\frac{(1+K_\varepsilon)(1+C_\kappa)}{c(1-\varepsilon)^4} \log(p_0) \right)^2 . \tag{A.45}$$

Thus, we obtain that

$$\mathcal{V}(s,[r',u,v]) \;=\; \left(\frac{1}{e^2}\right)^{\log(C_\kappa n)} + \left(\frac{r'^2}{e^2\, C_{\mathcal{V}}^2}\right)^{C_{\mathcal{V}}} + \left(\frac{\log(C_\kappa n)}{e^2\, C_{\mathcal{V}}}\right)^{\frac{2r'^2 \; p_0^2}{\log(C_\kappa n)}} .$$

Using (A.42), (A.43) and (A.44), we obtain that

$$\mathbb{P}\left(\|R_s H R_s\| \geq r'\right) \;\leq\; 2\times 36\times 3\times \kappa s \left(\left(\frac{1}{e^2}\right)^{\log(C_\kappa n)} + \left(\frac{r'^2}{e^2\, C_{\mathcal{V}}^2}\right)^{C_{\mathcal{V}}} + \left(\frac{\log(C_\kappa n)}{e^2\, C_{\mathcal{V}}}\right)^{\frac{2r'^2 \; p_0^2}{\log(C_\kappa n)}} \right).$$

Take

$$C_{\mathcal{V}} \;=\; \log(C_\kappa n) \tag{A.46}$$

and, since $p_0 > 1$ and $r \in (0,1)$, we obtain

$$\begin{aligned} &\mathbb{P}\left(\|R_s H R_s\| \geq r'\right) \\ &\leq 2\times 36\times 3\times \kappa s \left(\left(\frac{1}{e^2}\right)^{\log(C_\kappa n)} + \left(\frac{r'^2}{e^2\, \log^2(C_\kappa n)}\right)^{\log(C_\kappa n)} + \left(\frac{1}{e^2}\right)^{\frac{2r'^2 \; p_0^2}{\log(C_\kappa n)}} \right). \end{aligned} \tag{A.47}$$

Replace r' by $r/2$. Since it is assumed that $n \geq \exp(r/2)/C_\kappa$ and $p_0 \geq \sqrt{2}\log(C_\kappa n)/r$, it is sufficient to impose that

$$C_\kappa^2 n^2 \;\geq\; (2\times 36\times 3\times \kappa s\times 3)^{\frac{1}{\log(e^2)}} ,$$

in order for the right hand side of (A.47) to be less than one. Since $\kappa s \leq C_\kappa n$, it is sufficient to impose that

$$C_\kappa^2 n^2 \;\geq\; 2\times 36\times 3\times \; C_\kappa n\times 3,$$

or equivalently,

$$C_\kappa n \;\geq\; 2\times 36\times 3\times 3.$$

This is implied by (2.8) in the assumptions. On the other hand, combining (A.45) and (A.46) implies that one can take

$$\kappa \;=\; \frac{4e^3}{r^2} \left(\frac{(1+K_\varepsilon)(1+C_\kappa)}{c(1-\varepsilon)^4} \right)^2 \log^2(p_0) \log(C_\kappa n),$$

which is nothing but (2.7) in the assumptions.

References

[1] Akaike, H., A new look at the statistical model identification, IEEE Trans. Automat. Control, (1974) 19, 716–723.
[2] Bickel, P. J., Ritov, Y., Tsybakov, A. B., Simultaneous analysis of lasso and Dantzig selector. Ann. Statist. 37 (2009), no. 4, 1705–1732.
[3] Bunea, F., Tsybakov, A. and Wegkamp, M., Sparsity oracle inequalities for the LASSO, 2007, The Electronic Journal of Statistics, 169 – 194.
[4] Bühlmann, P., van de Geer, S., Statistics for High-Dimensional Data, Methods, Theory and Applications, Series: Springer Series in Statistics (2011).
[5] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics (2009) 3, 1360–1392.
[6] Candès, E. J. and Plan, Y. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 2145–2177.
[7] Chrétien, S. and Darses, S., Invertibility of random submatrices via tail decoupling and a Matrix Chernoff Inequality, Stat. and Prob. Lett., (2012), 82, no. 7, 1479–1487.
[8] Chrétien, S. and Darses, S., Sparse recovery with unknown variance: a LASSO-type approach, IEEE Trans. Info. Th., to appear.

[9] Dalalyan, A. and Tsybakov, A., Sparse Regression Learning by Aggregation and Langevin Monte-Carlo, J. Comput. System Sci. 78 (2012), pp. 1423-1443.
[10] A. d'Aspremont, F. Bach and L. El Ghaoui, Optimal Solutions for Sparse Principal Component Analysis. Journal of Machine Learning Research, 9 (2008), pp. 1269-1294.
[11] Dubhashi, D. P. and Panconesi, A., Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, Cambridge, 2009.
[12] The Risk InïňĆation Criterion for Multiple Regression. Dean P. Foster; Edward I. George. Annals of Statistics, Volume 22, Issue 4 (1994), 1947-1975.
[13] Muirhead, R., Aspects of multivariate statistical theory. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1982. xix+673 pp.
[14] Qi, F, Bounds for the ratio of two gamma functions. J. Inequal. Appl. 2010.
[15] Rigollet, P. and Tsybakov, A., Exponential Screening and optimal rates of sparse estimation, Ann. Statist., 39(2), 731-771.
[16] Rudelson, M. and Vershynin, R., Smallest singular value of a random rectangular matrix. Comm. Pure Appl. Math. 62 (2009), no. 12, 1707–1739.
[17] Estimating the dimension of a model, The Ann. of Stat., (1978) 6, 461–464.
[18] Tibshirani, R. Regression shrinkage and selection via the LASSO, J.R.S.S. Ser. B, 58, no. 1 (1996), 267–288.
[19] Tropp, J., User friendly tail bounds for sums of random matrices, Foundations of Computational Mathematics, (2012), 12, no.4, 389–434.
[20] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. Electron. J. Stat. 3 (2009) 1360–1392.
[21] Wainwright, Martin J., Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55 (2009), no. 5, 2183–2202.
[22] Zhao, P. and Yu, B., On model selection consistency of Lasso. J. Mach. Learn. Res. 7 (2006), 2541–2563.