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**

2 Classical theory of Linear Regression **3**

2.1 Linear models 3

2.2 The least squares method 4

2.3 Maximum likelihood estimation 7

3 Theory for LASSO in high dimensions **11**

3.1 Assuming the truth is linear 12

3.2 Linear approximation of the truth 20

Chapter 1

Introduction

Chapter 2

Classical theory of Linear Regression

This chapter introduces the reader to the basics of Linear Regression. In section 2.1, we start by defining the problem and propose a framework to solve it. Section 2.2 presents the least-squares method step by step, determining the estimator and proving fundamental results on it. Finally, section 2.3 builds up from the concepts introduced with the least squares method in order to teach the maximum likelihood estimation, another method to estimate the parameters of a linear model.
Some parts of this chapter follow the book *Introduction to Linear Regression Analysis*, fifth edition by Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining.

2.1 Linear models

We consider the setting of having a sample of n observations, where each observation consists of p components. The observations are

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

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

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. Such a model is of the form*

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

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 considering each observation individually we can deal with all of them together by expressing the linear model with matrices. We use the following 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* $p \times 1$*.* $\boldsymbol{\beta}$ *is the vector we want to estimate.*

(c) $\boldsymbol{\varepsilon}$ *is called the* ***error vector****. It has dimension* $n \times 1$*.* $\boldsymbol{\varepsilon}$ *represents the difference between the linear model and the observations.*

(d) $\mathbf{Y}$ *is called the* ***response vector****. It has dimension* $n \times 1$*.* $\mathbf{Y}$ *can be seen as the outcome of the observations.*

2.2 The least squares method

The least squares method consists of finding a vector minimizing a function called the objective function. This function represents the distance between the observed data and the linear model. The smaller the objective function, the better our model approximates the observations.

Definition 2.3. *We define the* ***objective function*** S *as follows*

$$S(\boldsymbol{\beta}) := \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \tag{2.3}$$

We may rewrite the objective function as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

The goal of the least squares method is to find the vector $\hat{\boldsymbol{\beta}}$ minimizing the objective function S. In other words, we want to find

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

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

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0 \\
\implies &\frac{\partial}{\partial \boldsymbol{\beta}} \left(\mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0 \\
\implies &-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \\
\implies &\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y} \qquad \text{(LSNE)}
\end{aligned}$$

where equation (LSNE) is called the least squares normal equations. If we assume that $\mathbf{X}^T\mathbf{X}$ is invertible, then (LSNE) yields

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.4}$$

So far we found a way to compute the least square estimator $\hat{\boldsymbol{\beta}}$. Now we would like to take a look at a few of its properties, namely the expected value and the variance.
First of all we show that $\hat{\boldsymbol{\beta}}$ is unbiased, that is $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Proposition 2.4. *The least squares estimator $\hat{\boldsymbol{\beta}}$ is unbiased.*

Proof.

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

where the transition to the last line occurs because $\boldsymbol{\varepsilon}$ is centered, therefore it has null expected value, and $\boldsymbol{\beta}$ is constant, hence equal to its expected value. This completes the proof.

□

Subsequently, we want to study 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

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

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 where $U = V$, the diagonal of the covariance matrix is nothing else than the variance of U, that is:

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

Proposition 2.5. *For $i, j = 1, \dots, p$, we have that:*

(i) $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j) = \boldsymbol{\sigma}^2 \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]_{ij}$

(ii) $\mathrm{Var}(\hat{\boldsymbol{\beta}}_i) = \boldsymbol{\sigma}^2 \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]_{ii}$

Proof. (i) One can note that

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

(ii) This is a direct consequence of the first point.

□

We have now confirmed that $\hat{\boldsymbol{\beta}}$ is unbiased and we found an expression for its variance. Now we move onto estimating the quality of our prediction. The residuals can help us do that.

Definition 2.6 (Residuals)**.** *For a given set of observations* $\mathbf{Y}$*, the* ***residuals*** *(also called* ***vector of residuals****) is the difference between the model's prediction and the observed value, that is*

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

Building up from the residuals, 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 2.7. *For a given set of observations* $\mathbf{Y}$*, the* ***prediction error*** *(also called* ***residual sum of squares****) is the squared* ℓ^2*-norm of the difference between the prediction of our model and the observed value, that is*

$$\mathcal{E} = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 = \|\boldsymbol{\varepsilon}\|_2^2$$

The prediction error can be rewritten as follows:

$$\begin{aligned}
\mathcal{E} &= \|\boldsymbol{\varepsilon}\|_2^2 \\
&= \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \\
&= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \\
&= \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{Y}^T\mathbf{Y} - 2\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\underbrace{\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}}_{=\mathbf{X}^T\mathbf{Y}} \\
&= \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

We will make further use of the prediction error in chapter 3. For now, let us introduce another method to estimate parameters: the maximum likelihood estimation.

2.3 Maximum likelihood estimation

The maximum likelihood estimation has similarities with the least squares method from the previous section. However, unlike the least squares method, the goal of the maximum likelihood estimation is to maximise a so-called likelihood function. Doing so makes the observed data most probable.
As previously, we work with a normally and independently distributed error vector $\boldsymbol{\varepsilon}$ with constant variance $\boldsymbol{\sigma}^2$, in other words, $\boldsymbol{\varepsilon}$ is $\mathcal{N}(0, \boldsymbol{\sigma}^2 I)$. Then, the maximum-likelihood estimation model is given by

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

where each component of the error vector has density function:

$$f\left(\boldsymbol{\varepsilon}_i\right) = \frac{1}{\boldsymbol{\sigma}\sqrt{2\pi}} \exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}\boldsymbol{\varepsilon}_i^2\right), \quad i = 1, \ldots, n$$

Subsequently, the likelihood function is the joint density of $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$, therefore it is given by

$$\begin{aligned}
L\left(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2\right) &= \prod_{i=1}^{n} f\left(\boldsymbol{\varepsilon}_i\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}}\right)^n \exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i^2\right) \\
&= \frac{1}{(2\pi)^{n/2}\boldsymbol{\sigma}^n} \exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right)
\end{aligned}$$

which, using that $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, we may rewrite as a function of $\mathbf{Y}$ and $\mathbf{X}$, rather than of $\boldsymbol{\varepsilon}$. It also remains a function of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. Thus it becomes:

$$L\left(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2\right) = \frac{1}{(2\pi)^{n/2}\boldsymbol{\sigma}^n} \exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right)$$

We can apply the natural logarithm function on both sides of the above equation. Doing so cancels out the exponential and make our lives simpler for the rest of the analysis. Since the natural logarithm $\ln : \mathbb{R}^{>0} \to \mathbb{R}$ is an increasing function, it will not disturb the search for the argument maximizing L. We call the new function the log-likelihood function.

$$\begin{aligned}
&\ln\left[L\left(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2\right)\right] \\
&= \ln\left[\frac{1}{(2\pi)^{n/2}\boldsymbol{\sigma}^n} \exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right)\right] \\
&= -\ln[(2\pi)^{n/2}\boldsymbol{\sigma}^n] - \frac{1}{2\boldsymbol{\sigma}^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2}\ln(2\pi) - n\ln(\boldsymbol{\sigma}) - \frac{1}{2\boldsymbol{\sigma}^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}$$

Now, with a fixed $\boldsymbol{\sigma}$, the only term that can vary is $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. One can easily notice that we need to minimize this term in order to maximize the log-likelihood function. We will not repeat this argument since we already went through it in equation (LSNE). We had thereafter obtained in equation (2.4) that, given that $\mathbf{X}^T\mathbf{X}$ is invertible, the result is

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

We inject this result in the above and we get that the maximum-likelihood estimator of $\boldsymbol{\sigma}^2$ is given by

$$\tilde{\boldsymbol{\sigma}}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n}$$

In this chapter, we introduced the fundamentals of the classical theory of linear regression. First, we defined and introduced linear models. We saw that they give us a structure to work with observed data. Subsequently, we presented two major methods of the field of linear regression: the least squares method and the maximum likelihood estimation. The former aims at minimizing the squared euclidian distance between the observed data and the model's estimation. The latter considers a so-called maximum likelihood function (more precisely its natural logarithm) and maximizes it. The goal being to create a framework such that the observed data is the most probable outcome.
Now we move onto chapter 3 where we study yet another method of linear regression analysis, namely LASSO. We will have the occasion to understand how it differs from the previous two techniques seen in chapter 2, what its strengths are and when to choose it over other methods.

Chapter 3

Theory for LASSO in high dimensions

In chapter 2, we considered square or almost-square systems ($n \approx p$). In this chapter however, we work with underdetermined systems, which means that we have few observations but each with many components. In fact, this difference may even be substantial and we will introduce general methods to deal with such cases. Several linear regresion methods exist in order to work for underdetermined systems but we will introduce only one, namely LASSO. Our exploration will consist of two parts. In the first one, we will work with the heuristic assumption that the underlying true model is linear. In the second part, we will acknowledge that the underlying true model is not linear, but we will argue and show that it can be well-approximated by a linear model, accounting for the additional nonlinear part.
LASSO (Least Absolute Shrinkage and Selection Operator) is a linear regression method which takes into account the ℓ_1-norm of the estimator $\boldsymbol{\beta}$ as a penalization parameter. LASSO aims at minimizing the following quantity over $\boldsymbol{\beta}$:

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

Because of the ℓ_1 term, the model is able to indentify and select the relevant parameters, while the other parameters can be set to zero. This makes LASSO shine, in comparison to other (non-penalized) linear regression techniques, when there is a great number of parameters.
Some parts of this chapter follow *Statistics for High-Dimensional Data* by Peter Bühlmann and Sara van de Geer.

In this section, we assume that there exists some underlying "true value" for the parameter $\boldsymbol{\beta}$ that would make the model ideally fit the observations to the predictions. We call this parameter vector $\boldsymbol{\beta}^0$. We want to be as close as possible

from such a parameter.
One of the roadblocks that prevents us from finding $\boldsymbol{\beta}^0$ is that we do not know which of the $\boldsymbol{\beta}_j$ are non-zero. Therefore, one of our goals is to identify the so-called *active set*, which is the set of non-zero components of $\boldsymbol{\beta}^0$. We write it S_0 and we define it as follows:

$$S_0 := \left\{1 \leq j \leq p : \boldsymbol{\beta}_j^0 \neq 0\right\}$$

Let us also introduce the number of non-zero parameters s_0. We call s_0 the sparsity index of $\boldsymbol{\beta}^0$ and it is defined as the cardinality of S_0, that is $s_0 := |S_0|$.
We must account for the fact that we do not know S_0. Using a regularization parameter fulfills that purpose as it ensures that the components of our estimator will remain "small". Thus, keeping the results from chapter 2 in mind, we define $\hat{\boldsymbol{\beta}}$ as follows:

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

3.1 Assuming the truth is linear

We will build several results from the following lemma, which we logically call the Basic Inequality.

Lemma 3.1 (Basic Inequality)**.** *The following holds:*

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

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. $\square$

In the case of quadratic loss, we call the term

$$\frac{2\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)}{n}$$

the empirical process. It depends on the measurement error, however we can easily bound it from above by replacing each component of $\boldsymbol{\varepsilon}^T\mathbf{X}$ by the maximum over $j = 1, \ldots, p$. We get

$$\frac{2\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)}{n} \leq \left(\max_{1\leq j\leq p} 2|\boldsymbol{\varepsilon}^T\mathbf{X}^{(j)}|\right)\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_1$$

Now let us define $\mathscr{T}$ as follows:

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

And in order to account for the random part of the problem, we assume that $\lambda \geq 2\lambda_0$.
We also define the Gram matrix, scaled by $1/n$, as follows:

$$\hat{\boldsymbol{\Sigma}} := \frac{\mathbf{X}^T\mathbf{X}}{n}$$

and we write its diagonal elements as

$$\hat{\boldsymbol{\sigma}}_j := \hat{\boldsymbol{\Sigma}}_{jj}, \quad j = 1, \ldots, p$$

$\mathscr{T}$ gives us a useful upper bound if we can find a value of λ_0 such that $\mathscr{T}$ has probability close to 1. In Lemma 3.2, will find such an upper bound and show that $\mathscr{T}$ has indeed probability close to 1.

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

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

we have

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

Proof. We define

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

Then we have

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

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

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

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

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

$\square$

We extend lemma 3.2 with the following corollary, which gives us an upper bound for the empirical process, which takes the value of $\boldsymbol{\sigma}$ into account.

Corollary 3.3 (Consistency of the LASSO)**.** *Assume $\hat{\boldsymbol{\sigma}}_j^2 = 1$ for all $j = 1, \ldots, p$. For some $t > 0$, let the regularization parameter be:*

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

where $\hat{\boldsymbol{\sigma}}$ is an estimator of $\boldsymbol{\sigma}$.
Then with probability at least $1-\alpha$, where

$$\alpha := 2\exp\left[\frac{-t^2}{2}\right] + \mathbb{P}(\{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\})$$

we have

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

Proof. Recall that we defined

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

and lemma 3.2, we know that

$$\lambda_0 = 2\boldsymbol{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}} \implies \mathbb{P}(\mathscr{T}) \geq 1 - 2\exp\left[\frac{-t^2}{2}\right]$$

So if $\hat{\boldsymbol{\sigma}} > \boldsymbol{\sigma}$, replacing the latter by the former will result in a weaker statement, which therefore still holds, *i.e.*

$$\lambda_0 = 2\hat{\boldsymbol{\sigma}}\sqrt{\frac{t^2 + 2\log p}{n}} \implies \mathbb{P}(\mathscr{T}) \geq 1 - 2\exp\left[\frac{-t^2}{2}\right]$$

Now if we are on $\mathscr{T}$, starting with the basic inequality 3.1:

$$\begin{aligned}
&\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} + 2\lambda_0\|\hat{\boldsymbol{\beta}}\|_1 \leq 2\frac{\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n} + 2\lambda_0\|\boldsymbol{\beta}^0\|_1 \\
\implies &\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} \leq \underbrace{2\frac{\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{n}}_{\text{bounded on } \mathscr{T}} + 2\lambda_0\left(\|\boldsymbol{\beta}^0\|_1 - \|\hat{\boldsymbol{\beta}}\|_1\right) \\
\implies &\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} \leq \lambda_0(\underbrace{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 - \|\hat{\boldsymbol{\beta}}\|_1}_{\leq\|\boldsymbol{\beta}^0\|_1} + 2\|\boldsymbol{\beta}^0\|_1) \\
\implies &\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2}{n} \leq 3\lambda_0\|\boldsymbol{\beta}^0\|_1
\end{aligned}$$

Therefore with $\lambda = 2\lambda_0 = 4\hat{\boldsymbol{\sigma}}\sqrt{\frac{t^2+2\log p}{n}}$, we obtain the desired result. This occurs if we are on $\mathscr{T}$ and $\hat{\boldsymbol{\sigma}} > \boldsymbol{\sigma}$, which means that it doesn't occur with probability

$$\begin{aligned}
&\mathbb{P}(\neg\mathscr{T} \cup \{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\}) \\
&= \mathbb{P}(\neg\mathscr{T}) + \mathbb{P}(\{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\}) - \mathbb{P}(\neg\mathscr{T} \cap \{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\}) \\
&\leq \mathbb{P}(\neg\mathscr{T}) + \mathbb{P}(\{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\}) \\
&= 1 - \mathbb{P}(\mathscr{T}) + \mathbb{P}(\{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\}) \\
&\leq 1 - \left(1 - 2\exp\left[\frac{-t^2}{2}\right]\right) + \mathbb{P}(\{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\}) \\
&= 2\exp\left[\frac{-t^2}{2}\right] + \mathbb{P}(\{\hat{\boldsymbol{\sigma}} \leq \boldsymbol{\sigma}\})
\end{aligned}$$

which concludes the proof. □

Now we want to go one step further and we need to introduce some supplementary notation. We write $\boldsymbol{\beta}_{j,S}$ for the vector $\boldsymbol{\beta}$, where we keep only the components which belong to S and set the other ones to 0. That is:

$$\boldsymbol{\beta}_{j,S} := \boldsymbol{\beta}_j \mathbf{1}_{j\in S}$$

Naturally, since any given index either belongs to a set or to its complement, we have that

$$\boldsymbol{\beta} = \boldsymbol{\beta}_{j,S} + \boldsymbol{\beta}_{j,S^C}$$

We now have the necessary notation so we proceed with the lemma.

Lemma 3.4. *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 we assumed that $\lambda \geq 2\lambda_0$

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

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

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

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

□

We now introduce an important assumption that we will require in order to deduce several results in the future. Let us first define it, then we will discuss its importance.

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*

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

The compatibility condition gives us a permission to cross the bridge between the ℓ_1 world and the ℓ_2 world. That bridge is obtained thanks the the Cauchy-Scharz inequality, and it is:

$$\left\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\right\|_1 \leq \sqrt{s_0}\left\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\right\|_2$$

Moreover we have that

$$\begin{aligned}\frac{\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\right\|_2^2}{n} &= \frac{\left[\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\right]^T\left[\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\right]}{n} \\ &= \left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)^T\frac{\mathbf{X}^T\mathbf{X}}{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right) \\ &= \left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)^T\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\end{aligned}$$

And we hope that for some constant $\phi_0 > 0$:

$$\left\|\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\right\|_2^2 \leq \frac{\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)^T\hat{\boldsymbol{\Sigma}}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)}{\phi_0^2}$$

however, this is not the case for all $\boldsymbol{\beta}$. But lemma 3.4 tells us that

$$\left\|\hat{\beta}_{S_0^c}\right\|_1 \leq 3\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_1$$

as long as we are on $\mathscr{T}$.
We may now use the compatibility condition to get the following theorem.

Theorem 3.6. *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

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

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

3.2 Linear approximation of the truth

In this section, we take into account the fact that the expected value of $\mathbf{Y}$ may not be 0. Several points run in a similar fashion to section 3.1 but carry extra terms, which account for that difference.
From now on, we consider the expected value of $\mathbf{Y}$ to be $\mathbb{E}[\mathbf{Y}] := \mathbf{f}^0$. The basic inequality still holds, but carries extra terms

Lemma 3.7 (Basic Inequality for non-zero expected value)**.** *For all $\boldsymbol{\beta}^* \in \mathbb{R}^p$ we have*

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

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

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

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

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

□

We can also adapt lemma 3.4 for non-zero expected value.

Lemma 3.8. *We have on $\mathscr{T}$, with $\lambda \geq 4\lambda_0$:*

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

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

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 $4\lambda_0 \leq \lambda$

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

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

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

Injecting these two results, we get that

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

□

The compatibility condition in its form for non-zero expected value is given as follows.

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 use the notation $\mathscr{S}$ for the collection of sets S for which the compatibility condition holds. Then, let us define the oracle, which is the best possible value for the estimator $\boldsymbol{\beta}$.

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 $\lambda \geq \lambda_0$.