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

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

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

where $X_i \in \mathscr{X} \subseteq \mathbb{R}^p$, $i = 1, \ldots, n$ and $Y_i \in \mathscr{Y} \subseteq \mathbb{R}$, $i = 1, \ldots, n$. In other words, each of the observations contains p covariates. In the real world this could mean having n patients, p observations per patient and trying to predict an outcome such as having a certain type of cancer.

A bit out of context

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*

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

Any assumptions on ε_i ?

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

$$S(\boldsymbol{\beta}) = \underbrace{\sum_{i=1}^{n} \varepsilon_i^2}_{\text{I would skip it}} = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.3}$$

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.

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

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

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

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*

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

Delete this sentence

However, since the residuals take into account the sign of the difference, they may partially cancel out. We would like a measure that indicates how far our predictions are from the measurements. We will use the prediction error for 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*

$$\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 \cancel{\in \mathbb{R}}$$

- What about asymptotic properties of $\hat{\beta}$ or $X\hat{\beta}$?
- What about optimality ?

Chapter 3

Theory for LASSO in high dimensions

3.1 Assuming the truth is linear

• Here you need a long motivation

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

• Why exactly this definition?

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

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

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

mention that this is the true parameter of the model

In particular for $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ we have

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

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 \le \frac{\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^0\|_2^2}{n}+\lambda\|\boldsymbol{\beta}^0\|_1 \\
\implies &\frac{\|(\mathbf{X}\boldsymbol{\beta}^0+\boldsymbol{\varepsilon})-\mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \le \frac{\|(\mathbf{X}\boldsymbol{\beta}^0+\boldsymbol{\varepsilon})-\mathbf{X}\boldsymbol{\beta}^0\|_2^2}{n}+\lambda\|\boldsymbol{\beta}^0\|_1 \\
\implies &\left(\frac{\|\mathbf{X}(\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}})+\boldsymbol{\varepsilon}\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \le \frac{\|\mathbf{X}(\boldsymbol{\beta}^0-\boldsymbol{\beta}^0)+\boldsymbol{\varepsilon}\|_2^2}{n}+\lambda\|\boldsymbol{\beta}^0\|_1\right) \text{ omit this line} \\
\implies &\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 \le \frac{\|\boldsymbol{\varepsilon}\|_2^2}{n}+\lambda\|\boldsymbol{\beta}^0\|_1 \\
\implies &\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 \le \frac{\|\boldsymbol{\varepsilon}\|_2^2}{n}+\lambda\|\boldsymbol{\beta}^0\|_1 \\
\implies &\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \le \frac{2\langle\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0),\boldsymbol{\varepsilon}\rangle}{n}+\lambda\|\boldsymbol{\beta}^0\|_1 \\
\implies &\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \le \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.

□

Some explanations are needed

Let

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

Lemma 3.2 (Lemma 6.2.). *For all $t>0$ and* ← assumption missing?

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

we have

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

Proof. We define

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

Then we have

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

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

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

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

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

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

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

omit this line

Thus by definition of ζ

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

omit this

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

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

proof of it ?

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

$$2\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\|_1 \tag{3.3}$$

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 $2\lambda_0 \leq \lambda$

why does it hold ?

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

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

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

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

□

Why do we need this condition ?

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*

$\hat{\Sigma}$?

$$\|\boldsymbol{\beta}_{S_0}\|_1^2 \leq (\boldsymbol{\beta}^T\hat{\boldsymbol{\sigma}}\boldsymbol{\beta})\,\frac{s_0}{\phi_0^2} \tag{3.4}$$

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

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

L what is lemma 3.3 ?

Proof. Using ~~lemma~~ Lemma 3.3 we have that

$$
\begin{aligned}
&2\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_1\\
&=2\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}_{S_0}+\hat{\boldsymbol{\beta}}_{S_0^c}-\boldsymbol{\beta}^0_{S_0}-\underbrace{\boldsymbol{\beta}^0_{S_0^c}}_{=0}\|_1\\
&=2\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\|_1+\lambda\|\hat{\boldsymbol{\beta}}_{S_0^c}\|_1 \quad \textit{(by lemma 3.3)}\\
&\le 4\lambda\|\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\|_1\\
&\le 4\lambda\sqrt{\left(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\right)^T\hat{\boldsymbol{\sigma}}\left(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\right)s_0/\phi_0^2}\\
&\le \sqrt{\left(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\right)^T\mathbf{X}^T\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\right)}\frac{4\lambda\sqrt{s_0}}{\phi_0\sqrt{n}}\\
&\le \|\mathbf{X}(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0})\|_2\frac{4\lambda\sqrt{s_0}}{\phi_0\sqrt{n}}\\
&\le \|\mathbf{X}(\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0})\|_2^2+\frac{4\lambda^2 s_0}{\phi_0^2 n}
\end{aligned}
$$

? why $\hat{\Sigma}$?

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

w

3.2 Linear approximation of the truth

Introduction is needed here.

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*

$$
\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{f}^0\|_2^2}{n}+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \le \frac{2\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)}{n}+\lambda\|\boldsymbol{\beta}^*\|_1+\frac{\|\mathbf{X}\boldsymbol{\beta}^*-\mathbf{f}^0\|_2^2}{n} \tag{3.5}
$$

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

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

We since $\mathbf{Y} = \mathbf{f}^0 + \boldsymbol{\varepsilon}$

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

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

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

$$\Longrightarrow \frac{\langle(\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}, (\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}\rangle}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\langle(\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*) + \boldsymbol{\varepsilon}, (\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*) + \boldsymbol{\varepsilon}\rangle}{n} + \lambda\|\boldsymbol{\beta}^*\|_1$$

$$\Longrightarrow \frac{\|\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 + 2\langle\mathbf{f}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}, \boldsymbol{\varepsilon}\rangle}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\|\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 + 2\langle\mathbf{f}^0 - \mathbf{X}\boldsymbol{\beta}^*, \boldsymbol{\varepsilon}\rangle}{n} + \lambda\|\boldsymbol{\beta}^*\|_1$$

$$\Longrightarrow \left(\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{2\langle\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \boldsymbol{\varepsilon}\rangle}{n} + \lambda\|\boldsymbol{\beta}^*\|_1 + \frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n}\right) \text{ omit}$$

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

□

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

$$\frac{4\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + 3\lambda\|\hat{\boldsymbol{\beta}}_{S_*^c}\|_1 \leq 5\lambda\|\hat{\boldsymbol{\beta}}_{S_*} - \boldsymbol{\beta}^*_{S_*}\|_1 + \frac{4\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n} \tag{3.6}$$

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 $\mathscr{T}$ and since $4\lambda_0 \leq \lambda$

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

$$\Longrightarrow \left(\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda_0\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda\|\boldsymbol{\beta}^*\|_1 + \frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n}\right) \text{ omit}$$

$$\Longrightarrow 4\frac{\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{f}^0\|_2^2}{n} + 4\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + 4\lambda\|\boldsymbol{\beta}^*\|_1 + 4\frac{\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{f}^0\|_2^2}{n}$$

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

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

Injecting these two results, we get that

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

□

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 $\lambda \geq \lambda_0$.