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

Chapter 1

Introduction

1.1 Notes for chapter 1

Chapter 2

Classical theory of Linear Regression

2.1 Linear models

We consider the setting of having a sample of n observations

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

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

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

$$i = 1, \ldots, n \qquad \mathbf{Y}_i = \sum_{j=1}^{p} \boldsymbol{\beta}_j \mathbf{X}_i^{(j)} + \boldsymbol{\varepsilon}_i \tag{2.1}$$

where $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are independent and identically distributed (i.i.d.). Moreover, $\forall\, i = 1, \ldots, n$, we have that $\mathbb{E}[\boldsymbol{\varepsilon}_i] = 0$ and each $\boldsymbol{\varepsilon}_i$ is independent of all of the $X_j,\ j = 1, \ldots, n$.

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

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) $\boldsymbol{\beta}$ *is called the* ***parameter vector****. It has dimension* $p \times 1$.

(c) $\boldsymbol{\varepsilon}$ *is called the* ***error vector****. It has dimension* $n \times 1$.

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

2.2 The least squares method

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

$$S(\boldsymbol{\beta}) := \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.3}$$

which may be rewritten as

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

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

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

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

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

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 $\hat{\boldsymbol{\beta}}$ is given by

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

Now, we can verify that our estimator has some fundamental properties. Namely, we want to make sure that it is unbiased, *i.e.* $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

$$\begin{aligned}\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}+\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\right] \\ &= \boldsymbol{\beta}\end{aligned}$$

Moreover, we want to take a look at the estimator's variance. For this purpose we will use the covariance matrix. Let $U, V \in \mathbb{R}^p$, recall that the covariance matrix is defined as

$$\mathrm{Cov}(U,V) := \mathbb{E}\left[\left(U-\mathbb{E}(U)\right)\left(V-\mathbb{E}(V)\right)^T\right] \in \mathcal{M}_{p\times p}(\mathbb{R})$$

where $\forall\, i,j = 1,\dots,p$, $\mathrm{Cov}(U,V)_{ij}$ is the covariance between U_i and V_j. In the particular case $U = V$, the diagonal of the covariance matrix is nothing else than the variance of U,that is

$$\mathrm{Var}(U)_i = \mathrm{Cov}(U,U)_{ii} \quad i = 1,\dots,p$$

Therefore the variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\mathrm{Var}(\mathbf{Y})\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]^T \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\end{aligned}$$

Now that we confirmed the above properties of $\hat{\boldsymbol{\beta}}$, 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*

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

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, that is*

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

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 $\boldsymbol{\beta}$ fit the observations to the predictions perfectly. We call this ideal parameter vector $\boldsymbol{\beta}^0$. However, we work with an underdetermined system: there are more variables than equations, or in our context, there are more parameters than observations (*i.e.* $p > n$).

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

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

Lemma 3.1 (Basic Inequality)**.**

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

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 $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ we have

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

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

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

This completes the proof. □

Let

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

Lemma 3.2 (Lemma 6.2.). *Suppose $\forall j=1,\ldots,p, \hat{\boldsymbol{\sigma}}_j^2=1$ and for all $t>0$ and*

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

we have

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

Proof. We define

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

Then we have

$$
\begin{aligned}
\mathbb{P}(\mathscr{T}) &= \mathbb{P}\left(\max_{1\le j\le p} 2\frac{\left|\varepsilon^T \mathbf{X}^{(j)}\right|}{n} \le 2\boldsymbol{\sigma}\sqrt{\frac{t^2+2\log p}{n}}\right) \\
&= \mathbb{P}\left(\max_{1\le j\le p} \left|\frac{\varepsilon^T \mathbf{X}^{(j)}}{\sqrt{n\boldsymbol{\sigma}^2}}\right| \le \sqrt{t^2+2\log p}\right) \\
&= \mathbb{P}\left(\max_{1\le j\le p} |V_j| \le \sqrt{t^2+2\log p}\right) \\
&= 1-\mathbb{P}\left(\max_{1\le j\le p} |V_j| > \sqrt{t^2+2\log p}\right) \\
&= 1-\mathbb{P}\left(\bigcup_{j=1}^{p}\left\{|V_j| > \sqrt{t^2+2\log p}\right\}\right) \\
&\ge 1-\sum_{j=1}^{p}\mathbb{P}\left(|V_j| > \sqrt{t^2+2\log p}\right) \\
&\ge 1-p\,\mathbb{P}\left(|V_j| > \sqrt{t^2+2\log p}\right) \qquad (3.2)
\end{aligned}
$$

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

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

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

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

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

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

□

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

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

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

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

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

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

Proof. We start with the Basic Inequality

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

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

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

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

Let $\boldsymbol{\beta}_{j,S} := \boldsymbol{\beta}_j 1\{j \in S\}$. We use the triangle inequality on the left hand side

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

whereas on the right hand side

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

Injecting these two results, we get that

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

□

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 $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 3\|\boldsymbol{\beta}_{S_0}\|_1$, it holds that*

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

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

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

Proof. Using Lemma 3.4 we have that

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

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

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*

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

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

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

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

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

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

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

□

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

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

where $S_* := \{j : \boldsymbol{\beta}^*_j \ne 0\}$.

Proof. We start with the Basic Inequality

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

Now since we are on $\mathscr{T}$ and since $4\lambda_0 \le \lambda$

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

We use the triangle inequality on the left hand side

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

whereas on the right hand side

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

Injecting these two results, we get that

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

□

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 $\boldsymbol{\beta}$, with $\|\boldsymbol{\beta}_{S^c}\|_1 \leq 3\|\boldsymbol{\beta}_S\|_1$, one has*

$$\|\boldsymbol{\beta}_S\|_1^2 \leq \left(\boldsymbol{\beta}^T \hat{\boldsymbol{\sigma}} \boldsymbol{\beta}\right) \frac{|S|}{\phi^2(S)}$$

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

Definition 3.10 (The oracle)**.** *We define the oracle $\boldsymbol{\beta}^*$ as*

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

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