



□ FACULTY OF SCIENCE,
TECHNOLOGY
AND MEDICINE

UNIVERSITY OF LUXEMBOURG

THESIS FOR THE BACHELOR OF MATHEMATICS

High Dimensional Regression Models

BY JORIS LIMONIER

Supervisor: Mark PODOLSKIJ

Submitted: June 4, 2021

Abstract

Abstract goes here.

Contents

1	Introduction	1
1.1	Notes for chapter 1	1
2	Classical theory of Linear Regression	3
2.1	Linear models	3
2.2	The least squares method	4
3	Theory for LASSO in high dimensions	7
3.1	Assuming the truth is linear	7
3.2	Linear approximation of the truth	12

Chapter 1

Introduction

1.1 Notes for chapter 1

Chapter 2

Classical theory of Linear Regression

To be added

- how to get \hat{b} on page 101.
- where the χ^2 distribution comes from in page 101

2.1 Linear models

We consider the setting of having a sample of n observations

$$(\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_n, \mathbf{Y}_n)$$

where $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$, $i = 1, \dots, n$ and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$, $i = 1, \dots, n$.

Definition 2.1 (The linear model). *The relationship between an observation $\mathbf{X}_i \in \mathcal{X}$ and its outcome $\mathbf{Y}_i \in \mathcal{Y}$ can be established by a linear model, that is*

$$i = 1, \dots, n \quad \mathbf{Y}_i = \sum_{j=1}^p \beta_j \mathbf{X}_i^{(j)} + \varepsilon_i \quad (2.1)$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed (i.i.d.). Moreover, $\forall i = 1, \dots, n$, we have that $\mathbb{E}[\varepsilon_i] = 0$ and each ε_i is independent of all of the X_j , $j = 1, \dots, n$.

Instead of seeing each observation individually we can deal with all of them together by expressing the linear model in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.2)$$

Definition 2.2.

(a) \mathbf{X} is called the **design matrix**. It has dimension $n \times p$. \mathbf{X} consists of stacking the vectors relative to each observation inside of a matrix

$$X = \begin{bmatrix} - & X_1^T & - \\ & \vdots & \\ - & X_n^T & - \end{bmatrix}$$

(b) β is called the **parameter vector**. It has dimension $p \times 1$.

(c) ϵ is called the **error vector**. It has dimension $n \times 1$.

(d) \mathbf{Y} is called the **response vector**. It has dimension $n \times 1$.

Without loss of generality, and after centering and scaling if necessary, we can make the following assumptions $\forall j = 1, \dots, p$:

1. The mean of the response vector is 0:

$$\bar{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i = 0$$

2. The variance of the observations is 1:

$$\hat{\sigma}_j^2 := \frac{1}{n} \sum_{i=1}^n \left(\mathbf{X}_i^{(j)} - \bar{\mathbf{X}}^{(j)} \right)^2 = 1$$

2.2 The least squares method

We define the objective function $S(\beta)$ as follows

$$S(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon^T \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \quad (2.3)$$

which may be rewritten as

$$\begin{aligned} S(\beta) &= (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X}\beta - \beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X}\beta \\ &= \mathbf{Y}^T \mathbf{Y} - 2\beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X}\beta \end{aligned}$$

The least squares method aims at finding the vector $\hat{\beta}$ minimizing S , that is

$$\hat{\beta} := \arg \min_{\beta} S(\beta)$$

We find $\hat{\beta}$ by differentiating S with respect to β and setting the result to 0.

$$\begin{aligned} \frac{\partial}{\partial \beta} S(\hat{\beta}) &= 0 \\ \implies \frac{\partial}{\partial \hat{\beta}} \left(\mathbf{Y}^T \mathbf{Y} - 2\hat{\beta}^T \mathbf{X}^T \mathbf{Y} + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} \right) &= 0 \\ \implies -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \hat{\beta} &= 0 \\ \implies \mathbf{X}^T \mathbf{X} \hat{\beta} &= \mathbf{X}^T \mathbf{Y} \end{aligned} \tag{2.4}$$

where equation (2.4) is called the least squares normal equations.

If we assume that $\mathbf{X}^T \mathbf{X}$ is invertible, then (2.4) yields that our least squares estimator $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \tag{2.5}$$

We are interested in estimating the quality of our prediction. The residuals can help us do that.

Definition 2.3 (Residuals). *For a given set of observations \mathbf{Y} , the **residuals** (or **vector of residuals**) is the difference between the prediction of our model and the observed value, that is*

$$X(\hat{\beta} - \beta) \in \mathbb{R}^n$$

We would like a measure that indicates how far our predictions are from the measurements. We will use the prediction error for that purpose.

Definition 2.4 (Prediction error). *For a given set of observations \mathbf{Y} , the **prediction error** is the squared ℓ^2 -norm of the difference between the prediction of our model and the observed value. In other words, it is the squared residuals, that is*

$$\|X(\hat{\beta} - \beta)\|_2^2$$

Chapter 3

Theory for LASSO in high dimensions

3.1 Assuming the truth is linear

In this section, we assume that there exists some “true value” that would make the parameter β fit the observations to the predictions perfectly. We call this ideal parameter vector β^0 .

We work with an underdetermined system: there are more variables than equations, or in our context, there are more parameters than observations ($p > n$).

We define $\hat{\beta}$ as follows

$$\hat{\beta} := \arg \min_{\beta} \left\{ \frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\} \quad (3.1)$$

Lemma 3.1 (Basic Inequality).

$$\frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda \|\hat{\beta}\|_1 \leq 2 \frac{\varepsilon^T \mathbf{X}(\hat{\beta} - \beta^0)}{n} + \lambda \|\beta^0\|_1$$

Proof. By definition of $\hat{\beta}$, we have that

$$\forall \beta \quad \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2}{n} + \lambda \|\hat{\beta}\|_1 \leq \frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \|\beta\|_1$$

In particular for $\beta = \beta^0$ we have

$$\frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2}{n} + \lambda \|\hat{\beta}\|_1 \leq \frac{\|\mathbf{Y} - \mathbf{X}\beta^0\|_2^2}{n} + \lambda \|\beta^0\|_1$$

We now replace \mathbf{Y} using equation (2.2):

$$\begin{aligned}
& \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^0\|_2^2}{n} + \lambda\|\boldsymbol{\beta}^0\|_1 \\
\Rightarrow & \frac{\|(\mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}) - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|(\mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}) - \mathbf{X}\boldsymbol{\beta}^0\|_2^2}{n} + \lambda\|\boldsymbol{\beta}^0\|_1 \\
\Rightarrow & \frac{\langle \mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}, \mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\boldsymbol{\varepsilon}\|_2^2}{n} + \lambda\|\boldsymbol{\beta}^0\|_1 \\
\Rightarrow & \frac{\|\mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 + 2\langle \mathbf{X}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}), \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\boldsymbol{\varepsilon}\|_2^2}{n} + \lambda\|\boldsymbol{\beta}^0\|_1 \\
\Rightarrow & \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\langle \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0), \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\boldsymbol{\beta}^0\|_1 \\
\Rightarrow & \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\boldsymbol{\varepsilon}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n} + \lambda\|\boldsymbol{\beta}^0\|_1
\end{aligned}$$

This completes the proof. □

Let

$$\mathcal{T} := \left\{ \max_{1 \leq j \leq p} 2 \frac{|\boldsymbol{\varepsilon}^T \mathbf{X}^{(j)}|}{n} \leq \lambda_0 \right\}$$

Lemma 3.2 (Lemma 6.2.). *For all $t > 0$ and*

$$\lambda_0 := 2\sigma \sqrt{\frac{t^2 + 2 \log p}{n}}$$

we have

$$\mathbb{P}(\mathcal{T}) \geq 1 - 2 \exp[-t^2/2]$$

Proof. We define

$$V_j := \frac{\boldsymbol{\varepsilon}^T \mathbf{X}^{(j)}}{\sqrt{n\sigma^2}}$$

Then we have

$$\begin{aligned}
\mathbb{P}(\mathcal{T}) &= \mathbb{P}\left(\max_{1 \leq j \leq p} 2 \frac{|\varepsilon^T \mathbf{X}^{(j)}|}{n} \leq 2\sigma \sqrt{\frac{t^2 + 2 \log p}{n}}\right) \\
&= \mathbb{P}\left(\max_{1 \leq j \leq p} \left| \frac{\varepsilon^T \mathbf{X}^{(j)}}{\sqrt{n\sigma^2}} \right| \leq \sqrt{t^2 + 2 \log p}\right) \\
&= \mathbb{P}\left(\max_{1 \leq j \leq p} |V_j| \leq \sqrt{t^2 + 2 \log p}\right) \\
&= 1 - \mathbb{P}\left(\max_{1 \leq j \leq p} |V_j| > \sqrt{t^2 + 2 \log p}\right) \\
&= 1 - \mathbb{P}\left(\bigcup_{j=1}^p \left\{|V_j| > \sqrt{t^2 + 2 \log p}\right\}\right) \\
&\geq 1 - \sum_{j=1}^p \mathbb{P}\left(|V_j| > \sqrt{t^2 + 2 \log p}\right) \\
&\geq 1 - p \mathbb{P}\left(|V_j| > \sqrt{t^2 + 2 \log p}\right) \tag{3.2}
\end{aligned}$$

Now, let us define $\zeta := \sqrt{t^2 + 2 \log p}$. Since V_j is $\mathcal{N}(0, 1)$ -distributed and $\zeta > 0$, it follows that

$$\begin{aligned}
\mathbb{P}(V_j > \zeta) &= \frac{1}{\sqrt{2\pi}} \int_{\zeta}^{\infty} e^{-y^2/2} dy \\
&< \frac{1}{\sqrt{2\pi}} \int_{\zeta}^{\infty} \frac{y}{\zeta} e^{-y^2/2} dy \\
&= \frac{1}{\zeta \sqrt{2\pi}} \int_{\zeta}^{\infty} y e^{-y^2/2} dy \\
&= \frac{1}{\zeta \sqrt{2\pi}} e^{-\zeta^2/2}
\end{aligned}$$

We note that $p \geq 2 \implies \zeta \sqrt{2\pi} \geq 1$ therefore

$$\mathbb{P}(V_j > \zeta) < e^{-\zeta^2/2}$$

Moreover by symmetry of the $\mathcal{N}(0, 1)$ distribution,

$$\begin{aligned}
\mathbb{P}(|V_j| > \zeta) &= 2\mathbb{P}(V_j > \zeta) \\
&< 2e^{-\zeta^2/2}
\end{aligned}$$

Inserting this result into (3.2) we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{T}) &\geq 1 - p \mathbb{P}\left(|V_j| > \sqrt{t^2 + 2 \log p}\right) \\ &\geq 1 - p \frac{2}{p} \exp\left[\frac{-t^2}{2}\right] \\ &= 1 - 2 \exp\left[\frac{-t^2}{2}\right]\end{aligned}$$

□

Corollary 3.3 (Consistency of the LASSO). *Assume $\sigma^2 = 1$ for all j . We define the regularization parameter as*

$$\lambda = 4\hat{\sigma}^2 \sqrt{\frac{t^2 + 2 \log p}{n}}$$

where $\hat{\sigma}$ is some estimator of σ .

Then with probability at least $1 - \alpha$, where $\alpha := 2 \exp(-t^2/2) + \mathbb{P}(\hat{\sigma} \leq \sigma)$ we have

$$2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} \leq 3\lambda \|\beta^0\|_1$$

Lemma 3.4 (Lemma 6.3.). *We have on \mathcal{T} , with $\lambda \geq 2\lambda_0$,*

$$2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda \|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1$$

Proof. We start with the Basic Inequality

$$\frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda \|\hat{\beta}\|_1 \leq 2 \frac{\varepsilon^T \mathbf{X}(\hat{\beta} - \beta^0)}{n} + \lambda \|\beta^0\|_1$$

Now since we are on \mathcal{T} and since $2\lambda_0 \leq \lambda$

$$\begin{aligned}\frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda \|\hat{\beta}\|_1 &\leq \lambda_0 \|\hat{\beta} - \beta^0\|_1 + \lambda \|\beta^0\|_1 \\ 2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + 2\lambda \|\hat{\beta}\|_1 &\leq \lambda \|\hat{\beta} - \beta^0\|_1 + 2\lambda \|\beta^0\|_1\end{aligned}$$

Let $\beta_{j,S} := \beta_j 1\{j \in S\}$. We use the triangle inequality on the left hand side

$$\begin{aligned}\|\hat{\beta}\|_1 &= \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \\ &= \|\beta_{S_0}^0 - \beta_{S_0}^0 + \hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \\ &\geq \|\beta_{S_0}^0\|_1 - \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + \|\hat{\beta}_{S_0^c}\|_1\end{aligned}$$

whereas on the right hand side

$$\begin{aligned}\|\hat{\beta} - \beta^0\|_1 &= \|(\hat{\beta}_{S_0} + \hat{\beta}_{S_0^c}) - (\beta_{S_0}^0 + \underbrace{\beta_{S_0^c}^0}_{=0})\|_1 \\ &= \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + \|\hat{\beta}_{S_0^c}\|_1\end{aligned}$$

Injecting these two results, we get that

$$\begin{aligned}& 2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + 2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta} - \beta^0\|_1 + 2\lambda\|\beta^0\|_1 \\ \implies & 2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + 2\lambda \left(\|\beta_{S_0}^0\|_1 - \|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \right) \\ & \leq \lambda \left(\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \right) + 2\lambda\|\beta^0\|_1 \\ \implies & 2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + 2\lambda \underbrace{\|\beta_{S_0}^0\|_1}_{=\beta^0} + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1 + 2\lambda\|\beta^0\|_1 \\ \implies & 2 \frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta_{S_0}^0\|_1\end{aligned}$$

□

Definition 3.5 (Compatibility condition). *We say that the compatibility condition is met for the set S_0 , if for some $\phi_0 > 0$, and for all β satisfying $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$, it holds that*

$$\|\beta_{S_0}\|_1^2 \leq \left(\beta^T \hat{\Sigma} \beta \right) \frac{s_0}{\phi_0^2} \quad (3.3)$$

Theorem 3.6 (Theorem 6.1.). *Suppose the compatibility condition holds for S_0 . Then on \mathcal{T} , we have for $\lambda \geq 2\lambda_0$,*

$$\frac{\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 \frac{s_0}{\phi_0^2}$$

Proof. Using Lemma 3.4 we have that

$$\begin{aligned}
& 2 \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \\
&= 2 \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda \|\hat{\boldsymbol{\beta}}_{S_0} + \underbrace{\hat{\boldsymbol{\beta}}_{S_0^c} - \boldsymbol{\beta}_{S_0}^0 - \boldsymbol{\beta}_{S_0^c}^0}_{=0}\|_1 \\
&= 2 \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + \lambda \|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0\|_1 + \lambda \|\hat{\boldsymbol{\beta}}_{S_0^c}\|_1 \quad (\text{by lemma 3.4}) \\
&\leq 4\lambda \|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0\|_1 \\
&= 4\lambda \sqrt{\left(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0\right)^T \hat{\boldsymbol{\Sigma}} \left(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0\right) s_0 / \phi_0^2} \\
&\leq \sqrt{\left(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0\right)^T \mathbf{X}^T \mathbf{X} \left(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0\right)} \frac{4\lambda \sqrt{s_0}}{\phi_0 \sqrt{n}} \\
&\leq \|\mathbf{X}(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0)\|_2 \frac{4\lambda \sqrt{s_0}}{\phi_0 \sqrt{n}} \\
&\leq \|\mathbf{X}(\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}_{S_0}^0)\|_2^2 + \frac{4\lambda^2 s_0}{\phi_0^2 n}
\end{aligned}$$

Where the last inequality follows from $4uv \leq u^2 + 4v^2$. □

3.2 Linear approximation of the truth

Now $\mathbf{Y} := \mathbf{f}^0 + \boldsymbol{\varepsilon}$, therefore $\mathbb{E}[\mathbf{Y}] := \mathbf{f}^0$.

Lemma 3.7 (New version of the Basic Inequality). $\forall \boldsymbol{\beta}^* \in \mathbb{R}^p$ we have

$$\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + \lambda \|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\boldsymbol{\varepsilon}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} + \lambda \|\boldsymbol{\beta}^*\|_1 + \frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n} \quad (3.4)$$

Proof. By definition of $\hat{\boldsymbol{\beta}}$, we have that

$$\forall \boldsymbol{\beta} \quad \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{n} + \lambda \|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{n} + \lambda \|\boldsymbol{\beta}\|_1$$

In particular for $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ we have

$$\forall \boldsymbol{\beta}^* \quad \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{n} + \lambda \|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2}{n} + \lambda \|\boldsymbol{\beta}^*\|_1$$

Since $\mathbf{Y} = \mathbf{f}^0 + \boldsymbol{\varepsilon}$:

$$\begin{aligned}
& \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 \\
\Rightarrow & \frac{\|(\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|(\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*) + \boldsymbol{\varepsilon}\|_2^2}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 \\
\Rightarrow & \frac{\langle (\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}, (\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
& \leq \frac{\langle (\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*) + \boldsymbol{\varepsilon}, (\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*) + \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 \\
\Rightarrow & \frac{\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 + 2\langle \mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}, \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \\
& \leq \frac{\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 + 2\langle \mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*, \boldsymbol{\varepsilon} \rangle}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 \\
\Rightarrow & \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\boldsymbol{\varepsilon}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 + \frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n}
\end{aligned}$$

□

Lemma 3.8 (New version of Lemma 6.3.). *We have on \mathcal{T} , with $\lambda \geq 4\lambda_0$,*

$$\frac{4\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + 3\lambda\|\hat{\boldsymbol{\beta}}_{S_*^c}\|_1 \leq 5\lambda\|\hat{\boldsymbol{\beta}}_{S_*} - \boldsymbol{\beta}_{S_*}^*\|_1 + \frac{4\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n} \quad (3.5)$$

where $S_* := \{j : \beta_j^* \neq 0\}$.

Proof. We start with the Basic Inequality

$$\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\boldsymbol{\varepsilon}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 + \frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n}$$

Now since we are on \mathcal{T} and since $4\lambda_0 \leq \lambda$

$$\begin{aligned}
& \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\boldsymbol{\varepsilon}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 + \frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n} \\
\Rightarrow & 4\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + 4\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + 4\lambda\|\boldsymbol{\beta}^*\|_1 + 4\frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n}
\end{aligned}$$

We use the triangle inequality on the left hand side

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}\|_1 &= \|\hat{\boldsymbol{\beta}}_{S_*}\|_1 + \|\hat{\boldsymbol{\beta}}_{S_*^c}\|_1 \\
&= \|\boldsymbol{\beta}_{S_*}^* - \boldsymbol{\beta}_{S_*}^* + \hat{\boldsymbol{\beta}}_{S_*}\|_1 + \|\hat{\boldsymbol{\beta}}_{S_*^c}\|_1 \\
&\geq \|\boldsymbol{\beta}_{S_*}^*\|_1 - \|\hat{\boldsymbol{\beta}}_{S_*} - \boldsymbol{\beta}_{S_*}^*\|_1 + \|\hat{\boldsymbol{\beta}}_{S_*^c}\|_1
\end{aligned}$$

whereas on the right hand side

$$\begin{aligned}\|\hat{\beta} - \beta^*\|_1 &= \|(\hat{\beta}_{S_*} + \hat{\beta}_{S_*^c}) - (\beta_{S_*}^* + \underbrace{\beta_{S_*^c}^*}_{=0})\|_1 \\ &= \|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 + \|\hat{\beta}_{S_*^c}\|_1\end{aligned}$$

Injecting these two results, we get that

$$\begin{aligned}& 4 \frac{\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2}{n} + 4\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta} - \beta^*\|_1 + 4\lambda\|\beta^*\|_1 + 4 \frac{\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2}{n} \\ \implies & 4 \frac{\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2}{n} + 4\lambda \left(\|\beta_{S_*}^*\|_1 - \|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 + \|\hat{\beta}_{S_*^c}\|_1 \right) \\ & \leq \lambda \left(\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 + \|\hat{\beta}_{S_*^c}\|_1 \right) + 4\lambda\|\beta^*\|_1 + 4 \frac{\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2}{n} \\ \implies & 4 \frac{\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2}{n} + 4\lambda \underbrace{\|\beta_{S_*}^*\|_1}_{=\beta^*} + 3\lambda\|\hat{\beta}_{S_*^c}\|_1 \\ & \leq 5\lambda\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 + 4\lambda\|\beta^*\|_1 + 4 \frac{\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2}{n} \\ \implies & 4 \frac{\|\mathbf{X}\hat{\beta} - \mathbf{f}^0\|_2^2}{n} + 3\lambda\|\hat{\beta}_{S_*^c}\|_1 \leq 5\lambda\|\hat{\beta}_{S_*} - \beta_{S_*}^*\|_1 + 4 \frac{\|\mathbf{X}\beta^* - \mathbf{f}^0\|_2^2}{n}\end{aligned}$$

□

Definition 3.9 (Compatibility condition for general sets). *We say that the compatibility condition holds for the set S , if for some constant $\phi(S) > 0$, and for all β , with $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, one has*

$$\|\beta_S\|_1^2 \leq (\beta^T \hat{\sigma} \beta) \frac{|S|}{\phi^2(S)}$$

We define \mathcal{S} as the collection of sets S for which the compatibility condition holds.

Definition 3.10 (The oracle). *We define the oracle β^* as*

$$\beta^* = \arg \min_{\beta: S_\beta \in \mathcal{S}} \left\{ \frac{\|\mathbf{X}\beta - \mathbf{f}^0\|_2^2}{n} + \frac{4\lambda^2 s_\beta}{\phi^2(S_\beta)} \right\}$$

where $S_\beta := \{j : \beta_j \neq 0\}$, $s_\beta := |S_\beta|$ denotes the cardinality of S_β and the factor 4 in the right hand side comes from choosing $\lambda \geq \lambda_0$.