Tutorial on Lasso

Statistics Student Seminar @ MSU

Honglang Wang

1 Introduction

1.1 Bias-Variance Trade-off Perspective

Consider a small simulation study with $n = 50$ and $p = 30$. The entries of the predictor matrix $\mathbf{X} \in \mathbb{R}^{50\times 30}$ were all drawn IID from $\mathrm{N}(0, 1)$. The true model is

$$Y_i = \mathbf{X}_i^\intercal \boldsymbol{\beta}^0 + \epsilon_i$$

where ϵ_i IID $\mathrm{N}(0, 1)$ and the true coefficient vector $\boldsymbol{\beta}^0$ has 10 large components (between 0.5 and 1) and 20 small components (between 0 and 0.3). For the linear regression estimator $\hat{\boldsymbol{\beta}}^{\mathrm{ols}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbb{Y}$, we have the expected test error

$$\mathbb{E}\{Y_{\mathrm{new}} - \mathbf{X}_{\mathrm{new}}^\intercal \hat{\boldsymbol{\beta}}^{\mathrm{ols}}\}^2 = \sigma^2 + \sigma^2 p/n,$$

where the first term is the irreducible error, the second term comes entirely from the variance of the linear regression estimate and its bias is exactly zero. Thus if we add another predictor variable into the mix, then it will add the same amount of variance, σ^2/n, regardless of whether its true coefficient is large or small (or zero). So in the example, we were "spending" variance in trying to fit truly small coefficients—there were 20 of them, out of 30 total.

One might think therefore that we can we do better by shrinking small coefficients towards zero, which potentially introduces some bias, but also potentially reduces the variance. In other words, this is trying to ignore some "small details" in order to get a more stable "big picture". If done properly, this will actually work.

1.2 Compressive Sensing Perspective

In many practical problems of science and technology, one encounters the task of inferring quantities of interest from measured information. For instance, in signal and image processing, one would like to reconstruct a signal from measured data. When the information acquisition process is linear, the problem reduces to solving a linear system of equations. In mathematical terms, the observed data $\mathbb{Y} \in \mathbb{R}^n$ is connected to the signal $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ of interest via

$$\mathbb{Y} = \mathbf{X}\boldsymbol{\beta}^0, \tag{1}$$

where the matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ models the linear measurement process. Then one tries to recover the vector $\boldsymbol{\beta}^0$ by solving the above linear system. In compressive sensing setup, the linear system will be under-determined, i.e. $n < p$. That is we have too many (infinitely many) solutions to the system, and we need to pick one out with the property we need. The property we need is just the additional structure assumption for the truth. And people found that surprisingly under the sparsity assumption, it is possible to reconstruct signals in the under-determined situation. And the research area associated with this phenomenon has become known as compressive sensing or sparse recovery.

Looking closer at the standard compressive sensing problem consisting in the reconstruction of a sparse vector $\boldsymbol{\beta}^0$ from under-determined measurements $\mathbb{Y} = \mathbf{X}\boldsymbol{\beta}^0$ one essentially identifies two questions:

1. How should one design the linear measurement process? In other words, what matrix $\mathbf{X}$ are suitable? (Note that this is important since we know for sure that for some design matrix $\mathbf{X}$ the reconstruction is impossible.)

2. How can one reconstruct $\boldsymbol{\beta}^0$ from under-determined measurements $\mathbb{Y} = \mathbf{X}\boldsymbol{\beta}^0$? In other words, what are efficient reconstruction algorithms?

Producing adequate design (measurement/information) matrices $\mathbf{X}$ is a remarkably intriguing endeavor. To data, it is an open problem to construct explicit matrices which are provably optimal in a conpressive sensing setting. A breakthrough is achieved by resorting to random matrices—this discovery can be viewed as the birth of compressive sensing. A key result in compressive sensing states that with high probability on the random draw of an $n \times p$ Gaussian matrix (with IID entries from N(0,1)) $\mathbf{X}$, all s-sparse vector $\boldsymbol{\beta}^0$ can be reconstructed from $\mathbb{Y} = \mathbf{X}\boldsymbol{\beta}^0$ using a variety of algorithms provided

$$n \geq Cs \log(p/s), \tag{2}$$

where $C > 0$ is a universal constant. This bound is in fact optimal.

Foucart and Rauhut (2013) discussed the way to connect the fixed design with the random design described above. First of all, we introduce one popular algorithm is called basis pursuit linear program, given by

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{\beta}\|_1, \ s.t. \ \mathbf{X}\boldsymbol{\beta} = \mathbb{Y}. \tag{3}$$

In this noiseless observation models, the so called restricted nullspace property on the design matrix $\mathbf{X}$ is both necessary and sufficient for the basis pursuit linear program to recover $\boldsymbol{\beta}^0$ exactly.

Definition 1. For a given subset $S \subseteq \{1, 2, \cdots, p\}$ and constant $\alpha \geq 1$, define the set

$$\mathscr{C}(S, \alpha) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S^c}\|_1 \leq \alpha \|\boldsymbol{\beta}_S\|_1\}.$$

For a given sparsity index $s \leq p$, we say that the matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ satisfies the restricted nullspace condition of order s if $\text{null}(\mathbf{X}) \cap \mathscr{C}(S, 1) = \{\mathbf{0}\}$ for all subsets S of cardinality s. Note that $\text{null}(\mathbf{X}) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \mathbf{X}\boldsymbol{\beta} = \mathbf{0}\}$.

And finally people answered that for the Gaussian matrices, the restricted nullspace condition holds with high probability. Thus we got the whole picture behind this beautiful theory:

1. One algorithm: basis pursuit linear programming.
2. One property: restricted null space property.
3. One random matrix theory: for some random matrices, the restricted nullspace condition holds with high probability.

But is statistical and machine learning society, we are in the noise observation world. Consider a linear regression model:

$$\mathbb{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbb{Y} = (Y_1, Y_2, \cdots, Y_n)^\intercal \in \mathbb{R}^n$ is a response vector, $\mathbf{X} = ((X_{ij})) \in \mathbb{R}^{n\times p}$ is a design matrix with columns $\{\mathbf{X}_j\}_{j=1}^p$ and rows $\{\mathbb{X}_i\}_{i=1}^n$, and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is a vector of unknown true regression coefficients. $\boldsymbol{\epsilon}$ independent with $\mathbf{X}$ has mean $\mathbf{0}$, and variance $\sigma^2\mathbf{I}$.

In the case of noisy observation, exact recovery of $\boldsymbol{\beta}^0$ is no longer possible. And we will discuss to control the L_2 error under the so called restricted eigenvalue condition for the design matrix later. And people can also show that for for some random matrices, the restricted eigenvalue condition holds with high probability.

2 Basic Properties of Lasso

For the liner model (4), the lasso solution is defined as

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\lambda) \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1, \lambda > 0. \tag{5}$$

Karush-Kuhn-Tucker (KKT) Conditions By optimization theory, we have that the solution to (5) if and only if the following KKT optimality conditions (i.e. the subgradient of the objective is 0) are satisfied

$$\mathbf{X}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = \lambda \text{sign}(\hat{\boldsymbol{\beta}}) := \lambda\boldsymbol{\gamma}, \tag{6}$$

where $\boldsymbol{\gamma} = \text{sign}(\hat{\boldsymbol{\beta}}) = (\text{sign}(\hat{\beta}_1), \cdots, \text{sign}(\hat{\beta}_p))^\intercal$ with $\gamma_j = \text{sign}(\hat{\beta}_j) \in \begin{cases} \text{sign}(\hat{\beta}_j), & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1], & \text{if } \hat{\beta}_j = 0 \end{cases}$,
for $j \in \{1, 2, \cdots, p\}$.

Let

$$\mathcal{E} := \left\{ j \in \{1, 2, \cdots, p\} : |\mathbb{X}_j^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})|/n = \lambda \right\} = \left\{ j \in \{1, 2, \cdots, p\} : |\gamma_j| = 1 \right\}$$

be the equicorrelation set which contains the variables that have maximal equal absolute correlation with the residual. And let the equicorrelation signs $\mathbf{s}$ as

$$\mathbf{s} := \text{sign}(\mathbb{X}_{\mathcal{E}}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})) = \boldsymbol{\gamma}_{\mathcal{E}}.$$

Now the lasso solution can be written as

$$\hat{\boldsymbol{\beta}}_{\setminus\mathcal{E}} = 0, \tag{7}$$

$$\hat{\boldsymbol{\beta}}_{\mathcal{E}} : \mathbb{X}_{\mathcal{E}}^{\intercal}(\mathbb{Y} - \mathbb{X}_{\mathcal{E}}\hat{\boldsymbol{\beta}}_{\mathcal{E}})/n = \lambda \mathbf{s}. \tag{8}$$

By the following theorem about the uniqueness of the lasso solution Osborne et al. (2000); Tibshirani et al. (2013), we have the nice description of the lasso solution which is also given in the following theorem

Theorem 1. *If the entries of* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *are drawn from a continuous probability distribution on* $\mathbb{R}^{np}$*, then for any* $\mathbb{Y}$ *and* $\lambda > 0$*, with probability one, the lasso solution is unique and is given by*

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\setminus\mathcal{E}} &= 0, \\ \hat{\boldsymbol{\beta}}_{\mathcal{E}} &= (\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{X}_{\mathcal{E}})^{-1}[\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{Y} - n\lambda\mathbf{s}]. \end{aligned} \tag{9}$$

And the solution has at most $n \wedge p$ *nonzero components.*

Remark. From this Proposition, we could see that the lasso solution shrink all of the coefficients towards 0. For those in $\mathcal{E}$, $\hat{\boldsymbol{\beta}}_{\mathcal{E}} = (\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{X}_{\mathcal{E}})^{-1}\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{Y} - n\lambda(\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{X}_{\mathcal{E}})^{-1}\mathbf{s}$ shrinking $(\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{X}_{\mathcal{E}})^{-1}\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{Y}$ by $n\lambda(\mathbb{X}_{\mathcal{E}}^{\intercal}\mathbb{X}_{\mathcal{E}})^{-1}\mathbf{s}$. For those in $\mathcal{E}^c$, they are just shrinked to 0.

3 Algorithm

3.1 Single Linear Regression

With a single predictor (i.e. $p = 1$), $L(\beta) = \|\mathbb{Y} - \mathbb{X}\beta\|_2^2/(2n) + \lambda|\beta|$, the lasso solution is very simple, and is a soft-thresholded version of the least squares estimate $\hat{\beta}^{\text{ols}}$. In fact, by

$$L'(\hat{\beta}) = (\mathbb{X}^{\intercal}\mathbb{X}\hat{\beta} - \mathbb{X}^{\intercal}\mathbb{Y})/n + \lambda \text{sign}(\hat{\beta}) = 0,$$

we know if $\hat{\beta} > 0$, then $(\mathbb{X}^{\intercal}\mathbb{X}\hat{\beta} - \mathbb{X}^{\intercal}\mathbb{Y})/n + \lambda = 0$, i.e. $\hat{\beta} = (\mathbb{X}^{\intercal}\mathbb{X})^{-1}\mathbb{X}^{\intercal}\mathbb{Y} - n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda = \hat{\beta}^{\text{ols}} - n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda$; if $\hat{\beta} < 0$, then $(\mathbb{X}^{\intercal}\mathbb{X}\hat{\beta} - \mathbb{X}^{\intercal}\mathbb{Y})/n - \lambda = 0$, i.e. $\hat{\beta} = (\mathbb{X}^{\intercal}\mathbb{X})^{-1}\mathbb{X}^{\intercal}\mathbb{Y} + n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda = \hat{\beta}^{\text{ols}} + n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda$; and otherwise $\hat{\beta} = 0$. In summary, we have

$$\hat{\beta}(\lambda) = \begin{cases} \hat{\beta}^{\text{ols}} - n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda, & \text{if } \hat{\beta}^{\text{ols}} > 0 \text{ and } \lambda < n^{-1}\mathbb{X}^{\intercal}\mathbb{X}\hat{\beta}^{\text{ols}} = n^{-1}\mathbb{X}^{\intercal}\mathbb{X}|\hat{\beta}^{\text{ols}}| \\ \hat{\beta}^{\text{ols}} + n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda, & \text{if } \hat{\beta}^{\text{ols}} < 0 \text{ and } \lambda < -n^{-1}\mathbb{X}^{\intercal}\mathbb{X}\hat{\beta}^{\text{ols}} = n^{-1}\mathbb{X}^{\intercal}\mathbb{X}|\hat{\beta}^{\text{ols}}| \\ 0, & \text{if } \hat{\beta}^{\text{ols}} = 0 \text{ or } \lambda \geq n^{-1}\mathbb{X}^{\intercal}\mathbb{X}|\hat{\beta}^{\text{ols}}| \end{cases}.$$

And since $\hat{\beta}^{\text{ols}} = 0$ is included in $\lambda \geq n^{-1}\mathbb{X}^{\intercal}\mathbb{X}|\hat{\beta}^{\text{ols}}|$, we finally have

$$\hat{\beta}(\lambda) = \begin{cases} \hat{\beta}^{\text{ols}} - n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda, & \text{if } \hat{\beta}^{\text{ols}} > 0 \text{ and } |\hat{\beta}^{\text{ols}}| > n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda \\ \hat{\beta}^{\text{ols}} + n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda, & \text{if } \hat{\beta}^{\text{ols}} < 0 \text{ and } |\hat{\beta}^{\text{ols}}| > n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda \\ 0, & \text{if } |\hat{\beta}^{\text{ols}}| \leq n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda \end{cases} \tag{10}$$

$$= \text{sign}(\hat{\beta}^{\text{ols}})(|\hat{\beta}^{\text{ols}}| - n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda)_+ := \mathfrak{S}(\hat{\beta}^{\text{ols}}, n(\mathbb{X}^{\intercal}\mathbb{X})^{-1}\lambda), \tag{11}$$

where $\mathfrak{S}(z, \lambda) := \text{sign}(z)(|z| - \lambda)_+$ is the soft threshold function.

3.2 Coordinate Descent Algorithm

The simple optimization theory suggests that for $f(\mathbf{x}) = g(\mathbf{x}) + \sum_{i=1}^n h_i(x_i)$ with g convex, differentiable and each h_i convex, we can use coordinate descent to find a minimizer: start with some initial guess $\mathbf{x}^{(0)}$, and repeat for $k = 1, 2, 3, \cdots$

$$\begin{aligned}
x_1^{(k)} &\in \arg\min_{x_1} f(x_1, x_2^{(k-1)}, x_3^{(k-1)}, \cdots, x_n^{(k-1)}) \\
x_2^{(k)} &\in \arg\min_{x_2} f(x_1^{(k)}, x_2, x_3^{(k-1)}, \cdots, x_n^{(k-1)}) \\
x_3^{(k)} &\in \arg\min_{x_3} f(x_1^{(k)}, x_2^{(k)}, x_3, \cdots, x_n^{(k-1)}) \\
&\vdots \\
x_n^{(k)} &\in \arg\min_{x_n} f(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \cdots, x_n)
\end{aligned}$$

The above "one-at-a-time" updating scheme is critical, and "all-at-once" scheme does not necessarily converge Tseng (2001).

Now for our lasso problem (5), the objective function $\|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1$ have the separable non-smooth part $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. Thus we can use the above coordinate descent algorithm.

And the solution expression we obtained for one single predictor is useful for the general lasso solution since the objective function has the separable non-smooth part. In fact the coordinate wise solution is given by for $j \in \{1, 2, \cdots, p\}$

$$\hat{\beta}_j = \mathfrak{S}((\mathbb{X}_j^\intercal \mathbb{X}_j)^{-1}\mathbb{X}_j(\mathbb{Y} - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\beta}}_{\setminus j}), n(\mathbb{X}^\intercal\mathbb{X})^{-1}\lambda). \tag{12}$$

Friedman et al. (2007) explored the "one-at-a-time" coordinate-wise descent algorithms for some convex optimization problems from statistical analysis, including lasso.

4 Some Consistency Results

4.1 Prediction Consistency and L_1 Consistency

Recall the lasso solution

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1, \lambda > 0.$$

Assume $\boldsymbol{\beta}^0$ is the truth. By the minimization property, we have

$$\|\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \le \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}^0\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}^0\|_1,$$

i.e.

$$\begin{aligned}
\|\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 &= \|\mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&= \|\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2/(2n) + \|\boldsymbol{\epsilon}\|_2^2/(2n) - 2\boldsymbol{\epsilon}^\intercal\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)/(2n) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&\le \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}^0\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}^0\|_1 = \|\boldsymbol{\epsilon}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}^0\|_1,
\end{aligned}$$

which implies the following basic inequality

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \boldsymbol{\epsilon}^\intercal \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)/n + \lambda\|\boldsymbol{\beta}^0\|_1. \tag{13}$$

From (13), we could see that the random part involved there is $\boldsymbol{\epsilon}^\intercal \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)/n$, which can be bounded by

$$|\boldsymbol{\epsilon}^\intercal \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)/n| \leq \max_{1\leq j\leq p}\{|\boldsymbol{\epsilon}^\intercal \mathbb{X}_j|/n\}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1. \tag{14}$$

Now let's introduce an event (which could be proved to have high probability for this event to hold)

$$\mathscr{T} := \Big\{ \max_{1\leq j\leq p}\{|\boldsymbol{\epsilon}^\intercal \mathbb{X}_j|/n \leq \lambda_0\Big\}. \tag{15}$$

By working on this event, we can get rid of randomness.

On $\mathscr{T}$, we have $|\boldsymbol{\epsilon}^\intercal \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)/n| \leq \lambda_0\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1$, and hence

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda_0\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 + \lambda\|\boldsymbol{\beta}^0\|_1.$$

Let $S := \{j : \beta_j^0 \neq 0\}$ be the non-zero position set for the truth. By $\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \geq \|\boldsymbol{\beta}_S^0\|_1 - \|\hat{\boldsymbol{\beta}}_S\|_1$, we have

$$\|\hat{\boldsymbol{\beta}}\|_1 = \|\hat{\boldsymbol{\beta}}_S\|_1 + \|\hat{\boldsymbol{\beta}}_{S^c}\|_1 \geq \|\boldsymbol{\beta}_S^0\|_1 - \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + \|\hat{\boldsymbol{\beta}}_{S^c}\|_1.$$

And then

$$\begin{aligned}
\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) &\leq \lambda_0\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 + \lambda\|\boldsymbol{\beta}^0\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
&\leq \lambda_0\{\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + \|\hat{\boldsymbol{\beta}}_{S^c}\|_1\} + \lambda\{\|\boldsymbol{\beta}_S^0\|_1 - \|\boldsymbol{\beta}_S^0\|_1 + \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 - \|\hat{\boldsymbol{\beta}}_{S^c}\|_1\} \\
&= (\lambda_0 + \lambda)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + (\lambda_0 - \lambda)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1.
\end{aligned}$$

Then if $\lambda_0 \leq \lambda/2$, we have

$$\begin{aligned}
\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) &\leq (\lambda_0 + \lambda)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + (\lambda_0 - \lambda)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 \\
&\leq (3\lambda/2)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 - (\lambda/2)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1,
\end{aligned}$$

that is

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + (\lambda/2)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 \leq (3\lambda/2)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1.$$

Note that this inequality is true on the event $\mathscr{T}$ with $\lambda_0 \leq \lambda/2$, and it also implies that

$$\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 \leq 3\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1,$$

i.e.

$$\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)_{S^c}\|_1 = \|\hat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}_{S^c}^0\|_1 \leq 3\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 = 3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)_S\|_1,$$

that is the error vector $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ belongs to a very specific region. Such nice property is actually from the two facts: the first one is the decomposability of the norm based regularizer $\mathcal{R}(\cdot) = \|\cdot\|_1$, and the other is the choice of the regularization penalty λ.

We now define the following version of compatibility condition:

Definition 2. (Compatibility Condition) We say that the compatibility condition is met for the set S, if for some $\phi_0 > 0$, and for all $\boldsymbol{\beta} \in \mathscr{C}(S,3) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S^c}\|_1 \leq 3\|\boldsymbol{\beta}_S\|_1\}$, it holds that

$$\|\boldsymbol{\beta}_S\|_1^2 \leq s(\boldsymbol{\beta}^\intercal \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta})/\phi_0^2,$$

where $\hat{\boldsymbol{\Sigma}} = \mathbf{X}^\intercal \mathbf{X}/n$. And we denote $\phi^2(\hat{\boldsymbol{\Sigma}}, S) := \min_{\boldsymbol{\beta} \in \mathscr{C}(S,3)} \frac{s\boldsymbol{\beta}^\intercal \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}}{\|\boldsymbol{\beta}_S\|_1^2}$.

Since we have already obtained that on the event $\mathscr{T}$ with $\lambda_0 \leq \lambda/2$,

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + (\lambda/2)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 \leq (3\lambda/2)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1,$$

we have

$$\begin{aligned}
\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + (\lambda/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 &= \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + (\lambda/2)\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 + (\lambda/2)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \\
&\leq (3\lambda/2)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 + (\lambda/2)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \\
&= 2\lambda\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1.
\end{aligned}$$

Since we know that on the event $\mathscr{T}$ with $\lambda_0 \leq \lambda/2$, the error vector $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0$ satisfies $\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)_{S^c}\|_1 \leq 3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)_S\|_1$, by compatibility condition, we have

$$\begin{aligned}
\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + (\lambda/2)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 &\leq 2\lambda\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^0\|_1 \leq 2\lambda\sqrt{s([\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0]^\intercal \hat{\boldsymbol{\Sigma}}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0])/\phi_0^2} \\
&= 2\lambda\sqrt{s}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2/(\sqrt{n}\phi_0) \\
&\leq \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(4n) + 4\lambda^2 s/\phi_0^2,
\end{aligned}$$

which implies that

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq 8\lambda^2 s/\phi_0^2.$$

We summarize the above analysis as the following theorem.

Theorem 2. *Suppose the compatibility holds for S. Then on the event $\mathscr{T}$ with $\lambda_0 \leq \lambda/2$, we have*

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/(2n) + \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq 8\lambda^2 s/\phi_0^2. \tag{16}$$

The only thing left is when the event $\mathscr{T}$ with $\lambda_0 \leq \lambda/2$ holds? This is answered by the following lemma.

Theorem 3. *Suppose that $\hat{\sigma}_j = \hat{\boldsymbol{\Sigma}}_{j,j} = 1$ where $\hat{\boldsymbol{\Sigma}} = \mathbf{X}^\intercal \mathbf{X}/n$ for all j. Then we have for all $t > 0$, and for $\lambda_0 = \sigma\sqrt{\frac{t^2 + 2\log p}{n}}$,*

$$\mathbb{P}(\mathscr{T}) \geq 1 - 2\exp\{-t^2/2\}.$$

Proof. Recall that $\mathscr{T} = \Big\{ \max_{1 \leq j \leq p}\{|\boldsymbol{\epsilon}^\intercal \mathbb{X}_j|/n \leq \lambda_0 \Big\}$. We have

$$\begin{aligned}
1 - \mathbb{P}(\mathscr{T}) &= \mathbb{P}\Big\{ \max_{1 \leq j \leq p} |\boldsymbol{\epsilon}^\intercal \mathbb{X}_j|/n > \lambda_0 \Big\} = \mathbb{P}\Big\{ \max_{1 \leq j \leq p} |\boldsymbol{\epsilon}^\intercal \mathbb{X}_j|/\sqrt{n\sigma^2} > \sqrt{t^2 + 2\log p} \Big\} \\
&\leq p\mathbb{P}\Big\{ |\boldsymbol{\epsilon}^\intercal \mathbb{X}_j|/\sqrt{n\sigma^2} > \sqrt{t^2 + 2\log p} \Big\} = 2p\mathbb{P}\Big\{ \boldsymbol{\epsilon}^\intercal \mathbb{X}_j/\sqrt{n\sigma^2} > \sqrt{t^2 + 2\log p} \Big\} \\
&\leq 2p\exp\{-\frac{t^2 + 2\log p}{2}\} = 2\exp\{-t^2/2\}.
\end{aligned}$$

□

4.2 L_2 Consistency

First, we introduce a stronger condition than compatibility condition, the restricted eigenvalue conditions.

Definition 3. We say that the $p \times p$ sample covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{X}^\intercal \mathbf{X}/n$ satisfies the restricted eigenvalue condition over S with parameters $\phi_0 \in (0, \infty)$ if

$$\boldsymbol{\beta}^\intercal \mathbf{\Sigma} \boldsymbol{\beta} = \|\mathbf{X}\boldsymbol{\beta}\|_2^2/n \geq \phi_0^2 \|\boldsymbol{\beta}\|_2^2, \forall \boldsymbol{\beta} \in \mathscr{C}(S, 3),$$

where $\mathscr{C}(S, 3)$ as defined in compatibility condition. If this condition holds uniformly for all subsets S with cardinality s, we say that $\hat{\mathbf{\Sigma}}$ satisfies a restricted eigenvalue condition of order s with parameters ϕ_0. On occasion, we will say also that a deterministic $p \times p$ covariance matrix $\mathbf{\Sigma}$ satisfies a restricted eigenvalue condition, by which we mean that $\|\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 \geq \phi_0 \|\boldsymbol{\beta}\|_2$ for all $\boldsymbol{\beta} \in \mathscr{C}(S, 3)$.

Now the question is in the setting of linear regression with random design, for what ensembles of design matrices do the restricted eigenvalue condition hold with high probability. Raskutti et al. (2010) discussed the correlated Gaussian designs since in reality it's not reasonable to assume that different covariates are IID (except in compressive sensing setting).

Theorem 4. *For any Gaussian random design $\mathbf{X} \in \mathbb{R}^{n \times p}$ with IID $N(\mathbf{0}, \mathbf{\Sigma})$ rows, there are universal positive constants C_1, C_2 such that*

$$\|\mathbf{X}\boldsymbol{\beta}\|_2/\sqrt{n} \geq \|\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}\|_2/4 - 9\sqrt{\max_{j \in \{1,2,\cdots,p\}} \mathbf{\Sigma}_{j,j}}\sqrt{\log p/n}\|\boldsymbol{\beta}\|_1, \forall \boldsymbol{\beta} \in \mathbb{R}^p \tag{17}$$

with probability at least $1 - C_1 \exp(-C_2 n)$.

Remark. For this probability inequality, we are not going to prove it here (Raskutti et al. (2010)). And the constants 1/4 and 9 are not meant to be sharp.

Theorem 5. *Suppose that $\mathbf{\Sigma}$ satisfies the restricted eigenvalue condition of order s with parameter ϕ_0. Then for universal positive constants C_1, C_2, C_3, if the sample size satisfies*

$$n > C_3 \frac{16 \max_{j \in \{1,2,\cdots,p\}} \mathbf{\Sigma}_{j,j}}{\phi_0^2} s \log p,$$

then the matrix $\hat{\mathbf{\Sigma}} = \mathbf{X}^\intercal \mathbf{X}/n$ satisfies the restricted eigenvalue condition with parameters $\phi_0/8$ with probability at least $1 - C_1 \exp(-C_2 n)$.

Proof. Let S be an arbitrary subset of cardinality s, and suppose $\boldsymbol{\beta} \in \mathscr{C}(S, 3)$. By definition and Cauchy-Schwarts inequality, we have

$$\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}_S\|_1 + \|\boldsymbol{\beta}_{S^c}\|_1 \leq 4\|\boldsymbol{\beta}_S\|_1 \leq 4\sqrt{s}\|\boldsymbol{\beta}_S\|_2 \leq 4\sqrt{s}\|\boldsymbol{\beta}\|_2.$$

By assumption that $\mathbf{\Sigma}$ satisfies the restricted eigenvalue condition of order s with parameter ϕ_0, we have

$$\|\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}\|_2 \geq \phi_0 \|\boldsymbol{\beta}\|_2, \forall \boldsymbol{\beta} \in \mathscr{C}(S, 3).$$

By substituting these two into the bound in (17), we have

$$\begin{aligned}\|\mathbf{X}\boldsymbol{\beta}\|_2/\sqrt{n} &\geq \phi_0\|\boldsymbol{\beta}\|_2/4 - 36\sqrt{\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}\sqrt{s\log p/n}\|\boldsymbol{\beta}\|_2 \\ &= \Big\{\phi_0/4 - 36\sqrt{\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}\sqrt{s\log p/n}\Big\}\|\boldsymbol{\beta}\|_2.\end{aligned}$$

Now by $n > C_3\frac{16\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}{\phi_0^2}s\log p$, we have

$$\begin{aligned}&36\sqrt{\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}\sqrt{s\log p/n} \\ \leq&36\sqrt{\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}\sqrt{s\log p/\Big(C_3\frac{16\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}{\phi_0^2}s\log p\Big)} \\ =&9\phi_0/\sqrt{C_3}.\end{aligned}$$

Thus,

$$\|\mathbf{X}\boldsymbol{\beta}\|_2/\sqrt{n} \geq \{1/4 - 9/\sqrt{C_3}\}\phi_0\|\boldsymbol{\beta}\|_2.$$

By taking $C_3 = 72^2$, we have $1/4 - 9/\sqrt{C_3} = 1/8$ and we finished the proof. □

(A2) The rows of $\mathbf{X}$ are IID realizations from a Gaussian distribution whose $p-$dimensional covariance matrix $\boldsymbol{\Sigma}$ has strictly positive smallest eigenvalue $\Lambda^2_{\min}$ satisfying $1/\Lambda^2_{\min} = O(1)$. Furthermore, $\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j} = O(1)$.

We now ready to get the general L_2 bounds with random design by the following theorem.

Theorem 6. *Under (A2), $s = o(n/\log p)$ and $\lambda \asymp \sqrt{\log p/n}$, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 = O_p(\sqrt{s\log p/n}).$$

Proof. In fact, by Theorem 4, there are universal positive constants C_1, C_2 such that

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2/\sqrt{n} \geq \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2/4 - 9\sqrt{\max_{j\in\{1,2,\cdots,p\}}\boldsymbol{\Sigma}_{j,j}}\sqrt{\log p/n}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_1 \quad (18)$$

with probability at least $1 - C_1\exp(-C_2 n)$. That is with probability tending to one, for a suitably chosen C, we have

$$\|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2 \leq C\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2/\sqrt{n} + C\sqrt{\log p/n}\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_1.$$

By (A2) and $s = o(n/\log p)$, by the first fact, with probability tending to one we have the compatibility condition holds. So we can appeal to the Theorem 2 together with the Lemma 3, and then we have

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2/n = O_p(s\lambda^2), \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 = O_p(s\lambda).$$

So together with $\sqrt{s}\lambda = o(1)$, we have

$$\|\mathbf{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2 \leq C\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2/\sqrt{n}+C\sqrt{\log p/n}\|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_1 = O_p(\sqrt{s}\lambda)+O_p((\sqrt{s}\lambda)^2) = O_p(\sqrt{s}\lambda).$$

Thus by (A2) with $\frac{\|\mathbf{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_2^2} \geq \Lambda_{\min}^2$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 = O_p(\sqrt{s\log p/n}).$$

□

5 Statistical Inference

I found an interesting post from the blog "Empirical Filtration", talking about the difference between statistics and statistical engineering. It is common that many people in statistical engineering try to find the bounds on convergence rate, which we have done so far for lots of consistency results. The bounds are like their destination; they usually not go further for the distribution. In contrast, people in statistics will not stop at the rate; statisticians are targeting at the asymptotic distributions.

The reason why statisticians care about asymptotic distribution may be related to the statistical inference. The statistical inference such as confidence intervals, hypothesis tests, requires knowledge about the distribution of a certain statistics. Knowing the bounds is not sufficient for carrying out the inference. Both confidence intervals (or more general, confidence sets) and hypothesis test require the distributions.

This might also be the reason why courses in ML emphasizes more on the Hoeffding's inequality, Bernstein's inequality while in statistics, the courses focus more on the central limit theory and chi-square approximation.

Although we statistician should not limit ourselves to those methods that are capable of statistical inference, since many methods though have no asymptotic distribution, are still very useful in prediction, especially those with guarantees from probability bounds, we need to consider statistical inference now for our lasso solution.

5.1 Knight & Fu

We discuss the paper Knight and Fu (2000), which derived the asymptotic distribution for lasso type estimators in low dimension case.

Let's define the random function

$$\mathcal{L}_n(\boldsymbol{\beta}) := \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_n\|\boldsymbol{\beta}\|_1, \lambda_n > 0, \tag{19}$$

which is minimized at $\hat{\boldsymbol{\beta}}$. We assume that the dimension p and the truth $\boldsymbol{\beta}^0$ are fixed, independent of n. We also assume the following two regularity conditions

$$\hat{\mathbf{\Sigma}} \to \mathbf{\Sigma}, n \to \infty, \tag{20}$$

$$\frac{1}{n}\max_{1\leq i\leq n}\mathbf{X}_i^\intercal\mathbf{X}_i \to 0, n \to \infty, \tag{21}$$

which makes the asymptotic normality of the ordinary least square estimator

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{\text{ols}} - \boldsymbol{\beta}^0) \xrightarrow{d} \mathrm{N}(0, \sigma^2\boldsymbol{\Sigma}^{-1}).$$

For simplicity, we also assume $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ are non singular.

First note that we want to study the asymptotics of the minimizer of a convex random process. And actually the asymptotics of convex optimization has been studied by several authors (see Kato (2009)). They developed the convexity arguments: let $g_n(\mathbf{x})$ and $g_\infty(\mathbf{x})$ be random convex functions taking minimum values at $\mathbf{x}_n$ and $\mathbf{x}_\infty$ respectively. If all finite dimensional distributions of g_n converge weakly to those of g_∞ and $\mathbf{x}_\infty$ is the unique minimum point of g_∞ with probability one, then $\mathbf{x}_n$ converges weakly to $\mathbf{x}_\infty$.

Theorem 7. *Given the above regularity conditions (20), (21) and assuming $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ nonsingular, with $\lambda_n \to \lambda_0 \geq 0$, we have $\hat{\boldsymbol{\beta}} \xrightarrow{p} \arg\min \mathcal{L}_0(\boldsymbol{\beta})$ where $\mathcal{L}_0(\boldsymbol{\beta}) = (\boldsymbol{\beta}-\boldsymbol{\beta}^0)^\intercal\boldsymbol{\Sigma}(\boldsymbol{\beta}-\boldsymbol{\beta}^0) + \lambda_0\|\boldsymbol{\beta}\|_1$. Thus if $\lambda_n \to 0$, $\arg\min \mathcal{L}_0 = \boldsymbol{\beta}^0$, and so $\hat{\boldsymbol{\beta}}$ is consistent.*

Proof. This just follows from the above convexity arguments since $\mathcal{L}_n(\boldsymbol{\beta}) \xrightarrow{a.s.} \mathcal{L}_0(\boldsymbol{\beta}) + \sigma^2$ pointwisely and both are convex. □

Theorem 8. *Given the above regularity conditions (20), (21) and assuming $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ nonsingular, with $\sqrt{n}\lambda_n \to \lambda_0 \geq 0$, then we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} \arg\min \mathcal{V}(\boldsymbol{\beta}),$$

where $\mathcal{V}(\boldsymbol{\beta}) = -2\boldsymbol{\beta}^\intercal\mathbf{u} + \boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta} + 2\lambda_0\sum_{j=1}^p\{\beta_j sign(\beta_j^0)\mathbb{1}(\beta_j^0 \neq 0) + |\beta_j|\mathbb{1}(\beta_j^0 = 0)\}$ and $\mathbf{u}$ has a $N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$ distribution.

Proof. First observe that

$$\begin{aligned}
\mathcal{L}_n(\boldsymbol{\beta}) &= \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_n\|\boldsymbol{\beta}\|_1 \\
&= \|\mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda_n\|\boldsymbol{\beta}\|_1 \\
&= \|\boldsymbol{\epsilon} - \mathbf{X}\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)/\sqrt{n}\|_2^2/(2n) + \lambda_n\|\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)/\sqrt{n} + \boldsymbol{\beta}^0\|_1.
\end{aligned}$$

Now define

$$\mathcal{V}_n(\boldsymbol{\alpha}) = \|\boldsymbol{\epsilon} - \mathbf{X}\boldsymbol{\alpha}/\sqrt{n}\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2 + 2n\lambda_n\|\boldsymbol{\alpha}/\sqrt{n} + \boldsymbol{\beta}^0\|_1 - 2n\lambda_n\|\boldsymbol{\beta}^0\|. \tag{22}$$

Then we have $\hat{\boldsymbol{\alpha}} := \arg\min \mathcal{V}_n(\boldsymbol{\alpha}) = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. And since $\|\boldsymbol{\epsilon} - \mathbf{X}\boldsymbol{\alpha}/\sqrt{n}\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2 = -2\boldsymbol{\epsilon}^\intercal\mathbf{X}\boldsymbol{\alpha}/\sqrt{n} + \boldsymbol{\alpha}^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\alpha}/n \xrightarrow{d} -2\boldsymbol{\alpha}^\intercal\mathbf{u} + \boldsymbol{\alpha}^\intercal\boldsymbol{\Sigma}\boldsymbol{\alpha}$. And it's easy to check out that

$$2n\lambda_n\|\boldsymbol{\alpha}/\sqrt{n} + \boldsymbol{\beta}^0\|_1 - 2n\lambda_n\|\boldsymbol{\beta}^0\| \to 2\lambda_0\sum_{j=1}^p\{\alpha_j\text{sign}(\beta_j^0)\mathbb{1}(\beta_j^0 \neq 0) + |\alpha_j|\mathbb{1}(\beta_j^0 = 0)\}.$$

Thus we have proved that $\mathcal{V}_n(\boldsymbol{\alpha}) \xrightarrow{d} \mathcal{V}(\boldsymbol{\alpha})$ with the finite-dimensional convergence. And since $\mathcal{V}_n$ is convex and $\mathcal{V}$ has a unique minimum, by the convexity argument, we have

$$\arg\min \mathcal{V}_n = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} \arg\min \mathcal{V}.$$

Note that when $\lambda_0 = 0$, $\arg\min \mathcal{V} = \boldsymbol{\Sigma}^{-1}\mathbf{u} \sim \mathrm{N}(0, \sigma^2\boldsymbol{\Sigma}^{-1})$. □

Remark. 1. Theorem 8 shows that nonzero parameters are estimated with some asymptotic bias if $\lambda_0 > 0$ due the the term $2\lambda_0 \sum_{j=1}^{p} \{\beta_j \text{sign}(\beta_j^0)\mathbb{1}(\beta_j^0 \neq 0) + |\beta_j|\mathbb{1}(\beta_j^0 = 0)\}$ in $\mathcal{V}(\boldsymbol{\beta})$, and shrinking the estimates of zero regression parameters to 0 with positive probability: Suppose $\boldsymbol{\beta}_{S^c}^0 = \mathbf{0}$ and $\boldsymbol{\beta}_S^0$ are nonzero. In this case, we have

$$\mathcal{V}(\boldsymbol{\alpha}) = -2\boldsymbol{\alpha}^\intercal \mathbf{u} + \boldsymbol{\alpha}^\intercal \boldsymbol{\Sigma} \boldsymbol{\alpha} + 2\lambda_0 \sum_{j \in S} \alpha_j \text{sign}(\beta_j^0) + 2\lambda_0 \sum_{j \in S^c} |\alpha_j|.$$

And by the uniqueness of the solution to $\min \mathcal{V}(\boldsymbol{\alpha})$, we have $\boldsymbol{\alpha}_{S^c} = \mathbf{0}$ if and only if $|\boldsymbol{\Sigma}_{2,1}(S)(\boldsymbol{\Sigma}_{1,1}(S))^{-1}[\mathbf{u}_S - \lambda_0 \text{sign}(\boldsymbol{\beta}_S^0)] - \mathbf{u}_{S^c}| \leq \lambda_0$ (element wise) since KKT holds by taking $\boldsymbol{\alpha}_S = (\boldsymbol{\Sigma}_{1,1}(S))^{-1}[\mathbf{u}_S - \lambda_0 \text{sign}(\boldsymbol{\beta}_S^0)]$. Thus we have $\mathbb{P}(\boldsymbol{\alpha}_{S^c} = \mathbf{0}) = \mathbb{P}\big\{|\boldsymbol{\Sigma}_{2,1}(S)(\boldsymbol{\Sigma}_{1,1}(S))^{-1}[\mathbf{u}_S - \lambda_0 \text{sign}(\boldsymbol{\beta}_S^0)] - \mathbf{u}_{S^c}| \leq \lambda_0\big\} > 0$.

2. Except for some very special cases, no closed form formula for either the limiting random vector or the limiting distribution is available. As a result, the use of the asymptotic distribution of the lasso estimator for constructing confidence intervals or for conducting large sample tests is not very appealing in practice.—Bootstrapping! Now we refer to the following two papers, Chatterjee and Lahiri (2010, 2011) for the discussing of bootstrapping for lasso estimator.

5.2 van de Geer & Bühlmann & Ritov & Dezeure

We are going to discuss essentially the same estimator as in Zhang and Zhang (2014) from van de Geer et al. (2014).

Consider a linear regression model:

$$\mathbb{Y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}, \tag{23}$$

where $\mathbb{Y} = (Y_1, Y_2, \cdots, Y_n)^\intercal \in \mathbb{R}^n$ is a response vector, $\mathbf{X} = ((X_{ij})) \in \mathbb{R}^{n \times p}$ is a design matrix with columns $\{\mathbf{X}_j\}_{j=1}^p$ and rows $\{\mathbb{X}_i\}_{i=1}^n$, and $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is a vector of unknown true regression coefficients. $\boldsymbol{\epsilon}$ independent with $\mathbf{X}$ has normal mean $\mathbf{0}$, and variance $\sigma^2\mathbf{I}$, i.e. $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Recall the lasso solution

$$\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbb{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1, \lambda > 0,$$

and the Karush-Kuhn-Tucker (KKT) Condition

$$\mathbf{X}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = \lambda \text{sign}(\hat{\boldsymbol{\beta}}) := \lambda\hat{\boldsymbol{\kappa}},$$

where $\hat{\boldsymbol{\kappa}} = \text{sign}(\hat{\boldsymbol{\beta}}) = (\text{sign}(\hat{\beta}_1), \cdots, \text{sign}(\hat{\beta}_p))^\intercal$ with $\kappa_j = \text{sign}(\hat{\beta}_j) \in \begin{cases} \text{sign}(\hat{\beta}_j), & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1], & \text{if } \hat{\beta}_j = 0 \end{cases}$, for $j \in \{1, 2, \cdots, p\}$.

We re-write KKT as

$$\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \lambda\hat{\boldsymbol{\kappa}} = \mathbf{X}^\intercal\boldsymbol{\epsilon}/n.$$

The idea is now to use a "relaxed form" of an inverse of $\hat{\mathbf{\Sigma}}$ (which is similar to the idea of the "relaxed projection" in Zhang and Zhang (2014)). Suppose that $\hat{\mathbf{\Theta}}$ is a reasonable approximation for such an inverse, then

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + (\hat{\mathbf{\Theta}}\mathbf{\Sigma} - \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \hat{\mathbf{\Theta}}\lambda\hat{\boldsymbol{\kappa}} = \hat{\mathbf{\Theta}}\mathbf{X}^\intercal\boldsymbol{\epsilon}/n,$$

i.e.

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \hat{\mathbf{\Theta}}\lambda\hat{\boldsymbol{\kappa}} = \hat{\mathbf{\Theta}}\mathbf{X}^\intercal\boldsymbol{\epsilon}/n - \mathbf{\Delta}/\sqrt{n}, \tag{24}$$

where $\mathbf{\Delta} = \sqrt{n}(\hat{\mathbf{\Theta}}\mathbf{\Sigma} - \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. Thus suggests the following estimator

$$\hat{\mathbf{b}} := \hat{\boldsymbol{\beta}} + \hat{\mathbf{\Theta}}\lambda\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\beta}} + \hat{\mathbf{\Theta}}\mathbf{X}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n. \tag{25}$$

Look at the estimator closely, you will see the following aspect of interpretation. Based on the construction of the Lasso estimator, we observed that the Lasso solution is biased towards smaller L_1 norm. And $\mathbf{X}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n\lambda)$ is a subgradient of the L_1 norm at the lasso solution $\hat{\boldsymbol{\beta}}$. By adding a term which is a linear transformation (rotation and scaling) of this subgradient, our procedure compensates the bias introduced by the L_1 penalty in the Lasso (Javanmard and Montanari, 2013).

Now we are going to construct the approximate inverse $\hat{\mathbf{\Theta}}$ by the lasso for the nodewise regression on the design $\mathbf{X}$. For each $j \in \{1, 2, \cdots, p\}$, recall

$$\hat{\boldsymbol{\gamma}}_j := \hat{\boldsymbol{\gamma}}_j(\lambda_j) = \arg\min_{\mathbf{b}\in\mathbb{R}^{p-1}} \left\{\|\mathbb{X}_j - \mathbb{X}_{\setminus j}\mathbf{b}\|_2^2/(2n) + \lambda_j\|\mathbf{b}\|_1\right\}, \tag{26}$$

$$\mathbb{Z}_j := \mathbb{Z}_j(\lambda_j) = \mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\gamma}}_j(\lambda_j), \tag{27}$$

where $\hat{\boldsymbol{\gamma}}_j = \{\hat{\gamma}_{j,k} : k = 1, 2, \cdots, p, k \neq j\}$. By denoting

$$\hat{\mathbf{\Gamma}} := \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & \hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}, \text{ and } \hat{\mathbf{\Xi}}^2 := \text{diag}(\hat{\xi}_1^2, \cdots, \hat{\xi}_p^2)$$

where $\hat{\xi}_j^2 = \|\mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\gamma}}_j\|_2^2/n + \lambda_j\|\hat{\boldsymbol{\gamma}}_j\|_1 = \mathbb{X}_j^\intercal(\mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\gamma}}_j)/n - \hat{\boldsymbol{\gamma}}_j^\intercal\mathbb{X}_{\setminus j}^\intercal(\mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\gamma}}_j)/n + \lambda_j\|\hat{\boldsymbol{\gamma}}_j\|_1 = \mathbb{X}_j^\intercal\mathbb{Z}_j/n - \hat{\boldsymbol{\gamma}}_j^\intercal\lambda_j\text{sign}(\hat{\boldsymbol{\gamma}}_j) + \lambda_j\|\hat{\boldsymbol{\gamma}}_j\|_1 = \mathbb{X}_j^\intercal\mathbb{Z}_j/n$ by the KKT condition for the $j-$th node regression, we define

$$\hat{\mathbf{\Theta}} = \begin{pmatrix} \hat{\mathbf{\Theta}}_1^\intercal \\ \hat{\mathbf{\Theta}}_2^\intercal \\ \vdots \\ \hat{\mathbf{\Theta}}_p^\intercal \end{pmatrix} := \hat{\mathbf{\Xi}}^{-2}\hat{\mathbf{\Gamma}} = \begin{pmatrix} \hat{\mathbf{\Gamma}}_1^\intercal/\hat{\xi}_1^2 \\ \hat{\mathbf{\Gamma}}_2^\intercal/\hat{\xi}_2^2 \\ \vdots \\ \hat{\mathbf{\Gamma}}_p^\intercal/\hat{\xi}_p^2 \end{pmatrix}. \tag{28}$$

Then we can further have our estimator

$$\begin{aligned} \hat{b}_j &= \hat{\beta}_j + \hat{\mathbf{\Theta}}_j^\intercal\mathbf{X}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = \hat{\beta}_j + \hat{\xi}_j^{-2}\hat{\mathbf{\Gamma}}_j^\intercal\mathbf{X}^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n \\ &= \hat{\beta}_j + \hat{\xi}_j^{-2}(\mathbf{X}\hat{\mathbf{\Gamma}}_j)^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = \hat{\beta}_j + \hat{\xi}_j^{-2}(\mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\gamma}}_j)^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n \\ &= \hat{\beta}_j + \hat{\xi}_j^{-2}\mathbb{Z}_j^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = \hat{\beta}_j + (\mathbb{X}_j^\intercal\mathbb{Z}_j/n)^{-1}\mathbb{Z}_j^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n \\ &= \hat{\beta}_j + (\mathbb{X}_j^\intercal\mathbb{Z}_j)^{-1}\mathbb{Z}_j^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\beta}_j + \frac{\mathbb{Z}_j^\intercal(\mathbb{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\mathbb{X}_j^\intercal\mathbb{Z}_j}, \end{aligned} \tag{29}$$

which is exactly the same as the LDPE obtained in Zhang and Zhang (2014). From here, we could see another aspect of this estimator. Taking $j = 1$ as an example, we first regress $\mathbb{X}_1$ against $\mathbb{X}_{\setminus 1}$, and also regress $\mathbb{Y}$ against $\mathbb{X}_{\setminus 1}$, we got the linear regression estimator

$$\hat{\beta}_1^{(\text{lin})} = \frac{(\mathcal{Q}_{\setminus 1}\mathbb{X}_1)^\intercal \mathcal{Q}_{\setminus 1}\mathbb{Y}}{(\mathcal{Q}_{\setminus 1}\mathbb{X}_1)^\intercal \mathcal{Q}_{\setminus 1}\mathbb{X}_1} = \frac{(\mathcal{Q}_{\setminus 1}\mathbb{X}_1)^\intercal \mathbb{Y}}{(\mathcal{Q}_{\setminus 1}\mathbb{X}_1)^\intercal \mathbb{X}_1} = \frac{(\mathbb{X}_1^\perp)^\intercal \mathbb{Y}}{(\mathbb{X}_1^\perp)^\intercal \mathbb{X}_1} = \frac{\mathbb{Z}_1^\intercal \mathbb{Y}}{\mathbb{Z}_1^\intercal \mathbb{X}_1},$$

where the projection relaxed by using lasso, i.e. $\mathbb{X}_1^\perp = \mathbb{Z}_1$ due to high dimensionality. But instead of using $\hat{\beta}_1^{(\text{lin})}$ directly, we used $\hat{b}_1 = \frac{\mathbb{Z}_1^\intercal \mathbb{Y}}{\mathbb{Z}_1^\intercal \mathbb{X}_1} - \sum_{k \neq 1} \frac{\mathbb{Z}_1^\intercal \mathbb{X}_k \hat{\beta}_k^{(\text{init})}}{\mathbb{Z}_1^\intercal \mathbb{X}_1}$ since we need to do the bias correction. Note that the reason why we need to do the relaxed projection is that the corresponding regression is still high dimensional. Thus in general we need to have a projection onto a low dimensional covariates, a small subset $\mathcal{S}$ of the covariates which are more strongly correlated with target:

$$\hat{\beta}_1^{(\text{lin})} = \frac{(\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^\intercal (\mathcal{Q}_{\mathcal{S}}\mathbb{Y})}{(\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^\intercal \mathcal{Q}_{\mathcal{S}}\mathbb{X}_1}.$$

And actually since for the relaxed projection, we used lasso procedure, the projection is just onto the selected variables by lasso. That is we used nodewise lasso to select the low dimensional covariates to project on.

Note that from the KKT conditions for the nodewise lasso, we have

$$\begin{aligned}
\mathbb{X}_j^\intercal \mathbf{X}\hat{\mathbf{\Theta}}_j/n = \mathbb{X}_j^\intercal \hat{\xi}_j^{-2}(\mathbf{X}\hat{\mathbf{\Gamma}}_j)/n = \hat{\xi}_j^{-2}\mathbb{X}_j^\intercal \mathbb{Z}_j/n = \hat{\xi}_j^{-2}\hat{\xi}_j^2 = 1, \\
\|\mathbb{X}_{\setminus j}^\intercal \mathbf{X}\hat{\mathbf{\Theta}}_j/n\|_\infty = \|\mathbb{X}_{\setminus j}^\intercal \hat{\xi}_j^{-2}(\mathbf{X}\hat{\mathbf{\Gamma}}_j)/n\|_\infty = \|\hat{\xi}_j^{-2}\mathbb{X}_{\setminus j}^\intercal (\mathbb{X}_j - \mathbb{X}_{\setminus j}\hat{\boldsymbol{\gamma}}_j)/n\|_\infty \leq \lambda_j \hat{\xi}_j^{-2},
\end{aligned}$$

which lead to the following so called extended KKT condition

$$\|\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}_j - \mathbf{e}_j\|_\infty \leq \lambda_j \hat{\xi}_j^{-2}, \tag{30}$$

where $\mathbf{e}_j$ is the j-th unit column vector.

Recall the compatibility condition in Definition (2) defined in the prediction consistency section, and we make the following assumption:

(A1) The compatibility condition (2) holds for $\hat{\mathbf{\Sigma}}$ with compatibility constant ϕ_0^2. Furthermore, $\max_j \hat{\mathbf{\Sigma}}_{j,j} \leq M^2$ for some $0 < M < \infty$ (Note that sometimes for simplicity we assume it's normalized such that $\hat{\mathbf{\Sigma}}_{j,j} = 1$).

Theorem 9. *Consider the linear model (4) with Gaussian error $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, and assume (A1). Let $t > 0$ be arbitrary. When using the lasso estimator $\hat{\boldsymbol{\beta}}(\lambda)$ with $\lambda \geq 2M\sigma\sqrt{\frac{2(t^2+\log p)}{n}}$, we have*

$$\begin{aligned}
\sqrt{n}(\hat{\mathbf{b}} - \boldsymbol{\beta}^0) = \mathbf{W} + \mathbf{\Delta}, \\
\mathbf{W} := \hat{\mathbf{\Theta}}\mathbf{X}^\intercal \boldsymbol{\epsilon}/\sqrt{n} \sim N(\mathbf{0}, \sigma^2\hat{\mathbf{\Omega}}), \;\; \hat{\mathbf{\Omega}} := \hat{\mathbf{\Theta}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Theta}}^\intercal, \\
\mathbb{P}\Big\{\|\mathbf{\Delta}\|_\infty \geq 8\sqrt{n}\Big(\max_j \frac{\lambda_j}{\hat{\xi}_j^2}\Big)\frac{\lambda s}{\phi_0^2}\Big\} \leq 2\exp(-t^2).
\end{aligned}$$

Proof. Note that the only thing we need to prove here is the last probability control. First of all

$$\|\boldsymbol{\Delta}\|_\infty = \|\sqrt{n}(\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_\infty \leq \sqrt{n}\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1.$$

By the extended KKT condition (30), we have

$$\|\boldsymbol{\Delta}\|_\infty \leq \sqrt{n}\|\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}} - \mathbf{I}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq \sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \max_j \frac{\lambda_j}{\hat{\xi}_j^2}.$$

By the Theorem 2 together with the Lemma 3 (which are true under (A1)), with $\lambda \geq 2M\sigma\sqrt{\frac{2(t^2+\log p)}{n}}$, we have with probability at least $1 - 2\exp(-t^2)$

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \leq 8\lambda \frac{s}{\phi_0^2}.$$

Thus with $\lambda \geq 2M\sigma\sqrt{\frac{2(t^2+\log p)}{n}}$, we have

$$\mathbb{P}\Big\{\|\boldsymbol{\Delta}\|_\infty \leq \sqrt{n}8\lambda \frac{s}{\phi_0^2} \max_j \frac{\lambda_j}{\hat{\xi}_j^2}\Big\} \geq 1 - 2\exp(-t^2)$$

. □

By the above theorem, we can construct the confidence intervals and do the hypothesis testing accordingly.

5.3 Question

In the classical statistics situation, for assessing significance, we usually do testing in terms of fixed parameters. But nowadays, it's common to be in the situation that one selecting procedure designed somehow according to certain importance level has selected a certain amount of variables, and we want to know when to stop. In other words, we want to do the following test: conditional on what we have selected has already containing all of the signals, what is the chance for us to observe the next data? This is a meaningful conditional test. Refer to Lockhart et al. (2014).

References

A Chatterjee and S Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509, 2010.

Arindam Chatterjee and Soumendra Nath Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013.

Kengo Kato. Asymptotics for argmin processes: Convexity arguments. *Journal of Multivariate Analysis*, 100(8):1816–1829, 2009.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, Robert Tibshirani, et al. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

Michael R Osborne, Brett Presnell, and Berwin A Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Sara van de Geer, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 06 2014. doi: 10.1214/14-AOS1221. URL `http://dx.doi.org/10.1214/14-AOS1221`.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.