A LECTURE ON THE LASSO ESTIMATOR AND MODEL SELECTION
based on a paper by Candès and Plan (2009)

SIMONE MARIA GIANCOLA*

February 9, 2025

CONTENTS

1 INTRODUCTION

In this forty-five minutes lecture, we will review the paper in the title. After presenting the problem, we will explain the main statement. Then, we will compare it with another perspective we saw in class (Massart 2024). If time permits, we will discuss the proof and the key lemmas.
Emphasis is on intuition and quick understanding. References are to a minimum. Computations, when performed, are explicit.

NOTATION We use bold for vectors, and curly latex for random variables. For example $\boldsymbol{\theta}$ is a random vector, while $\boldsymbol{\theta}$ is deterministic. Matrices are uppercase. The rest of the symbols are either standard or defined when discussed first. The main takeaway is this explicit distinction between what is random and what is not.

1.1 *Motivation and Setting*

Consider the following simple linear model:

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{z}, \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ is Gaussian noise. In modern settings where data is abundant, we are interested in the case when $p > n$. If the vector $\boldsymbol{\beta}$ is sparse, it would be nice to solve:

$$\arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{2} \underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2}_{\text{error}} + \underbrace{\lambda_0}_{\text{hyperparameter}} \sigma^2 \underbrace{\|\boldsymbol{b}\|_0}_{\text{complexity penalty}}, \tag{1.2}$$

*email: simonegiancola09@gmail.com

for some $\lambda_0 > 0$, but we know this is NP-hard in general. A common fix is to study the lasso, which relaxes the ℓ_0 norm into the ℓ_1 norm, obtaining:

$$\hat{\beta} \coloneqq \arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda_p \|b\|_1, \qquad \lambda_p > 0. \tag{1.3}$$

The model is now convex and has an algorithmic solution. Intuitively, we still have a regularization that penalizes "large" solutions.

Main finding

We will see that under appropriate conditions this relaxation is not harmful. Indeed, the lasso finds vectors that attain an error approximately as good as an estimator obtained from sparse subsets of explanatory variables. Furthermore, it is approximately as good as a lasso that does not "see" the noise.

Let us then introduce our main assumption, which bounds the redundancy of the data matrix.

Definition 1.4 (Coherence). *Given a data matrix X, its coherence is the maximal alignment of columns, denoted as $\mu(X) \coloneqq \sup_{1 \leqslant i < j \leqslant p} |\langle x_i, x_j \rangle|$.*

Assumption 1.5. *The data matrix X is such that its columns have unit ℓ_2 norm and for some a_0 its coherence is bounded from above by $a_0/\log p$.*

This assumption is minimal: we can always rescale the matrix as to have unit columns. Moreover, in the classic modeling case where we take $\mathbf{X}$ to be random and Gaussian, after standardizing we have about $\sqrt{2\log p/n}$ coherence, which is below our assumption for moderately large sample size n.

2 Main result

What is the best benchmark to compare lasso with? If we had exponential time at disposal, we could check each subset of explanatory variables and project over it. Mathematically, we would take each subset $\mathbb{M} \subset [p]$ and compare $\beta_{\mathbb{M}} = P_{\mathbb{M}} y$, which on average attains an error:

$$\mathbb{E}\left[\|X\beta - X\beta_{\mathbb{M}}\|\right]_2^2 = \|(I_p - P_{\mathbb{M}})X\beta\|_2^2 + |\mathbb{M}|\sigma^2. \tag{2.1}$$

Then, we would search for the minimum over this 2^p dimensional space:

$$\min_{\mathbb{M} \subset [p]} \underbrace{\|(I_p - P_{\mathbb{M}})X\beta\|_2^2}_{\text{error on model subspace}} + \sigma^2 \underbrace{|\mathbb{M}|}_{\text{model size}}. \tag{2.2}$$

If we could find the best b in equation equation 1.2, we would have a solution that is close to the best subset, and attains the best bias variance trade-off. However, we want an algorithm.

To model this situation, we consider the *best dimensional subset model*. Let us assume that $\mathbb{J}$ is any set attaining the value in equation 2.2, we say:

- it is uniformly distributed over $[p]$;
- it has size $|\mathbb{J}| = s$;
- it is associated to a vector β_0 defined via the optimal projection $X\beta_0 = P_{\mathbb{J}} X\beta$.

Remark 2.3. *Equivalently, we do not place any prior information on where it might lie.*

Under this model, we can state the result of (Candès and Plan 2009, thm. 1.4):

Theorem 2.4. *Let X satisfy assumption 1.5 and β_0 be from the best s-dimensional subset model. Suppose there exist a constant such that $s \leqslant c_0 p/\|X\|_{\text{op}}^2 \log p$. Then, if we choose $\lambda_p = 2\sqrt{2\log p}$ the lasso solution of equation 1.3 satisfies the following inequality:*

$$\|X\beta - X\hat{\beta}\|_2^2 \leqslant C\left[\min_{\mathbb{M}\subset[p]} \|(I_p - P_{\mathbb{M}})X\beta\|_2^2 + \lambda_p^2\sigma^2 C_0|\mathbb{M}|\right], \tag{2.5}$$

implicit hyperparameter ↓ $\lambda_p^2\sigma^2$; *deterministic error* ↑ $\|X\beta - X\hat{\beta}\|_2^2$; *corrected complexity penalty* ↑ $C_0|\mathbb{M}|$

with probability larger than $1 - 6p^{-2\log 2} - 1/p\sqrt{2\pi\log p}$, *and* C, C_0 *explicit constants.*

Sharpness

The assumptions are nearly optimal. In (Candès and Plan 2009, sec. 2) the authors bring counterexamples to low coherence or probability one settings. In simple words, there are always vectors that can make the statement false if we have high redundancy or look at the whole sample space.

A different formulation of this result might be more enlightening. If we let $s^\star$ be the maximal sparsity allowed, and consider the signs $\mathbb{A}_s := \{r \in \{\pm 1, 0\}^p : \sum |r_j| = s\}$, we have that for some subset $\mathbb{B} \subset \mathbb{A}_s$ with probability larger than $1 - O(1/p)$:

$$\|X\beta - X\hat{\beta}\|_2^2 \leqslant \min_{s\leqslant s^\star} \min_{b:\operatorname{sgn}(b)\in\mathbb{B}} C\left[\|X\beta - Xb\| + C_0\lambda_p^2 s\sigma^2\right]. \tag{2.6}$$

In particular, the set $\mathbb{B}$ is almost $\mathbb{A}_s$, with an explicit bound on their ratio: $|\mathbb{B}|/|\mathbb{A}_s| \geqslant 1 - O(1/p)$. In simple words, with large probability the predictive power of the lasso estimator is almost as good as the predictive power of the best deterministic lasso estimator obtained among all sparse models, with a slightly larger regularization, outside of a vanishing set of signs. The statement is just the same but this result is implicit in the proof technique.

2.1 *Comparison with the literature*

With no intention to be exhaustive, we compare theorem 2.4 with the results in (Massart and Meynet 2010) which are also reported in (Massart 2024). In both, the setting is a lot more general, so we will adapt the notation and the results. Consider a dictionary $\mathbb{D}_p$ of explanatory features that is finite and of dimension p, all such that their ℓ_2 norm is bounded above by one. We define the projected norm in its span as:

$$\|h\|_{\mathbf{L}_1(\mathbb{D}_p)} := \inf_{\theta\in\mathbb{R}^p : \sum_{j=1}^p \theta_j x_j = h} \|\theta\|_1. \tag{2.7}$$

If the underlying function is linear, we have the same model. The authors define the lasso estimator in the general case as:

$$\hat{\mathbf{f}} := X\hat{\beta} := \operatorname*{arg\,min}_{h\in\mathbf{L}_1(\mathbb{D}_p)} \|\mathbf{y} - h\|_2^2 + \lambda_p\|h\|_{\mathbf{L}_1(\mathbb{D}_p)}. \tag{2.8}$$

Then, a computation (see (Massart and Meynet 2010, sec 3.2)) shows that this estimator coincides with the lasso of equation 1.3, in the sense that $\hat{\mathbf{f}} = X\hat{\beta}$, so we have the same representation but on the space of observations. The theorem we will report has no assumptions (apart from Gaussian noise), but does a different comparison. We take it from (Massart and Meynet 2010, thm. 3.2), a slightly different formulation is in (Massart 2024, thm. 24).

Theorem 2.9. *Suppose the explanatory features* $\{x_j\}_{j=1}^p$ *are such that* $\max_j \|x_j\|_2 \leqslant 1$. *Let* $\lambda_p \geqslant 4\sigma/\sqrt{n}(\sqrt{\log p} + 1)$. *Then, for all* $\rho > 0$ *with probability larger than* $1 - 3.4e^{-\rho}$ *it holds that:*

$$\|\mathbf{y} - X\hat{\beta}\|_2^2 + \lambda_p\|\hat{\beta}\|_1 = \left\|X\beta - \hat{\mathbf{f}}\right\|_2^2 + \lambda_p\left\|\hat{\mathbf{f}}\right\|_{\mathbf{L}_1(\mathbb{D}_p)} \tag{2.10}$$

$$\leqslant C\left[\inf_{h\in\mathbf{L}_1(\mathbb{D}_p)} \{\|X\beta - h\|\} + \lambda_p\|h\|_{\mathbf{L}_1(\mathbb{D}_p)}\right] + \frac{\lambda_p\sigma}{\sqrt{n}}(1+\rho) \tag{2.11}$$

$$= C\left[\inf_{\tilde{\beta}\in\mathbb{R}_p} \|X\beta - X\tilde{\beta}\|_2^2 + \lambda_p\|\tilde{\beta}\|_1\right] + \frac{\lambda_p\sigma}{\sqrt{n}}(1+\rho). \tag{2.12}$$

Moreover, we can integrate the probability bound in ρ to obtain a version in expectation:

$$\mathbb{E}\left[\left\|X\beta - \hat{\mathbf{f}}\right\|_2^2 + \lambda_p \left\|\hat{\mathbf{f}}\right\|_{\mathbf{L}_1(\mathbb{D}_p)}\right] \leqslant C\left[\inf_{h\in\mathbf{L}_1(\mathbb{D}_p)}\left\{\|X\beta - h\|_2^2 + \lambda_p\|h\|_{\mathbf{L}_1(\mathbb{D}_p)}\right\} + \frac{\lambda\sigma}{\sqrt{n}}\right]. \tag{2.13}$$

In words, without any coherence assumption, we get that up to constants the lasso estimator is almost as good as the deterministic lasso estimator.

Sharpness

From (Massart and Meynet 2010, rem. 3.3 point 4) we get that this bound is optimal. In particular, there exists a regime of parameters where we can upper bound the last result by a quantity of the same order as the min-max lower bound over ℓ_1 balls.

Remark 2.14. *It is important to notice that differently from theorem 2.4 we can get a bound on the expectation by integrating. In the former case, the dependence on p of the probability is implicit in the upper bound, while here z is a generic scalar.*

Theorems 2.4 and 2.9 are quite different in nature. In (Candès and Plan 2009) the bound sought is with respect to the oracle, while the objective of Massart and Meynet (2010) is to compare with a non-noisy ideal lasso estimator. We can see the two approaches as complementary results, telling us that under all the assumptions the lasso estimator in the linear case with a matching linear model is approximately model selection optimal and careless with respect to noise, with these statements getting asymptotically better as the dimension increases. Alternatively, we know there are barriers to these results thanks to the counterexamples: if either condition or scaling does not hold, then we are not guaranteed to claim that the lasso estimator performs well according to either of the two notions.

3 PROOF OF THE MAIN RESULT

We now present the proof assuming the lemmas hold. These auxiliary results are discussed later in section 4.

proof of theorem 2.4. **(Preliminary)** Without loss of generality, we let $\sigma = 1$. To restrict a matrix to a set of columns $\mathbb{M} \subset [p]$ we write $X_{\mathbb{M}}$. Throughout, $\lambda_p = \sqrt{2\log p}$. It is also useful to recall that the lasso functional

$$\mathcal{K}(\mathbf{y}, b) := \frac{1}{2}\|\mathbf{y} - Xb\|_2^2 + 2\lambda_p\|b\|_1 \tag{3.1}$$

has a nice sub-gradient with respect to b:

$$\partial_2\mathcal{K}(\mathbf{y}, b) = X^\dagger(Xb - \mathbf{y}) + 2\lambda_p\epsilon, \qquad \epsilon_j = \begin{cases}\operatorname{sgn}(b_i) & b_i \neq 0\\ (-1,1) & b_i = 0.\end{cases} \tag{3.2}$$

In particular, we denoted the sub-gradient at b with ∂_2. Without regard to the specific coherence assumption, we have some *a priori* bounds:

(B1) it holds that $\left\|X^\dagger(\mathbf{y} - X\hat{\beta})\right\|_\infty \leqslant 2\lambda_p$;

(B2) when noise is Gaussian $\left\|X^\dagger\mathbf{z}\right\|_\infty \leqslant \sqrt{2}\lambda_p$ with probability $1 - 1/p\sqrt{2\pi\log p}$.

In particular, combining (B1)-(B2) we have with high probability:

$$\left\|X^\dagger X(\beta - \hat{\beta})\right\|_\infty \leqslant (\sqrt{2} + 2)\lambda_p. \tag{3.3}$$

Coherence combined with the scalings instead give us two other high probability results. Suppose the best s-dimensional subset model has support $\mathbb{J}$, then:

(R1) *[invertibility]* $\left\|(X_{\mathbb{J}}X_{\mathbb{J}})^{-1}\right\|_\infty \leqslant 2$;

(R2) *[complementary size]* seeing sgn as a function acting on vectors entry-wise:

$$\left\|X_{\mathbb{J}^c}^\dagger X_{\mathbb{J}}\left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1}X_{\mathbb{J}}^\dagger\mathbf{z}\right\|_\infty + 2\lambda_p\left\|X_{\mathbb{J}^c}^\dagger X_{\mathbb{J}}\left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1}\operatorname{sgn}(\beta_{0,\mathbb{J}})\right\|_\infty \leqslant (2 - \sqrt{2})\lambda_p. \tag{3.4}$$

In what follows, we assume all of these hold, and continue the proof. In section 4 we will estimate the probability with which they hold jointly, which matches that in the statement of theorem 2.4.

(Main) Since $\hat{\beta}$ is a minimizer of the lasso functional we have that $\mathcal{K}(\mathbf{y}, \hat{\beta}) \leqslant \mathcal{K}(\mathbf{y}, \beta_0)$. From this, we can deduce with by opening $\|\mathbf{y} - X\hat{\beta}\|_2^2$ and $\|\mathbf{y} - X\beta_0\|_2^2$ that:

$$\frac{1}{2}\|X\hat{\beta} - X\beta\|_2^2 \leqslant \frac{1}{2}\|X\beta_0 - X\beta\|_2^2 + \langle \mathbf{z}, X\hat{\beta} - X\beta_0 \rangle + 2\lambda_p(\|\beta_0\|_1 - \|\hat{\beta}\|_1). \tag{3.5}$$

The parts where β_0 is active and not are crucial. Let us denote $\mathbf{h} := \hat{\beta} - \beta_0 = [\mathbf{h}_{\mathbb{J}} \mid \mathbf{h}_{\mathbb{J}^c}]$. On the inner product, we can observe that:

$$\langle \mathbf{z}, X\hat{\beta} - \beta_0 \rangle = \left\langle X^\dagger \mathbf{z}, \mathbf{h} \right\rangle = \left\langle X^\dagger \mathbf{z}, \mathbf{h}_{\mathbb{J}} \right\rangle + \left\langle X^\dagger \mathbf{z}, \mathbf{h}_{\mathbb{J}^c} \right\rangle. \tag{3.6}$$

In particular, we bound the latter term using the (B2), i.e. that $\left\|X^\dagger \mathbf{z}\right\|_\infty \leqslant 2\lambda_p$. In the ℓ_1 norms we seek a cancellation. We can use the difference vector to write:

$$\|\hat{\beta}\|_1 = \|\beta_{0,\mathbb{J}} + \mathbf{h}_{\mathbb{J}}\|_1 + \|\mathbf{h}_{\mathbb{J}^c}\|_1. \tag{3.7}$$

Now notice that by the fact that the support is of β_0 and the definition of $\mathbf{h}$:

$$\forall j \in \mathbb{J}, \qquad |\hat{\beta}_j| = |\beta_{0,j} + h_j| \geqslant |\beta_{0,j}| + \operatorname{sgn}(\beta_{0,j})h_j, \tag{3.8}$$

which is easily checked by just plugging the case $\beta_{0,j}$ positive or negative. Summing up this inequality across $j \in \mathbb{J}$ we can cancel the $\|\beta_0\|_1$ term above, obtaining in exhange an inner product $\langle \mathbf{h}, \operatorname{sgn}(\beta_0) \rangle$. Recollecting all esitmates in the main equation we have:

$$\frac{1}{2}\|X\hat{\beta} - X\beta\|_2^2 \leqslant \frac{1}{2}\|X\beta_0 - X\beta\|_2^2 + \langle \mathbf{h}_{\mathbb{J}}, \mathbf{v} \rangle - (2 - \sqrt{2})\lambda_p \|\mathbf{h}_{\mathbb{J}^c}\|_1, \qquad \mathbf{v} := X_{\mathbb{J}}^\dagger \mathbf{z} - 2\lambda_p \operatorname{sgn}(\beta_{0,\mathbb{J}}). \tag{3.9}$$

The rest of the argument is aimed at upper bounding the inner product using the other results to cancel out the annoying ℓ_1 norm of the difference vector.

(Upper bound on inner product) We seek an upper bound in terms of infinity norms that make the complementary size condition appear, as well as the lasso bound in equation 3.3. The rest is just a matter of reordering terms. We will then reuse the distinction between $\mathbb{J}$ and its complement but in reverse. Implicitly, the invertibility condition (R1) allows us to use the inverse of the matrix restricted to $\mathbb{J}$. By this, we can inject it in the inner product:

$$\langle \mathbf{h}_{\mathbb{J}}, \mathbf{v} \rangle = \left\langle X_{\mathbb{J}}^\dagger X_{\mathbb{J}} \mathbf{h}_{\mathbb{J}}, \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{v} \right\rangle \tag{3.10}$$

$$= \underbrace{\left\langle X_{\mathbb{J}}^\dagger X \mathbf{h}, \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{v} \right\rangle}_{:= a_{\mathsf{hard}}} + \underbrace{\left\langle X_{\mathbb{J}}^\dagger X_{\mathbb{J}^c} \mathbf{h}_{\mathbb{M}_0^c}, \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{v} \right\rangle}_{:= a_{\mathsf{easy}}}, \tag{3.11}$$

where we just used the decomposition $X\mathbf{h} = X_{\mathbb{J}}\mathbf{h}_{\mathbb{J}} + X_{\mathbb{J}^c}\mathbf{h}_{\mathbb{J}^c}$. By the definition of $\mathbf{v}$ the second term is easy and gives us what we wanted:

$$a_{\mathsf{easy}} \leqslant \|\mathbf{h}_{\mathbb{J}^c}\|_1 \left\| X_{\mathbb{J}}^\dagger X_{\mathbb{J}^c} \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{v} \right\|_\infty \qquad \mathbf{v} = X_{\mathbb{J}}^\dagger \mathbf{z} - 2\lambda_p \operatorname{sgn}(\beta_{0,\mathbb{J}}), \tag{3.12}$$

$$\leqslant \|\mathbf{h}_{\mathbb{J}^c}\|_1 \left\{ \left\| X_{\mathbb{J}}^\dagger X_{\mathbb{J}^c} \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{z} \right\|_\infty + 2\lambda_p \left\| X_{\mathbb{J}}^\dagger X_{\mathbb{J}^c} \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \operatorname{sgn}(\beta_{0,\mathbb{J}}) \right\|_\infty \right\} \tag{3.13}$$

$$\leqslant \|\mathbf{h}_{\mathbb{J}^c}\|_1 (2 - \sqrt{2})\lambda_p, \tag{3.14}$$

comparing with the intermediate estimate in equation 3.9 we have cancelled the last term. What is missing is the hard term. To bound it, we decompose it again into the lasso vs true model contribution and best-dimensional model vs true model contribution. Namely, we rewrite $\mathbf{h} = \hat{\beta} - \beta_0 = \hat{\beta} - \beta + \beta - \beta_0$ and split again:

$$a_{\mathsf{hard}} = a_{\mathsf{hard}}^{\mathsf{old}} + a_{\mathsf{hard}}^{\mathsf{new}}, \tag{3.15}$$

$$a_{\mathsf{hard}}^{\mathsf{old}} := \left\langle X_{\mathbb{J}}^\dagger X(\hat{\beta} - \beta), \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{v} \right\rangle \tag{3.16}$$

$$a_{\mathsf{hard}}^{\mathsf{new}} := \left\langle X_{\mathbb{J}}^\dagger X(\beta - \beta_0), \left[X_{\mathbb{J}}^\dagger X_{\mathbb{J}}\right]^{-1} \mathbf{v} \right\rangle. \tag{3.17}$$

The notation old, new refers to the fact that the old term is part of the proof of an earlier theorem in (Candès and Plan 2009). We bound the absolute value of the two terms. In both cases this reduces to applying a simple inequality to $\mathbf{v}$. For the old component:

$$\begin{aligned}
\mathsf{a}_{\mathsf{hard}}^{\mathsf{old}} &\leqslant \left\| \boldsymbol{X}_{\mathbb{J}}^{\dagger} \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right\|_{\infty} \left\| \left[\boldsymbol{X}_{\mathbb{J}}^{\dagger} \boldsymbol{X}_{\mathbb{J}} \right]^{-1} \right\|_{\mathsf{op}} \|\mathbf{v}\|_{\infty} && (3.18)\\
&\leqslant 2(2+\sqrt{2}) s \lambda_p \|\mathbf{v}\|_{\infty} && \text{by equation 3.3 and (R1);} \quad (3.19)\\
&\leqslant 2(2+\sqrt{2}) s \lambda_p (\sqrt{2}\lambda_p + 2\lambda_p) && \text{by rough bound on } \mathbf{v}. \quad (3.20)
\end{aligned}$$

For the new term, we make non-stochastic difference between to the true model and the best-dimensional model appear:

$$\begin{aligned}
\mathsf{a}_{\mathsf{hard}}^{\mathsf{new}} &\leqslant \sqrt{2} \|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2 \|\mathbf{v}\|_2 && (3.21)\\
&\leqslant \frac{\sqrt{2}}{2} \left[\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2 + \|\mathbf{v}\|_2^2 \right] && \text{Young's inequality;} \quad (3.22)\\
&\leqslant \frac{\sqrt{2}}{2} \left[\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_2^2 + (2+\sqrt{2})^2 s + \lambda_p^2 \right] && \text{by rough bound on } \mathbf{v}. \quad (3.23)
\end{aligned}$$

(Finalization) Using the results in the step above, we simplify equation 3.9 with the following upper bound:

$$\frac{1}{2} \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leqslant \frac{1+\sqrt{2}}{2} \|\boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + (4+\sqrt{2})(1+\sqrt{2})^2 \lambda_p^2 s. \tag{3.24}$$

Substituting the scaling of λ_p we have the bound for some explicit constants. □

4 AUXILIARY STATEMENT

In this section we report the proof of (B1)-(B2)-equation 3.3-(R1)-(R2) in the premise of Candès and Plan (2009, thm. 1.4).

The preliminary bounds follow by standard techniques.

Lemma 4.1. *The following three facts are true.*

1. *The lasso solution satisfies:*
$$\left\| \boldsymbol{X}^{\dagger}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right\|_{\infty} \leqslant 2\lambda_p. \tag{4.2}$$
2. *If noise is Gaussian, the following inequality is true with probability* $1 - 1/p\sqrt{2\pi \log p}$. *:*
$$\left\| \boldsymbol{X}^{\dagger}\mathbf{z} \right\|_{\infty} \leqslant \sqrt{2}\lambda_p. \tag{4.3}$$
3. *The lasso satisfies a bound a priori:*
$$\left\| \boldsymbol{X}^{\dagger}\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right\|_{\infty} \leqslant (\sqrt{2}+2)\lambda_p, \qquad \textit{with probability } 1 - \frac{1}{p\sqrt{2\pi \log p}}. \tag{4.4}$$

The three results are a rewriting of (B1)-(B2) and equation 3.3.

Proof. **(Claim #1)** The claim follows by the form of the lasso subgradient in equation 3.2. In particular, it suffices to notice that $\|\boldsymbol{\epsilon}\|_{\infty} \leqslant 1$.
(Claim #2) We write a rough union bound on the random variables $\langle \boldsymbol{x}_j, \mathbf{z} \rangle$ for $j \in [p]$, which are all standard Gaussians. Indeed, we have:

$$\mathbb{P}\left[\left\| \boldsymbol{X}^{\dagger}\mathbf{z} \right\|_{\infty} \geqslant t \right] \leqslant \sum_{j=1}^{p} \mathbb{P}\left[|\langle \boldsymbol{x}_j, \mathbf{z} \rangle| \geqslant t \right] \leqslant 2p \frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}t}. \tag{4.5}$$

Plugging $t = \sqrt{2}\lambda_p$ allows us to conclude.
(Claim #3) Substituting $\boldsymbol{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{z}$ and using #1-#2 we obtain the desired inequality. □

To prove (R1) and (R2) we use a result of Tropp (2008) and some lemmas. Let us introduce first some useful notation. Let $\{\mathrm{r}_j\}_{j=1}^p \sim \mathsf{Ber}(\frac{s}{p})^{\otimes p}$ be a collection of independent Bernoulli random variables. Using their realizations, we construct a set of predictors $\mathbb{M} = \{j : \mathrm{r}_j = 1\}$, with the property that $\mathbb{E}\left[|\mathbb{M}|\right] = s$. We also construct a matrix:

$$\mathbf{R} := \mathsf{diag}(\mathrm{r}_1, \ldots, \mathrm{r}_p). \tag{4.6}$$

For this subsection only we use the shorthand $q := 2\log p$. Adapting the notations, at the end of (Tropp 2008, sec. 4) we find a bound that holds for matrices $\mathbf{A}$ decomposed into their diagonal and off diagonal part as $\boldsymbol{A} = \boldsymbol{H} + \boldsymbol{D}$. For us, it is instrumental to apply it to $\boldsymbol{I}_p - \boldsymbol{X}^\dagger\boldsymbol{X}$ which has $\boldsymbol{D} = \mathbf{0}_{p\times p}, \boldsymbol{A} = \boldsymbol{H}$ in this decomposition. Reporting the result:

$$\mathsf{E}\left[\|\mathbf{R}\boldsymbol{H}\mathbf{R}\|_{\mathsf{op}}^q\right]^{\frac{1}{q}} \leqslant 15q\mathbb{E}\left[(\|\mathbf{R}\boldsymbol{H}\mathbf{R}\|_\infty)^q\right]^{\frac{1}{q}} + 12\sqrt{\delta q}\mathsf{E}\left[\max_{\text{cols } j\in[p]}\left\|(\mathbf{R}\boldsymbol{H})_j\right\|_2\right] + 2\delta\mathsf{E}\left[\|\boldsymbol{H}\|_{\mathsf{op}}\right], \qquad \delta := \frac{s}{p}. \tag{4.7}$$

Furthermore, Tropp (2008) justifies the following inequality:

$$\left(\max_{\text{cols } j\in[p]}\left\|(\mathbf{R}\boldsymbol{H})_j\right\|_2\right) \leqslant \|\boldsymbol{X}\|_{\mathsf{op}}. \tag{4.8}$$

We combine these two arguments with the following observations:

- almost surely $\|\mathbf{R}\boldsymbol{H}\mathbf{R}\|_{\mathsf{op}} \leqslant \mu(\boldsymbol{X})$ the coherence by simply unrolling the definitions and bounding the Bernoullis by one;
- $\|\boldsymbol{H}\|_{\mathsf{op}} \leqslant \max\{\|\boldsymbol{X}\|_{\mathsf{op}}^2 - 1, 1\} \leqslant \|\boldsymbol{X}\|_{\mathsf{op}}^2$ since $\|\boldsymbol{X}\|_{\mathsf{op}} \geqslant 1$ by the unit norm assumption.

Using these observations and supposing $s\|\boldsymbol{X}\|_{\mathsf{op}}^2/p \leqslant 1/4$ gives us after some algebraic manipulations that:

$$\mathsf{E}\left[\|\mathbf{R}\boldsymbol{H}\mathbf{R}\|_{\mathsf{op}}^q\right]^{\frac{1}{q}} \leqslant 30\mu(\boldsymbol{X})\log p + (12\sqrt{2}\log p + 1)\sqrt{\frac{s\|\boldsymbol{X}\|_{\mathsf{op}}^2}{p}}. \tag{4.9}$$

Observing that $\mathbf{R}$ just selects the random set $\mathbb{M}$ it takes some moments to realize that we can rewrite the left hand side and make a slightly worse bound as follows:

$$\mathsf{E}\left[\left\|\boldsymbol{X}_{\mathbb{M}}^\dagger\boldsymbol{X}_{\mathbb{M}} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q\right]^{\frac{1}{q}} \leqslant 30\mu(\boldsymbol{X})\log p + 13\sqrt{\frac{s\sqrt{2}\log p\|\boldsymbol{X}\|_{\mathsf{op}}^2}{p}}. \tag{4.10}$$

Moreover, we also borrow (Tropp 2008, cor. 5.1), which states that:

$$\mathsf{E}\left[\max_{j\in\mathbb{M}^c}\left\|\boldsymbol{X}_{\mathbb{M}}^\dagger\boldsymbol{x}_j\right\|_2^q\right]^{\frac{1}{q}} \leqslant 4\mu(\boldsymbol{X})\sqrt{\log p} + \sqrt{\frac{s\|\boldsymbol{X}\|_{\mathsf{op}^2}}{p}} \tag{4.11}$$

Corollary 4.12. *A Poissonization argument letting the set be random gives the same bounds of equations 4.10-4.11 with an added $2^{1/q}$ on the right hand side. To see the full proof strategy, we refer to (Candès and Plan 2009, lem. 3.6).*

Proof. We only show the one result. The other is analogous. Let $\mathbb{J}$ be now a random set of dimension s taken uniformly, and $\mathbb{M}$ be the Bernoulli model. It suffices to show that:

$$\mathsf{E}\left[\left\|\boldsymbol{X}_{\mathbb{J}}^\dagger\boldsymbol{X}_{\mathbb{J}} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q\right] \leqslant 2\mathbb{E}\left[\left\|\boldsymbol{X}_{\mathbb{M}}^\dagger\boldsymbol{X}_{\mathbb{M}} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q\right]. \tag{4.13}$$

Then we can express the left-hand side as:

$$\begin{aligned}
\mathsf{E}\left[\left\|\boldsymbol{X}_{\mathbb{M}}^\dagger\boldsymbol{X}_{\mathbb{M}} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q\right] &= \sum_{k=0}^p \mathsf{P}\left[|\mathbb{M}| = k\right]\mathsf{E}\left[\left\|\boldsymbol{X}_{\mathbb{M}}^\dagger\boldsymbol{X}_{\mathbb{M}} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q \mid |\mathbb{M}| = k\right] && (4.14)\\
&\geqslant \sum_{k=s}^p \mathsf{P}\left[|\mathbb{M}| = k\right]\mathsf{E}\left[\left\|\boldsymbol{X}_{\mathbb{M}}^\dagger\boldsymbol{X}_{\mathbb{M}} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q \mid |\mathbb{M}| = k\right] && (4.15)\\
&\geqslant \sum_{k=s}^p \mathsf{P}\left[|\mathbb{M}| = k\right]\mathsf{E}\left[\left\|\boldsymbol{X}_{\mathbb{M}_k}^\dagger\boldsymbol{X}_{\mathbb{M}_k} - \boldsymbol{I}_p\right\|_{\mathsf{op}}^q\right] && (4.16)
\end{aligned}$$

where $\mathbb{M}_k$ is a uniform sample of a size k subset of $[p]$. We may conclude by observing that:

- the quantity $\left\| \boldsymbol{X}_{\mathbb{M}_k}^\dagger \boldsymbol{X}_{\mathbb{M}_k} - \boldsymbol{I}_p \right\|_{\mathsf{op}}^q$ is increasing in $|\mathbb{M}_k| = k$;
- by the symmetry of Bernoullis the median is $\mathbb{E}\left[|\mathbb{M}|\right] = s = \mathrm{Med}(|\mathbb{M}|)$;

Only the first claim needs further justification. We can see this simply by the Cauchy interlacing theorem. If we start from the full matrix $k = p$, at $k-1$ we will have removed one row and one column corresponding to the lost index j. By the Cauchy interlacing theorem the maximum eigenvalue of the original matrix upper bounds the new matrix. Iterating, this property is maintained. □

The invertibility condition is now an almost direct consequence. For the complementary size we will need some more work.

Proposition 4.17 (Invertibility)**.** *If the coherence satisfies assumption 1.5 and s satisfies the bound in theorem 2.4 we have that:*

$$\left\| \left[\boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}} \right]^{-1} \right\|_{\mathsf{op}} \leqslant 2, \qquad \text{with probability } 1 - p^{-2\log 2}. \tag{4.18}$$

Namely, condition (R1) holds with high probability.

Proof. Under our conditions, the right-hand side of equation 4.10 is bounded above by $1/4$. Let $\mathrm{z} \coloneqq \left\| \boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}} - \boldsymbol{I}_p \right\|_{\mathsf{op}}$. Since it is the operator norm of a Hermitian matrix it is its maximum eigenvalue, in the form $\max_{j\in[p]} |\lambda_j^{(\boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}})} - 1|$. Then, if $\mathrm{z} \leqslant 1/2$ the eigenvalues of the original matrix are bounded inside $[1/2, 3/2]$. By a simple Markov's inequality, we have that:

$$\mathbb{P}\left[\mathrm{z} \geqslant \frac{1}{2} \right] \leqslant 2^q \mathbb{E}\left[|\mathrm{z}|^q \right] \leqslant \frac{1}{2^q}, \tag{4.19}$$

where we used the assumptions to bound the estimate of before. This concludes the proof choosing $q = 2\log p$. □

Let us move to the irrepresentability condition (R2). We make use of the following result.

Lemma 4.20. *Let $(\boldsymbol{w}_j)_{j\in\mathbb{T}}, (\boldsymbol{v}_j)_{j\in\mathbb{T}}$ be collections of vectors with $\boldsymbol{w}_j \in \ell_2(\mathbb{M})$ and $\boldsymbol{v}_j \in \mathbb{R}^n$ respectively. They can also be random, the importance is that they are independent from the other random variables. Defining $\mathrm{z}_0 \coloneqq \max_{j\in\mathbb{T}} |\langle \boldsymbol{w}_j, \mathrm{sgn}(\boldsymbol{\beta}_{\mathbb{M}}) \rangle$ and $\mathrm{z}_1 \coloneqq \max_{j\in\mathbb{T}} |\langle \boldsymbol{v}_j, \boldsymbol{z} \rangle|$ the following two bounds hold:*

$$\mathbb{P}\left[\mathrm{z}_0 \geqslant t\right] \leqslant 2|\mathbb{T}| \mathrm{e}^{-\frac{t^2}{2\kappa^2}} \qquad \forall \kappa \geqslant \max_{j\in\mathbb{J}} \left\| \boldsymbol{w}_j \right\|_2; \tag{4.21}$$

$$\mathbb{P}\left[\mathrm{z}_1 \geqslant t\right] \leqslant 2|\mathbb{T}| \mathrm{e}^{-\frac{t^2}{2\eta^2}} \qquad \forall \eta \geqslant \max_{j\in\mathbb{J}} \left\| \boldsymbol{v}_j \right\|_2. \tag{4.22}$$

Proof. The first bound is an application of Hoeffding's inequality on each j, bounding the denominator by the maximum, and applying a union bound. The second inequality follows since each variable is Gaussian, bounding by the maximum, and applying a union bound again. □

Proposition 4.23 (Irrepresentability)**.** *If the coherence satisfies assumption 1.5 and s satisfies the bound in theorem 2.4 we have that with probability larger than $1 - 6p^{-2\log 2}$:*

$$\left\| \boldsymbol{X}_{\mathbb{J}^\mathrm{c}}^\dagger \boldsymbol{X}_{\mathbb{J}} \left[\boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}} \right]^{-1} \boldsymbol{X}_{\mathbb{J}}^\dagger \mathbf{z} \right\|_\infty + 2\lambda_p \left\| \boldsymbol{X}_{\mathbb{J}^\mathrm{c}}^\dagger \boldsymbol{X}_{\mathbb{J}} \left[\boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}} \right]^{-1} \mathrm{sgn}(\boldsymbol{\beta}_{0,\mathbb{J}}) \right\|_\infty \leqslant (2 - \sqrt{2})\lambda_p. \tag{4.24}$$

Namely, condition (R2) holds with high probability.

Proof. **(Defining objects)** Let us define for $j \in \mathbb{J}^\mathrm{c}$ the random variables:

$$\mathrm{z}_{0,j} \coloneqq \boldsymbol{x}_j^\dagger \boldsymbol{X}_{\mathbb{J}} \left[\boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}} \right]^{-1} \mathrm{sgn}(\boldsymbol{\beta}_{0,\mathbb{J}}), \qquad \mathrm{z}_{1,j} \coloneqq \boldsymbol{x}_j^\dagger \boldsymbol{X}_{\mathbb{J}} \left[\boldsymbol{X}_{\mathbb{J}}^\dagger \boldsymbol{X}_{\mathbb{J}} \right]^{-1} \boldsymbol{X}_{\mathbb{J}}^\dagger \mathbf{z}. \tag{4.25}$$

Now, we just want to show that with the claimed probability:

$$2\lambda_p \mathrm{z}_0 + \mathrm{z}_1 \leqslant (2 - \sqrt{2})\lambda_p, \qquad \mathrm{z}_0 \coloneqq \max_{j\in\mathbb{J}^\mathrm{c}} |\mathrm{z}_{0,j}|,\ \mathrm{z}_1 \coloneqq \max_{j\in\mathbb{J}^\mathrm{c}} |\mathrm{z}_{1,j}|. \tag{4.26}$$

To satisfy such inequality, we check that $z_0 \leqslant 1/4$ and $z_1 \leqslant (3/2 - \sqrt{2})\lambda_p$. Let us assign the following values to the placeholder variables of lemma 4.20:

$$w_j \coloneqq \left[\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{X}_{\mathbb{J}}\right]^{-1}\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{x}_j, \qquad \mathbf{v}_j \coloneqq \boldsymbol{X}_{\mathbb{J}}\left[\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{X}_{\mathbb{J}}\right]^{-1}\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{x}_j, \qquad \text{for all } j \in \mathbb{J}^{\mathrm{c}}. \tag{4.27}$$

We will also need to recall the definition of z in the proof of proposition 4.17. Then, defining the event $E \coloneqq \{z \leqslant 1/2\} \cap \left\{\max_{j\in\mathbb{J}^{\mathrm{c}}} \left\|\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{x}_j\right\|_2 \leqslant \gamma\right\}$ we know that by the first event considered all eigenvalues of $\boldsymbol{X}_{\mathbb{J}}$ are in the interval $[1/\sqrt{2}, \sqrt{3/2}]$ and that the operator norm of the inverse is bounded above by 2.
(Conditioning on a large probability event) Combining these facts, we get that:

$$\left\|\boldsymbol{X}_{\mathbb{J}}\left[\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{X}_{\mathbb{J}}\right]^{-1}\right\|_{\mathsf{op}} \leqslant \sqrt{2}, \tag{4.28}$$

and it implies by our definitions considering the second event that:

$$\left\|\mathbf{w}_j\right\|_2 \leqslant 2\gamma, \qquad \left\|\mathbf{v}_j\right\|_2 \leqslant \sqrt{2}\gamma. \tag{4.29}$$

We are now ready to apply the results of lemma 4.20. Let us inspect the following probability by the principle of "conditioning on a large probability event":

$$\mathbb{P}\left[\{z_0 \geqslant t\} \cup \{z_1 \geqslant u\}\right] \leqslant \mathbb{P}\left[\{z_0 \geqslant t\} \cup \{z_1 \geqslant u\} \mid E\right] + \mathbb{P}\left[E^{\mathrm{c}}\right] \tag{4.30}$$

$$\leqslant 2p\mathrm{e}^{-\frac{t^2}{8\gamma^2}} + 2p\mathrm{e}^{-\frac{t^2}{4\gamma^2}} + \mathbb{P}\left[E^{\mathrm{c}}\right] \qquad \text{union bound.} \tag{4.31}$$

The latter term can be further upper bounded by a union bound into the sum of the single probabilities:

$$\mathbb{P}\left[z > \frac{1}{2}\right], \qquad \mathbb{P}\left[\max_{j\in\mathbb{J}^{\mathrm{c}}} \left\|\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{x}_j\right\|_{\infty} > \gamma\right]. \tag{4.32}$$

The first term is bounded as in the proof of proposition 4.17 by $p^{-2\log 2}$.
(Applying the result of Tropp (2008)) The latter term is our terminal step. We use the second result of Tropp (2008) we mentioned, namely equation 4.11. Let:

$$t \coloneqq \frac{1}{4}, \qquad u \coloneqq \left(\frac{3}{2} - \sqrt{2}\right)\lambda_p, \qquad \gamma \coloneqq 2^{\frac{1}{q}} 4\mu(\boldsymbol{X})\sqrt{\log p} + 2^{\frac{1}{q}}\sqrt{\frac{s\|\boldsymbol{X}\|_{\mathsf{op}^2}}{p}}, \tag{4.33}$$

which is the right-hand side of equation 4.11 adjusted by the Poissonization of corollary 4.12. Under the conditions of theorem 2.4 $\gamma \leqslant c_0/\sqrt{\log p}$ for some positive c_0, and we can have the bounds:

$$\max\left\{2p\mathrm{e}^{-\frac{t^2}{8\gamma^2}}, 2p\mathrm{e}^{-\frac{u^2}{4\gamma^2}}\right\} \leqslant 2p^{-2\log 2}. \tag{4.34}$$

Moreover, we can also upper bound the remaining term by Markov's as:

$$\mathbb{P}\left[\max_{j\in\mathbb{J}^{\mathrm{c}}} \left\|\boldsymbol{X}_{\mathbb{J}}^{\dagger}\boldsymbol{x}_j\right\|_2 > \gamma\right] \leqslant \frac{1}{\gamma^q}\mathbb{E}\left[\max_{j\in\mathbb{J}^{\mathrm{c}}} \left\|\boldsymbol{X}_{\mathbb{J}}\boldsymbol{x}_j\right\|_2^q\right] \leqslant \left(\frac{\gamma_0}{\gamma}\right)^q. \tag{4.35}$$

By hypothesis $\gamma_0 \leqslant \gamma/2$, so the estimate is at most $p^{-2\log 2}$.
(Finalization) Summarizing, we have shown that:

$$\mathbb{P}\left[\left\{z_0 \geqslant \frac{1}{4}\right\} \cup \left\{z_1 \geqslant \left(\frac{3}{2} - \sqrt{2}\right)\right\}\right] \leqslant 4p^{-2\log 2} + p^{-2\log 2} + p^{-2\log 2} \tag{4.36}$$

So the opposite event holds with probability larger than $1 - 6p^{-2\log 2}$. □

If we combine the probabilities obtained in lemma 4.1, proposition 4.17 and proposition 4.23 we obtain the quantity in theorem 2.4.

Remark 4.37. *One interesting aspect is that in the development we would see that it implies the irrepresentability condition, which is mentioned in (Massart and Meynet 2010) as a standard hypothesis that is not used in their work. It is also one of the aspects of the critique of the classic assumptions on the lasso in (Geer and Bühlmann 2009).*

Candès, Emmanuel J. and Yaniv Plan (Aug. 2009). *Near-Ideal Model Selection by ℓ_1 Minimization*. DOI: 10.1214/08-AOS653. arXiv: 0801.0345 [math].

Geer, Sara A. van de and Peter Bühlmann (Jan. 2009). "On the Conditions Used to Prove Oracle Results for the Lasso". In: *Electronic Journal of Statistics* 3.none. ISSN: 1935-7524. DOI: 10.1214/09-EJS506. arXiv: 0910.0722 [math].

Massart, Pascal (2024). "Model Selection".

Massart, Pascal and Caroline Meynet (July 2010). *An L1-Oracle Inequality for the Lasso*. DOI: 10.48550/arXiv.1007.4791. arXiv: 1007.4791 [math].

Tropp, Joel A. (2008). "Norms of random submatrices and sparse approximation". In: *Comptes Rendus. Mathématique* 346.23-24, pp. 1271–1274. ISSN: 1778-3569. DOI: 10.1016/j.crma.2008.10.008.