CHAPTER 2

Linear Regression Model

In this chapter, we consider the following regression model:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is sub-Gaussian with variance proxy σ^2 and such that $\mathbb{E}[\varepsilon] = 0$. Our goal is to estimate the function f under a linear assumption. Namely, we assume that $x \in \mathbb{R}^d$ and $f(x) = x^\top \theta^*$ for some unknown $\theta^* \in \mathbb{R}^d$.

2.1 FIXED DESIGN LINEAR REGRESSION

Depending on the nature of the *design* points $X_1, \ldots, X_n$, we will favor a different measure of risk. In particular, we will focus either on *fixed* or *random* design.

Random design

The case of random design corresponds to the statistical learning setup. Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ be $n+1$ i.i.d. random couples. Given $(X_1, Y_1), \ldots, (X_n, Y_n)$ the goal is construct a function $\hat{f}_n$ such that $\hat{f}_n(X_{n+1})$ is a good predictor of Y_{n+1}. Note that when $\hat{f}_n$ is constructed, X_{n+1} is still unknown and we have to account for what value it is likely to take.

Consider the following example from [HTF01, Section 3.2]. The response variable Y is the log-volume of a cancerous tumor, and the goal is to predict it based on $X \in \mathbb{R}^6$, a collection of variables that are easier to measure (age of patient, log-weight of prostate, ...). Here the goal is clearly to construct f for *prediction* purposes. Indeed, we want to find an automatic mechanism that outputs a good prediction of the log-weight of the tumor given certain inputs for a new (unseen) patient.

A natural measure of performance here is the L_2-risk employed in the introduction:

$$R(\hat{f}_n) = \mathbb{E}[Y_{n+1} - \hat{f}_n(X_{n+1})]^2 = \mathbb{E}[Y_{n+1} - f(X_{n+1})]^2 + \|\hat{f}_n - f\|^2_{L^2(P_X)}\,,$$

where P_X denotes the marginal distribution of X_{n+1}. It measures how good the prediction of Y_{n+1} is in average over realizations of X_{n+1}. In particular, it does not put much emphasis on values of X_{n+1} that are not very likely to occur.

Note that if the ε_i are random variables with variance σ^2 then, one simply has $R(\hat{f}_n) = \sigma^2 + \|\hat{f}_n - f\|^2_{L^2(P_X)}$. Therefore, for random design, we will focus on the squared L_2 norm $\|\hat{f}_n - f\|^2_{L^2(P_X)}$ as a measure of accuracy. It measures how close $\hat{f}_n$ is to the unknown f *in average* over realizations of X_{n+1}.

Fixed design

In fixed design, the points (or vectors) $X_1, \ldots, X_n$ are *deterministic*. To emphasize this fact, we use lowercase letters $x_1, \ldots, x_n$ to denote fixed design. Of course, we can always think of them as realizations of a random variable but the distinction between fixed and random design is deeper and significantly affects our measure of performance. Indeed, recall that for random design, we look at the performance *in average* over realizations of X_{n+1}. Here, there is no such thing as a marginal distribution of X_{n+1}. Rather, since the design points $x_1, \ldots, x_n$ are considered deterministic, our goal is estimate f *only* at these points. This problem is sometimes called *denoising* since our goal is to recover $f(x_1), \ldots, f(x_n)$ given noisy observations of these values.

In many instances, fixed design can be recognized from their structured form. A typical example is the *regular design* on $[0, 1]$, given by $x_i = i/n, i = 1, \ldots, n$. Interpolation between these points is possible under smoothness assumptions.

Note that in fixed design, we observe $\mu^* + \varepsilon$, where $\mu^* = \big(f(x_1), \ldots, f(x_n)\big)^\top \in \mathbb{R}^n$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon)^\top \in \mathbb{R}^n$ is sub-Gaussian with variance proxy σ^2. Instead of a functional estimation problem, it is often simpler to view this problem as a vector problem in $\mathbb{R}^n$. This point of view will allow us to leverage the Euclidean geometry of $\mathbb{R}^n$.

In the case of fixed design, we will focus on the *Mean Squared Error* (MSE) as a measure of performance. It is defined by

$$\mathsf{MSE}(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^n \big(\hat{f}_n(x_i) - f(x_i)\big)^2\,.$$

Equivalently, if we view our problem as a vector problem, it is defined by

$$\mathsf{MSE}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^n \big(\hat{\mu}_i - \mu_i^*\big)^2 = \frac{1}{n}|\hat{\mu} - \mu^*|_2^2\,.$$

Often, the design vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ are stored in a $n \times d$ design matrix $\mathbb{X}$, whose jth row is given by $x_j^\top$. With this notation, the linear regression model can be written

$$Y = \mathbb{X}\theta^* + \varepsilon\,, \tag{2.2}$$

where $Y = (Y_1, \ldots, Y_n)^\top$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Moreover,

$$\mathsf{MSE}(\mathbb{X}\hat\theta) = \frac{1}{n}|\mathbb{X}(\hat\theta - \theta^*)|_2^2 = (\hat\theta - \theta^*)^\top \frac{\mathbb{X}^\top \mathbb{X}}{n}(\hat\theta - \theta^*)\,. \tag{2.3}$$

A natural example of fixed design regression is image denoising. Assume that $\mu_i^*, i \in 1, \ldots, n$ is the grayscale value of pixel i of an image. We do not get to observe the image μ^* but rather a noisy version of it $Y = \mu^* + \varepsilon$. Given a library of d images $\{x_1, \ldots, x_d\}, x_j \in \mathbb{R}^n$, our goal is to recover the original image μ^* using linear combinations of the images $x_1, \ldots, x_d$. This can be done fairly accurately (see Figure 2.1).

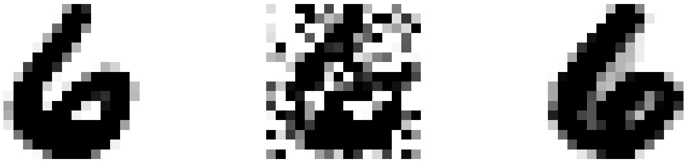

Figure 2.1. Reconstruction of the digit “6”: Original (left), Noisy (middle) and Reconstruction (right). Here $n = 16 \times 16 = 256$ pixels. Source [RT11].

As we will see in Remark 2.3, choosing fixed design properly also ensures that if $\mathsf{MSE}(\hat f)$ is small for some linear estimator $\hat f(x) = x^\top \hat\theta$, then $|\hat\theta - \theta^*|_2^2$ is also small.

In this chapter we only consider the fixed design case.

2.2 LEAST SQUARES ESTIMATORS

Throughout this section, we consider the regression model (2.2) with fixed design.

Unconstrained least squares estimator

Define the (unconstrained) *least squares estimator* $\hat\theta^{\text{LS}}$ to be any vector such that

$$\hat\theta^{\text{LS}} \in \mathop{\mathrm{argmin}}_{\theta \in \mathbb{R}^d} |Y - \mathbb{X}\theta|_2^2\,.$$

Note that we are interested in estimating $\mathbb{X}\theta^*$ and not θ^* itself, so by extension, we also call $\hat{\mu}^{\text{LS}} = \mathbb{X}\hat{\theta}^{\text{LS}}$ least squares estimator. Observe that $\hat{\mu}^{\text{LS}}$ is the projection of Y onto the column span of $\mathbb{X}$.

It is not hard to see that least squares estimators of θ^* and $\mu^* = \mathbb{X}\theta^*$ are maximum likelihood estimators when $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Proposition 2.1. The least squares estimator $\hat{\mu}^{\text{LS}} = \mathbb{X}\hat{\theta}^{\text{LS}} \in \mathbb{R}^n$ satisfies

$$\mathbb{X}^\top \hat{\mu}^{\text{LS}} = \mathbb{X}^\top Y\,.$$

Moreover, $\hat{\theta}^{\text{LS}}$ can be chosen to be

$$\hat{\theta}^{\text{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y\,,$$

where $(\mathbb{X}^\top \mathbb{X})^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbb{X}^\top \mathbb{X}$.

Proof. The function $\theta \mapsto |Y - \mathbb{X}\theta|_2^2$ is convex so any of its minima satisfies

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 = 0$$

Where ∇_θ is the gradient operator. Using matrix calculus, we find

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 = \nabla_\theta \big\{ |Y|_2^2 + -2Y^\top \mathbb{X}\theta + \theta^\top \mathbb{X}^\top \mathbb{X}\theta \big\} = -2(Y^\top \mathbb{X} - \theta^\top \mathbb{X}^\top \mathbb{X})^\top\,.$$

Therefore, solving $\nabla_\theta |Y - \mathbb{X}\theta|_2^2 = 0$ yields

$$\mathbb{X}^\top \mathbb{X}\theta = \mathbb{X}^\top Y\,.$$

It concludes the proof of the first statement. The second statement follows from the definition of the Moore-Penrose pseudoinverse. □

We are now going to prove our first result on the finite sample performance of the least squares estimator for fixed design.

Theorem 2.2. *Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Then the least squares estimator $\hat{\theta}^{\text{LS}}$ satisfies*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}})\big] = \frac{1}{n}\mathbb{E}|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma^2 \frac{r}{n}\,,$$

where $r = \text{rank}(\mathbb{X}^\top \mathbb{X})$. Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}) \lesssim \sigma^2 \frac{r + \log(1/\delta)}{n}\,.$$

Proof. Note that by definition

$$|Y - \mathbb{X}\hat{\theta}^{\text{LS}}|_2^2 \le |Y - \mathbb{X}\theta^*|_2^2 = |\varepsilon|_2^2\,. \tag{2.4}$$

Moreover,

$$|Y - \mathbb{X}\hat{\theta}^{\text{LS}}|_2^2 = |\mathbb{X}\theta^* + \varepsilon - \mathbb{X}\hat{\theta}^{\text{LS}}|_2^2 = |\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 - 2\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*) + |\varepsilon|_2^2\,.$$

Therefore, we get

$$|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \le 2\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*) = 2|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2 \frac{\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)}{|\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)|_2} \tag{2.5}$$

Note that it is difficult to control

$$\frac{\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)}{|\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)|_2}$$

as $\hat{\theta}^{\text{LS}}$ depends on ε and the dependence structure of this term may be complicated. To remove this dependency, a traditional technique is "sup-out" $\hat{\theta}^{\text{LS}}$. This is typically where maximal inequalities are needed. Here we have to be a bit careful.

Let $\Phi = [\phi_1, \dots, \phi_r] \in \mathbb{R}^{n\times r}$ be an orthonormal basis of the column span of $\mathbb{X}$. In particular, there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*) = \Phi\nu$. It yields

$$\frac{\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)}{|\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)|_2} = \frac{\varepsilon^\top \Phi\nu}{|\Phi\nu|_2} = \frac{\varepsilon^\top \Phi\nu}{|\nu|_2} = \tilde{\varepsilon}^\top \frac{\nu}{|\nu|_2} \le \sup_{u\in\mathcal{B}_2} \tilde{\varepsilon}^\top u\,,$$

where $\mathcal{B}_2$ is the unit ball of $\mathbb{R}^r$ and $\tilde{\varepsilon} = \Phi^\top\varepsilon$. Thus

$$|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \le 4 \sup_{u\in\mathcal{B}_2} (\tilde{\varepsilon}^\top u)^2\,,$$

Next, note that for any $u \in \mathcal{S}^{r-1}$, it holds $|\Phi u|_2^2 = u^\top\Phi^\top\Phi u = u^\top u = 1$ so that for any $s \in \mathbb{R}$, we have

$$\mathbb{E}[e^{s\tilde{\varepsilon}^\top u}] = \mathbb{E}[e^{s\varepsilon^\top \Phi u}] \le \mathrm{e}^{\frac{s^2\sigma^2}{2}}\,.$$

Therefore, $\tilde{\varepsilon} \sim \mathsf{subG}_r(\sigma^2)$.

To conclude the bound in expectation, observe that Lemma 1.4 yields

$$4\mathbb{E}\big[\sup_{u\in\mathcal{B}_2} (\tilde{\varepsilon}^\top u)^2\big] = 4\sum_{i=1}^r \mathbb{E}[\tilde{\varepsilon}_i^2] \le 16\sigma^2 r\,.$$

Moreover, with probability $1-\delta$, it follows from the last step in the proof[1] of Theorem 1.19 that

$$\sup_{u\in\mathcal{B}_2} (\tilde{\varepsilon}^\top u)^2 \le 8\log(6)\sigma^2 r + 8\sigma^2\log(1/\delta)\,.$$

□

Remark 2.3. If $d \le n$ and $B := \frac{\mathbb{X}^\top\mathbb{X}}{n}$ has rank d, then we have

$$|\hat{\theta}^{\text{LS}} - \theta^*|_2^2 \le \frac{\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}})}{\lambda_{\min}(B)}\,,$$

and we can use Theorem 2.2 to bound $|\hat{\theta}^{\text{LS}} - \theta^*|_2^2$ directly.

[1] we could use Theorem 1.19 directly here but at the cost of a factor 2 in the constant.

Constrained least squares estimator

Let $K \subset \mathbb{R}^d$ be a symmetric convex set. If we know *a priori* that $\theta^* \in K$, we may prefer a *constrained least squares* estimator $\hat\theta_K^{\text{LS}}$ defined by

$$\hat\theta_K^{\text{LS}} \in \mathop{\mathrm{argmin}}_{\theta \in K} |Y - \mathbb{X}\theta|_2^2 \,.$$

Indeed, the fundamental inequality (2.4) would still hold and the bounds on the MSE may be smaller. Indeed, (2.5) can be replaced by

$$|\mathbb{X}\hat\theta_K^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \le 2\varepsilon^\top \mathbb{X}(\hat\theta_K^{\text{LS}} - \theta^*) \le 2 \sup_{\theta \in K-K} (\varepsilon^\top \mathbb{X}\theta) \,,$$

where $K - K = \{x - y \,:\, x, y \in K\}$. It is easy to see that if K is symmetric ($x \in K \Leftrightarrow -x \in K$) and convex, then $K - K = 2K$ so that

$$2 \sup_{\theta \in K-K} (\varepsilon^\top \mathbb{X}\theta) = 4 \sup_{v \in \mathbb{X}K} (\varepsilon^\top v)$$

where $\mathbb{X}K = \{\mathbb{X}\theta \,:\, \theta \in K\} \subset \mathbb{R}^n$. This is a measure of the size (width) of $\mathbb{X}K$. If $\varepsilon \sim \mathcal{N}(0, I_d)$, the expected value of the above supremum is actually called *Gaussian width* of $\mathbb{X}K$. Here, ε is not Gaussian but sub-Gaussian and similar properties will hold.

ℓ_1 constrained least squares

Assume here that $K = \mathcal{B}_1$ is the unit ℓ_1 ball of $\mathbb{R}^d$. Recall that it is defined by

$$\mathcal{B}_1 = \Big\{ x \in \mathbb{R}^d \,:\, \sum_{i=1}^d |x_i| \le 1 \Big\} \,,$$

and it has exactly $2d$ vertices $\mathcal{V} = \{e_1, -e_1, \dots, e_d, -e_d\}$, where e_j is the j-th vector of the canonical basis of $\mathbb{R}^d$ and is defined by

$$e_j = (0, \dots, 0, \underbrace{\quad 1 \quad}_{j\text{th position}}, 0, \dots, 0)^\top \,.$$

It implies that the set $\mathbb{X}K = \{\mathbb{X}\theta, \theta \in K\} \subset \mathbb{R}^n$ is also a polytope with at most $2d$ vertices that are in the set $\mathbb{X}\mathcal{V} = \{\mathbb{X}_1, -\mathbb{X}_1, \dots, \mathbb{X}_d, -\mathbb{X}_d\}$ where $\mathbb{X}_j$ is the j-th column of $\mathbb{X}$. Indeed, $\mathbb{X}K$ is a obtained by rescaling and embedding (resp. projecting) the polytope K when $d \le n$ (resp., $d \ge n$). Note that some columns of $\mathbb{X}$ might not be vertices of $\mathbb{X}K$ so that $\mathbb{X}\mathcal{V}$ might be a strict superset of the set of vertices of $\mathbb{X}K$.

Theorem 2.4. *Let $K = \mathcal{B}_1$ be the unit ℓ_1 ball of $\mathbb{R}^d, d \ge 2$ and assume that $\theta^* \in \mathcal{B}_1$. Moreover, assume the conditions of Theorem 2.2 and that the columns of $\mathbb{X}$ are normalized in such a way that $\max_j |\mathbb{X}_j|_2 \le \sqrt{n}$. Then the constrained least squares estimator $\hat\theta_{\mathcal{B}_1}^{\text{LS}}$ satisfies*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat\theta_{\mathcal{B}_1}^{\text{LS}})\big] = \frac{1}{n}\mathbb{E}|\mathbb{X}\hat\theta_{\mathcal{B}_1}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma\sqrt{\frac{\log d}{n}} \,,$$

Moreover, for any $\delta > 0$, with probability $1-\delta$, it holds

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}_{\mathcal{B}_1}) \lesssim \sigma\sqrt{\frac{\log(d/\delta)}{n}}\,.$$

Proof. From the considerations preceding the theorem, we got that

$$|\mathbb{X}\hat\theta^{\text{LS}}_{\mathcal{B}_1} - \mathbb{X}\theta^*|_2^2 \le 4\sup_{v\in\mathbb{X}K}(\varepsilon^\top v)$$

Observe now that since $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$,then for any column $\mathbb{X}_j$ such that $|\mathbb{X}_j|_2 \le \sqrt{n}$, the random variable $\varepsilon^\top\mathbb{X}_j \sim \mathsf{subG}(n\sigma^2)$. Therefore, applying Theorem 1.16, we get the bound on $\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}_K)\big]$ and for any $t \ge 0$,

$$\mathbb{P}\big[\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}_K) > t\big] \le \mathbb{P}\big[\sup_{v\in\mathbb{X}K}(\varepsilon^\top v) > nt/4\big] \le 2de^{-\frac{nt^2}{32\sigma^2}}$$

To conclude the proof, we find t such that

$$2de^{-\frac{nt^2}{32\sigma^2}} \le \delta \;\Leftrightarrow\; t^2 \ge 32\sigma^2\frac{\log(2d)}{n} + 32\sigma^2\frac{\log(1/\delta)}{n}\,.$$

□

Note that the proof of Theorem 2.2 also applies to $\hat\theta^{\text{LS}}_{\mathcal{B}_1}$ (exercise!) so that $\hat\theta^{\text{LS}}_{\mathcal{B}_1}$ benefits from the best of both rates.

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}_{\mathcal{B}_1})\big] \lesssim \min\Big(\frac{r}{n}, \sqrt{\frac{\log d}{n}}\Big)\,.$$

This is called an *elbow effect*. The elbow takes place around $r \simeq \sqrt{n}$ (up to logarithmic terms).

ℓ_0 constrained least squares

We abusively call ℓ_0 norm of a vector $\theta \in \mathbb{R}^d$ it number of non-zero coefficient. It is denoted by

$$|\theta|_0 = \sum_{j=1}^d \mathbb{1}(\theta_j \neq 0)\,.$$

We call a vector θ with "small" ℓ_0 norm a *sparse* vector. More precisely, if $|\theta|_0 \le k$, we say that θ is a k-sparse vector. We also call *support* of θ the set

$$\mathrm{supp}(\theta) = \big\{j \in \{1,\ldots,d\} \,:\, \theta_j \neq 0\big\}$$

so that $|\theta|_0 = \mathrm{card}(\mathrm{supp}(\theta)) =: |\,\mathrm{supp}(\theta)|\,.$

Remark 2.5. The ℓ_0 terminology and notation comes from the fact that

$$\lim_{q\to 0^+}\sum_{j=1}^d |\theta_j|^q = |\theta|_0$$

Therefore it is really $\lim_{q\to0^+}|\theta|_q^q$ but the notation $|\theta|_0^0$ suggests too much that it is always equal to 1.

By extension, denote by $\mathcal{B}_0(k)$ the ℓ_0 ball of $\mathbb{R}^d$, i.e., the set of k-sparse vectors, defined by

$$\mathcal{B}_0(k) = \{\theta \in \mathbb{R}^d \,:\, |\theta|_0 \le k\}\,.$$

In this section, our goal is to control the MSE of $\hat\theta^{\text{LS}}_K$ when $K = \mathcal{B}_0(k)$. Note that computing $\hat\theta^{\text{LS}}_{\mathcal{B}_0(k)}$ essentially requires computing $\binom{d}{k}$ least squares estimators, which is an exponential number in k. In practice this will be hard (or even impossible) but it is interesting to understand the statistical properties of this estimator and to use them as a benchmark.

Theorem 2.6. *Fix a positive integer $k \le d/2$. Let $K = \mathcal{B}_0(k)$ be set of k-sparse vectors of $\mathbb{R}^d$ and assume that $\theta^* \in \mathcal{B}_0(k)$. Moreover, assume the conditions of Theorem 2.2. Then, for any $\delta > 0$, with probability $1-\delta$, it holds*

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2}{n}\log\binom{d}{2k} + \frac{\sigma^2 k}{n} + \frac{\sigma^2}{n}\log(1/\delta)\,.$$

Proof. We begin as in the proof of Theorem 2.2 to get (2.5):

$$|\mathbb{X}\hat\theta^{\text{LS}}_K - \mathbb{X}\theta^*|_2^2 \le 2\varepsilon^\top\mathbb{X}(\hat\theta^{\text{LS}}_K - \theta^*) = 2|\mathbb{X}\hat\theta^{\text{LS}}_K - \mathbb{X}\theta^*|_2 \frac{\varepsilon^\top\mathbb{X}(\hat\theta^{\text{LS}}_K - \theta^*)}{|\mathbb{X}(\hat\theta^{\text{LS}}_K - \theta^*)|_2}\,.$$

We know that both $\hat\theta^{\text{LS}}_K$ and θ^* are in $\mathcal{B}_0(k)$ so that $\hat\theta^{\text{LS}}_K - \theta^* \in \mathcal{B}_0(2k)$. For any $S \subset \{1,\dots,d\}$, let $\mathbb{X}_S$ denote the $n \times |S|$ submatrix of $\mathbb{X}$ that is obtained from the columns of $\mathbb{X}_j, j \in S$ of $\mathbb{X}$. Denote by $r_S \le |S|$ the rank of $\mathbb{X}_S$ and let $\Phi_S = [\phi_1,\dots,\phi_{r_S}] \in \mathbb{R}^{n\times r_S}$ be an orthonormal basis of the column span of $\mathbb{X}_S$. Moreover, for any $\theta \in \mathbb{R}^d$, define $\theta(S) \in \mathbb{R}^{|S|}$ to be the vector with coordinates $\theta_j, j \in S$. If we denote by $\hat S = \mathrm{supp}(\hat\theta^{\text{LS}}_K - \theta^*)$, we have $|\hat S| \le 2k$ and there exists $\nu \in \mathbb{R}^{r_{\hat S}}$ such that

$$\mathbb{X}(\hat\theta^{\text{LS}}_K - \theta^*) = \mathbb{X}_{\hat S}(\hat\theta^{\text{LS}}_K(\hat S) - \theta^*(\hat S)) = \Phi_{\hat S}\nu\,.$$

Therefore,

$$\frac{\varepsilon^\top\mathbb{X}(\hat\theta^{\text{LS}}_K - \theta^*)}{|\mathbb{X}(\hat\theta^{\text{LS}}_K - \theta^*)|_2} = \frac{\varepsilon^\top\Phi_{\hat S}\nu}{|\nu|_2} \le \max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}[\varepsilon^\top\Phi_S]u$$

where $\mathcal{B}_2^{r_S}$ is the unit ball of $\mathbb{R}^{r_S}$. It yields

$$|\mathbb{X}\hat\theta^{\text{LS}}_K - \mathbb{X}\theta^*|_2^2 \le 4\max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon_S^\top u)^2\,,$$

$\tilde\varepsilon_S = \Phi_S^\top\varepsilon \sim \mathsf{subG}_{r_S}(\sigma^2)$.

Using a union bound, we get for any $t > 0$,

$$\mathbb{P}\big(\max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon^\top u)^2 > t\big) \le \sum_{|S|=2k}\mathbb{P}\big(\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon^\top u)^2 > t\big)$$

It follows from the proof of Theorem 1.19 that for any $|S| \le 2k$,

$$\mathbb{P}\big(\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon^\top u)^2 > t\big) \le 6^{|S|}e^{-\frac{t}{8\sigma^2}} \le 6^{2k}e^{-\frac{t}{8\sigma^2}}\,.$$

Together, the above three displays yield

$$\mathbb{P}(|\mathbb{X}\hat{\theta}^{\text{LS}}_K - \mathbb{X}\theta^*|_2^2 > 4t) \le \binom{d}{2k} 6^{2k} e^{-\frac{t}{8\sigma^2}} \,. \tag{2.6}$$

To ensure that the right-hand side of the above inequality is bounded by δ, we need

$$t \ge C\sigma^2 \Big\{ \log \binom{d}{2k} + k\log(6) + \log(1/\delta) \Big\} \,.$$

□

How large is $\log \binom{d}{2k}$? It turns out that it is not much larger than k.

Lemma 2.7. *For any integers $1 \le k \le n$, it holds*

$$\binom{n}{k} \le \Big(\frac{en}{k}\Big)^k$$

Proof. Observe first that if $k = 1$, since $n \ge 1$, it holds,

$$\binom{n}{1} = n \le en = \Big(\frac{en}{1}\Big)^1$$

Next, we proceed by induction and assume that it holds for some $k \le n-1$.

$$\binom{n}{k} \le \Big(\frac{en}{k}\Big)^k$$

Observe that

$$\binom{n}{k+1} = \binom{n}{k}\frac{n-k}{k+1} \le \Big(\frac{en}{k}\Big)^k \frac{n-k}{k+1} = \frac{e^k n^{k+1}}{(k+1)^{k+1}}\Big(1+\frac{1}{k}\Big)^k \,,$$

where we used the induction hypothesis in the first inequality. To conclude, it suffices to observe that

$$\Big(1+\frac{1}{k}\Big)^k \le e$$

□

It immediately leads to the following corollary:

Corollary 2.8. *Under the assumptions of Theorem 2.6, for any $\delta > 0$, with probability $1-\delta$, it holds*

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2 k}{n}\log\Big(\frac{ed}{2k}\Big) + \frac{\sigma^2 k}{n}\log(6) + \frac{\sigma^2}{n}\log(1/\delta) \,.$$

Note that for any fixed δ, there exits a constant $C_\delta > 0$ such that for any $n \geq 2k$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)}) \leq C_\delta \frac{\sigma^2 k}{n} \log\Big(\frac{ed}{2k}\Big)\,.$$

Comparing this result with Theorem 2.2 with $r = k$, we see that the price to pay for not knowing the support of θ^* but only its size, is a logarithmic factor in the dimension d.

This result immediately leads the following bound in expectation.

Corollary 2.9. *Under the assumptions of Theorem 2.6,*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)})\big] \lesssim \frac{\sigma^2 k}{n} \log\Big(\frac{ed}{k}\Big)\,.$$

Proof. It follows from (2.6) that for any $H \geq 0$,

$$\begin{aligned}
\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)})\big] &= \int_0^\infty \mathbb{P}(|\mathbb{X}\hat{\theta}^{\text{LS}}_{K} - \mathbb{X}\theta^*|_2^2 > nu)\mathrm{d}u\\
&\leq H + \int_0^\infty \mathbb{P}(|\mathbb{X}\hat{\theta}^{\text{LS}}_{K} - \mathbb{X}\theta^*|_2^2 > n(u+H))\mathrm{d}u\\
&\leq H + \sum_{j=1}^{2k}\binom{d}{j}6^{2k}\int_0^\infty e^{-\frac{n(u+H)}{32\sigma^2}}\,,\\
&= H + \sum_{j=1}^{2k}\binom{d}{j}6^{2k}e^{-\frac{nH}{32\sigma^2}}\frac{32\sigma^2}{n}\mathrm{d}u\,.
\end{aligned}$$

Next, take H to be such that

$$\sum_{j=1}^{2k}\binom{d}{j}6^{2k}e^{-\frac{nH}{32\sigma^2}} = 1\,.$$

In particular, it yields

$$H \lesssim \frac{\sigma^2 k}{n}\log\Big(\frac{ed}{k}\Big)\,,$$

which completes the proof. □

2.3 THE GAUSSIAN SEQUENCE MODEL

The Gaussian Sequence Model is a toy model that has received a lot of attention, mostly in the eighties. The main reason for its popularity is that it carries already most of the insight of nonparametric estimation. While the model looks very simple it allows to carry deep ideas that extend beyond its framework and in particular to the linear regression model that we are interested in. Unfortunately we will only cover a small part of these ideas and the interested reader should definitely look at the excellent books by A. Tsybakov [Tsy09, Chapter 3] and I. Johnstone [Joh11]. The model is as follows:

$$Y_i = \theta^*_i + \varepsilon_i\,, \qquad i = 1, \ldots, d \tag{2.7}$$

where $\varepsilon_1, \ldots, \varepsilon_d$ are i.i.d $\mathcal{N}(0, \sigma^2)$ random variables. Note that often, d is taken equal to ∞ in this sequence model and we will also discuss this case. Its links to nonparametric estimation will become clearer in Chapter 3. The goal here is to estimate the unknown vector θ^*.

The sub-Gaussian Sequence Model

Note first that the model (2.7) is a special case of the linear model with fixed design (2.1) with $n = d$, $f(x) = x^\top \theta^*$, $x_1, \ldots, x_n$ form the canonical basis of $\mathbb{R}^n$ and ε has a Gaussian distribution. Therefore, $n = d$ is both the dimension of the parameter θ and the number of observation and it looks like we have chosen to index this problem by d rather than n somewhat arbitrarily. We can bring n back into the picture, by observing that this model encompasses slightly more general choices for the design matrix $\mathbb{X}$ as long as it satisfies the following assumption.

Assumption ORT The design matrix satisfies

$$\frac{\mathbb{X}^\top \mathbb{X}}{n} = I_d\,,$$

where I_d denotes the identity matrix of $\mathbb{R}^d$.

Assumption ORT allows for cases where $d \le n$ but not $d > n$ (high dimensional case) because of obvious rank constraints. In particular, it means that the d columns of $\mathbb{X}$ are orthogonal in $\mathbb{R}^n$ and all have norm $\sqrt{n}$.

Under this assumption, it follows from the linear regression model (2.2) that

$$\begin{aligned} y := \frac{1}{n}\mathbb{X}^\top Y &= \frac{\mathbb{X}^\top \mathbb{X}}{n}\theta^* + \frac{1}{n}\mathbb{X}^\top \varepsilon \\ &= \theta^* + \xi\,, \end{aligned}$$

where $\xi = (\xi_1, \ldots, \xi_d) \sim \mathsf{subG}_d(\sigma^2/n)$. As a result, under the assumption ORT, the linear regression model (2.2) is equivalent to the sub-Gaussian Sequence Model (2.7) up to a transformation of the data Y and a change of variable for the variance. Moreover, for any estimator $\hat\theta \in \mathbb{R}^d$, under ORT, it follows from (2.3) that

$$\mathsf{MSE}(\mathbb{X}\hat\theta) = (\hat\theta - \theta^*)^\top \frac{\mathbb{X}^\top \mathbb{X}}{n}(\hat\theta - \theta^*) = |\hat\theta - \theta^*|_2^2\,.$$

Furthermore, for any $\theta \in \mathbb{R}^d$, the assumption ORT yields,

$$\begin{aligned}
|y-\theta|_2^2 &= |\frac{1}{n}\mathbb{X}^\top Y - \theta|_2^2 \\
&= |\theta|_2^2 - \frac{2}{n}\theta^\top \mathbb{X}^\top Y + \frac{1}{n^2} Y^\top \mathbb{X}\mathbb{X}^\top Y \\
&= \frac{1}{n}|\mathbb{X}\theta|_2^2 - \frac{2}{n}(\mathbb{X}\theta)^\top Y + \frac{1}{n}|Y|_2^2 + Q \\
&= \frac{1}{n}|Y - \mathbb{X}\theta|_2^2 + Q\,, \qquad (2.8)
\end{aligned}$$

where Q is a constant that does not depend on θ and is defined by

$$Q = \frac{1}{n^2} Y^\top \mathbb{X}\mathbb{X}^\top Y - \frac{1}{n}|Y|_2^2$$

This implies in particular that the least squares estimator $\hat\theta^{\text{LS}}$ is equal to y.

We introduce a sightly more general model called *sub-Gaussian sequence model*:

$$y = \theta^* + \xi \quad \in \mathbb{R}^d \qquad (2.9)$$

where $\xi \sim \mathsf{subG}_d(\sigma^2/n)$.

In this section, we can actually completely "forget" about our original model (2.2). In particular we can define this model independently of Assumption ORT and thus for any values of n and d.

The sub-Gaussian sequence model, like the Gaussian sequence model are called *direct* (observation) problems as opposed to *inverse problems* where the goal is to estimate the parameter θ^* only from noisy observations of its image through an operator. The linear regression model one such inverse problem where the matrix $\mathbb{X}$ plays the role of a linear operator. However, in these notes, we never try to invert the operator. See [Cav11] for an interesting survey on the statistical theory of inverse problems.

Sparsity adaptive thresholding estimators

If we knew a priori that θ was k sparse, we could employ directly Corollary 2.8 to obtain that with probability $1-\delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}_{\mathcal{B}_0(k)}) \le C_\delta \frac{\sigma^2 k}{n}\log\Big(\frac{ed}{2k}\Big)\,.$$

As we will see, the assumption ORT gives us the luxury to not know k and yet *adapt* to its value. Adaptation means that we can construct an estimator that does not require the knowledge of k (the smallest such that $|\theta^*|_0 \le k$) and yet, perform as well as $\hat\theta^{\text{LS}}_{\mathcal{B}_0(k)}$, up to a multiplicative constant.

Let us begin with some heuristic considerations to gain some intuition. Assume the sub-Gaussian sequence model (2.9). If nothing is known about θ^*

it is natural to estimate it using the least squares estimator $\hat\theta^{\text{LS}} = y$. In this case,

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{LS}}) = |y-\theta^*|_2^2 = |\xi|_2^2 \le C_\delta \frac{\sigma^2 d}{n}\,,$$

where the last inequality holds with probability at least $1-\delta$. This is actually what we are looking for if $k = Cd$ for some positive constant $C \le 1$. The problem with this approach is that it does not use the fact that k may be much smaller than d, which happens when θ^* has many zero coordinate.

If $\theta^*_j = 0$, then, $y_j = \xi_j$, which is a sub-Gaussian random variable with variance proxy σ^2/n. In particular, we know from Lemma 1.3 that with probability at least $1-\delta$,

$$|\xi_j| \le \sigma\sqrt{\frac{2\log(2/\delta)}{n}} = \tau\,. \tag{2.10}$$

The consequences of this inequality are interesting. One the one hand, if we observe $|y_j| \gg \tau$, then it must correspond to $\theta^*_j \neq 0$. On the other hand, if $|y_j| \le \tau$ is smaller, then, θ^*_j cannot be very large. In particular, by the triangle inequality, $|\theta^*_j| \le |y_j| + |\xi_j| \le 2\tau$. Therefore, we loose at most 2τ by choosing $\hat\theta_j = 0$. It leads us to consider the following estimator.

Definition 2.10. The **hard thresholding** estimator with threshold $2\tau > 0$ is denoted by $\hat\theta^{\text{HRD}}$ and has coordinates

$$\hat\theta^{\text{HRD}}_j = \left\{\begin{array}{ll} y_j & \text{if } |y_j| > 2\tau\,, \\ 0 & \text{if } |y_j| \le 2\tau\,, \end{array}\right.$$

for $j = 1, \dots, d$. In short, we can write $\hat\theta^{\text{HRD}}_j = y_j \mathbb{1}(|y_j| > 2\tau)$.

From our above consideration, we are tempted to choose τ as in (2.10). Yet, this threshold is not large enough. Indeed, we need to choose τ such that $|\xi_j| \le \tau$ *simultaneously* for all j. This can be done using a maximal inequality. Namely, Theorem 1.14 ensures that with probability at least $1-\delta$,

$$\max_{1\le j\le d} |\xi_j| \le \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$$

It yields the following theorem.

Theorem 2.11. *Consider the linear regression model (2.2) under the assumption* $\textsf{\textit{ORT}}$ *or, equivalenty, the sub-Gaussian sequence model (2.9). Then the hard thresholding estimator* $\hat\theta^{\text{HRD}}$ *with threshold*

$$2\tau = 2\sigma\sqrt{\frac{2\log(2d/\delta)}{n}}\,, \tag{2.11}$$

enjoys the following two properties on the same event $\mathcal{A}$ *such that* $\mathbb{P}(\mathcal{A}) \ge 1-\delta$*:*

(i) If $|\theta^*|_0 = k$,

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\text{HRD}}) = |\hat\theta^{\text{HRD}} - \theta^*|_2^2 \lesssim \sigma^2 \frac{k\log(2d/\delta)}{n}\,.$$

(ii) if $\min_{j\in\text{supp}(\theta^*)}|\theta^*_j| > 3\tau$, *then*

$$\text{supp}(\hat\theta^{\text{HRD}}) = \text{supp}(\theta^*)\,.$$

Proof. Define the event

$$\mathcal{A} = \Big\{\max_j |\xi_j| \le \tau\Big\}\,,$$

and recall that Theorem 1.14 yields $\mathbb{P}(\mathcal{A}) \ge 1-\delta$. On the event $\mathcal{A}$, the following holds for any $j = 1,\dots,d$.

First, observe that

$$|y_j| > 2\tau \quad\Rightarrow\quad |\theta^*_j| \ge |y_j| - |\xi_j| > \tau \tag{2.12}$$

and

$$|y_j| \le 2\tau \quad\Rightarrow\quad |\theta^*_j| \le |y_j| + |\xi_j| \le 3\tau \tag{2.13}$$

It yields

$$\begin{aligned}|\hat\theta^{\text{HRD}}_j - \theta^*_j| &= |y_j - \theta^*_j|\mathbb{1}(|y_j| > 2\tau) + |\theta^*_j|\mathbb{1}(|y_j| \le 2\tau)\\ &\le \tau\mathbb{1}(|y_j| > 2\tau) + |\theta^*_j|\mathbb{1}(|y_j| \le 2\tau)\\ &\le \tau\mathbb{1}(|\theta^*_j| > \tau) + |\theta^*_j|\mathbb{1}(|\theta^*_j| \le 3\tau) && \text{by (2.12) and (2.13)}\\ &\le 4\min(|\theta^*_j|, \tau)\end{aligned}$$

It yields

$$|\hat\theta^{\text{HRD}} - \theta^*|_2^2 = \sum_{j=1}^d |\hat\theta^{\text{HRD}}_j - \theta^*_j|^2 \le 16\sum_{j=1}^d \min(|\theta^*_j|^2, \tau^2) \le 16|\theta^*|_0\tau^2\,.$$

This completes the proof of (i).

To prove (ii), note that if $\theta^*_j \neq 0$, then $|\theta^*_j| > 3\tau$ so that

$$|y_j| = |\theta^*_j + \xi_j| > 3\tau - \tau = 2\tau\,.$$

Therefore, $\hat\theta^{\text{HRD}}_j \neq 0$ so that $\text{supp}(\theta^*) \subset \text{supp}(\hat\theta^{\text{HRD}})$.

Next, if $\hat\theta^{\text{HRD}}_j \neq 0$, then $|\hat\theta^{\text{HRD}}_j| = |y_j| > 2\tau$. It yields

$$|\theta^*_j| \ge |y_j| - \tau > \tau$$

Therefore, $|\theta^*_j| \neq 0$ and $\text{supp}(\hat\theta^{\text{HRD}}) \subset \text{supp}(\theta^*)$. □

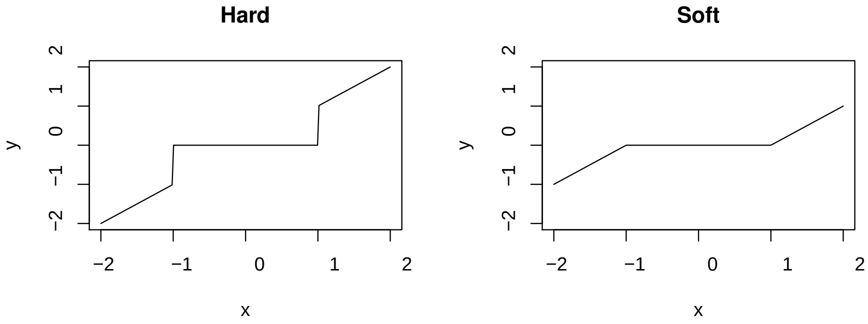

Figure 2.2. Transformation applied to y_j with $2\tau = 1$ to obtain the hard (left) and soft (right) thresholding estimators

Similar results can be obtained for the **soft thresholding** estimator $\hat{\theta}^{\text{SFT}}$ defined by

$$\hat{\theta}^{\text{SFT}}_j = \begin{cases} y_j - 2\tau & \text{if } y_j > 2\tau\,, \\ y_j + 2\tau & \text{if } y_j < -2\tau\,, \\ 0 & \text{if } |y_j| \le 2\tau\,, \end{cases}$$

In short, we can write

$$\hat{\theta}^{\text{SFT}}_j = \Big(1 - \frac{2\tau}{|y_j|}\Big)_+ y_j$$

2.4 HIGH-DIMENSIONAL LINEAR REGRESSION

The BIC and Lasso estimators

It can be shown (see Problem 2.5) that the hard and soft thresholding estimators are solutions of the following penalized empirical risk minimization problems:

$$\hat{\theta}^{\text{HRD}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \Big\{ |y - \theta|_2^2 + 4\tau^2 |\theta|_0 \Big\}$$

$$\hat{\theta}^{\text{SFT}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \Big\{ |y - \theta|_2^2 + 4\tau |\theta|_1 \Big\}$$

In view of (2.8), under the assumption ORT, the above variational definitions can be written as

$$\hat\theta^{\text{HRD}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + 4\tau^2|\theta|_0\Big\}$$

$$\hat\theta^{\text{SFT}} = \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + 4\tau|\theta|_1\Big\}$$

When the assumption ORT is not satisfied, they no longer correspond to thresholding estimators but can still be defined as above. We change the constant in the threshold parameters for future convenience.

Definition 2.12. Fix $\tau > 0$ and assume the linear regression model (2.2). The BIC[2] estimator of θ^* in is defined by any $\hat\theta^{\text{BIC}}$ such that

$$\hat\theta^{\text{BIC}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + \tau^2|\theta|_0\Big\}$$

Moreover the Lasso estimator of θ^* in is defined by any $\hat\theta^{\mathcal{L}}$ such that

$$\hat\theta^{\mathcal{L}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + 2\tau|\theta|_1\Big\}$$

Remark 2.13. Numerical considerations. Computing the BIC estimator can be proved to be NP-hard in the worst case. In particular, no computational method is known to be significantly faster than the brute force search among all 2^d sparsity patterns. Indeed, we can rewrite:

$$\min_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + \tau^2|\theta|_0\Big\} = \min_{0\le k\le d}\Big\{\min_{\theta\,:\,|\theta|_0=k}\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + \tau^2 k\Big\}$$

To compute $\min_{\theta\,:\,|\theta|_0=k}\frac{1}{n}|Y-\mathbb{X}\theta|_2^2$, we need to compute $\binom{d}{k}$ least squares estimators on a space of size k. Each costs $O(k^3)$ (matrix inversion). Therefore the total cost of the brute force search is

$$C\sum_{k=0}^{d}\binom{d}{k}k^3 = Cd^3 2^d\,.$$

Instead the the Lasso estimator is convex problem and there exists many efficient algorithms to compute it. We will not describe this optimization problem in details but only highlight a few of the best known algorithms:

1. Probably the most popular method among statisticians relies on coordinate gradient descent. It is implemented in the glmnet package in R [FHT10],

[2] Note that it minimizes the Bayes Information Criterion (BIC) employed in the traditional literature of asymptotic statistics if $\tau = \sqrt{\log(d)/n}$. We will use the same value below, up to multiplicative constants (it's the price to pay to get non asymptotic results).

2. An interesting method called LARS [EHJT04] computes the entire *regularization path*, i.e., the solution of the convex problem for all values of τ. It relies on the fact that, as a function of τ, the solution $\hat{\theta}^{\mathcal{L}}$ is a piecewise linear function (with values in $\mathbb{R}^d$). Yet this method proved to be too slow for very large problems and has been replaced by `glmnet` which computes solutions for values of τ on a grid much faster.

3. The optimization community has made interesting contribution to this field by using proximal methods to solve this problem. It exploits the structure of the form: smooth (sum of squares) + simple (ℓ_1 norm). A good entry point to this literature is perhaps the FISTA algorithm [BT09].

4. There has been recently a lot of interest around this objective for very large d and very large n. In this case, even computing $|Y - \mathbb{X}\theta|_2^2$ may be computationally expensive and solutions based on stochastic gradient descent are flourishing.

Note that by Lagrange duality computing $\hat{\theta}^{\mathcal{L}}$ is equivalent to solving an ℓ_1 *constrained* least squares. Nevertheless, the radius of the ℓ_1 constraint is unknown. In general it is hard to relate Lagrange multipliers to the size constraints. The name "Lasso" was given to the constrained version this estimator in the original paper of Robert Tibshirani [Tib96].

Analysis of the BIC estimator

While computationally hard to implement, the BIC estimator gives us a good benchmark for sparse estimation. Its performance is similar to that of $\hat{\theta}^{\text{HRD}}$ but without assumption ORT.

Theorem 2.14. *Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Then, the BIC estimator $\hat{\theta}^{\text{BIC}}$ with regularization parameter*

$$\tau^2 = 16\log(6)\frac{\sigma^2}{n} + 32\frac{\sigma^2\log(ed)}{n}\,. \tag{2.14}$$

satisfies

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{BIC}}) = \frac{1}{n}|\mathbb{X}\hat{\theta}^{\text{BIC}} - \mathbb{X}\theta^*|_2^2 \lesssim |\theta^*|_0\sigma^2\frac{\log(ed/\delta)}{n}$$

with probability at least $1-\delta$.

Proof. We begin as usual by noting that

$$\frac{1}{n}|Y - \mathbb{X}\hat{\theta}^{\text{BIC}}|_2^2 + \tau^2|\hat{\theta}^{\text{BIC}}|_0 \le \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + \tau^2|\theta^*|_0\,.$$

It implies

$$|\mathbb{X}\hat{\theta}^{\text{BIC}} - \mathbb{X}\theta^*|_2^2 \le n\tau^2|\theta^*|_0 + 2\varepsilon^\top\mathbb{X}(\hat{\theta}^{\text{BIC}} - \theta^*) - n\tau^2|\hat{\theta}^{\text{BIC}}|_0\,.$$

First, note that

$$\begin{aligned}2\varepsilon^\top \mathbb{X}(\hat\theta^{\text{BIC}}-\theta^*) &= 2\varepsilon^\top\Big(\frac{\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*}{|\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*|_2}\Big)|\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*|_2\\ &\le 2\Big[\varepsilon^\top\Big(\frac{\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*}{|\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*|_2}\Big)\Big]^2+\frac{1}{2}|\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*|_2^2\,,\end{aligned}$$

where we use the inequality $2ab \le 2a^2+\frac{1}{2}b^2$. Together with the previous display, it yields

$$|\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*|_2^2 \le 2n\tau^2|\theta^*|_0 + 4\big[\varepsilon^\top\mathcal{U}(\hat\theta^{\text{BIC}}-\theta^*)\big]^2 - 2n\tau^2|\hat\theta^{\text{BIC}}|_0 \qquad (2.15)$$

where

$$\mathcal{U}(\hat\theta^{\text{BIC}}-\theta^*) = \frac{\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*}{|\mathbb{X}\hat\theta^{\text{BIC}}-\mathbb{X}\theta^*|_2}$$

Next, we need to "sup out" $\hat\theta^{\text{BIC}}$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta\in\mathbb{R}^d} = \max_{1\le k\le d}\max_{|S|=k}\sup_{\text{supp}(\theta)=S}\,.$$

Applied to the above inequality, it yields

$$\begin{aligned}4\big[&\varepsilon^\top\mathcal{U}(\hat\theta^{\text{BIC}}-\theta^*)\big]^2 - 2n\tau^2|\hat\theta^{\text{BIC}}|_0\\ &\le \max_{1\le k\le d}\Big\{\max_{|S|=k}\sup_{\text{supp}(\theta)=S} 4\big[\varepsilon^\top\mathcal{U}(\theta-\theta^*)\big]^2-2n\tau^2k\Big\}\\ &\le \max_{1\le k\le d}\Big\{\max_{|S|=k}\sup_{u\in\mathcal{B}_2^{r_{S,*}}} 4\big[\varepsilon^\top\Phi_{S,*}u\big]^2-2n\tau^2k\Big\}\,,\end{aligned}$$

where $\Phi_{S,*}=[\phi_1,\dots,\phi_{r_{S,*}}]$ is an orthonormal basis of the set $\{\mathbb{X}_j, j\in S\cup \text{supp}(\theta^*)\}$ of columns of $\mathbb{X}$ and $r_{S,*}\le |S|+|\theta^*|_0$ is the dimension of this column span.

Using union bounds, we get for any $t>0$,

$$\begin{aligned}\mathbb{P}\Big(&\max_{1\le k\le d}\Big\{\max_{|S|=k}\sup_{u\in\mathcal{B}_2^{r_{S,*}}} 4\big[\varepsilon^\top\Phi_{S,*}u\big]^2-2n\tau^2k\Big\}\ge t)\\ &\le \sum_{k=1}^d\sum_{|S|=k}\mathbb{P}\Big(\sup_{u\in\mathcal{B}_2^{r_{S,*}}}\big[\varepsilon^\top\Phi_{S,*}u\big]^2\ge\frac{t}{4}+\frac{1}{2}n\tau^2k\Big)\end{aligned}$$

Moreover, using the ε-net argument from Theorem 1.19, we get for $|S|=k$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{u\in\mathcal{B}_2^{r_{S,*}}}\big[\varepsilon^\top\Phi_{S,*}u\big]^2\ge\frac{t}{4}+\frac{1}{2}n\tau^2k\Big) &\le 2\cdot 6^{r_{S,*}}\exp\Big(-\frac{\frac{t}{4}+\frac{1}{2}n\tau^2k}{8\sigma^2}\Big)\\ &\le 2\exp\Big(-\frac{t}{32\sigma^2}-\frac{n\tau^2k}{16\sigma^2}+(k+|\theta^*|_0)\log(6)\Big)\\ &\le \exp\Big(-\frac{t}{32\sigma^2}-2k\log(ed)+|\theta^*|_0\log(12)\Big)\end{aligned}$$

where, in the last inequality, we used the definition (2.14) of τ.

Putting everything together, we get

$$\begin{aligned}
&\mathbb{P}\Big(|\mathbb{X}\hat\theta^{\text{BIC}} - \mathbb{X}\theta^*|_2^2 \ge 2n\tau^2|\theta^*|_0 + t\Big) \le \\
&\quad \sum_{k=1}^{d}\sum_{|S|=k} \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0\log(12)\Big)\\
&\quad = \sum_{k=1}^{d}\binom{d}{k} \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0\log(12)\Big)\\
&\quad \le \sum_{k=1}^{d} \exp\Big(-\frac{t}{32\sigma^2} - k\log(ed) + |\theta^*|_0\log(12)\Big) \qquad \text{by Lemma 2.7}\\
&\quad = \sum_{k=1}^{d} (ed)^{-k} \exp\Big(-\frac{t}{32\sigma^2} + |\theta^*|_0\log(12)\Big)\\
&\quad \le \exp\Big(-\frac{t}{32\sigma^2} + |\theta^*|_0\log(12)\Big)\,.
\end{aligned}$$

To conclude the proof, choose $t = 32\sigma^2|\theta^*|_0\log(12) + 32\sigma^2\log(1/\delta)$ and observe that combined with (2.15), it yields with probability $1-\delta$,

$$\begin{aligned}
|\mathbb{X}\hat\theta^{\text{BIC}} - \mathbb{X}\theta^*|_2^2 &\le 2n\tau^2|\theta^*|_0 + t\\
&= 64\sigma^2\log(ed)|\theta^*|_0 + 64\log(12)\sigma^2|\theta^*|_0 + 32\sigma^2\log(1/\delta)\\
&\le 224|\theta^*|_0\sigma^2\log(ed) + 32\sigma^2\log(1/\delta)\,.
\end{aligned}$$

□

It follows from Theorem 2.14 that $\hat\theta^{\text{BIC}}$ *adapts* to the unknown sparsity of θ^*, just like $\hat\theta^{\text{HRD}}$. Moreover, this holds under no assumption on the design matrix $\mathbb{X}$.

Analysis of the Lasso estimator

Slow rate for the Lasso estimator

The properties of the BIC estimator are quite impressive. It shows that under no assumption on $\mathbb{X}$, one can mimic two oracles: (i) the oracle that knows the support of θ^* (and computes least squares on this support), up to a $\log(ed)$ term and (ii) the oracle that knows the sparsity $|\theta^*|_0$ of θ^*, up to a smaller logarithmic term $\log(ed/|\theta^*|_0)$ is replaced by $\log(ed)$. Actually the latter can even be removed by using a modified BIC estimator (see Problem 2.6).

The Lasso estimator is a bit more difficult because, by construction, it should more naturally adapt to the unknown ℓ_1-norm of θ^*. This can be easily shown as in the next theorem, analogous to Theorem 2.4.

Theorem 2.15. *Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume that the columns of $\mathbb{X}$ are normalized in such a way that $\max_j |\mathbb{X}_j|_2 \le \sqrt{n}$. Then, the Lasso estimator $\hat\theta^{\mathcal{L}}$ with regularization parameter*

$$2\tau = 2\sigma\sqrt{\frac{2\log(2d)}{n}} + 2\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,. \tag{2.16}$$

satisfies

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathcal{L}}) = \frac{1}{n}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \le 4|\theta^*|_1\sigma\sqrt{\frac{2\log(2d)}{n}} + 4|\theta^*|_1\sigma\sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1-\delta$. Moreover, there exists a numerical constant $C>0$ such that

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathcal{L}})\big] \le C|\theta^*|_1\sigma\sqrt{\frac{\log(2d)}{n}}\,.$$

Proof. From the definition of $\hat\theta^{\mathcal{L}}$, it holds

$$\frac{1}{n}|Y - \mathbb{X}\hat\theta^{\mathcal{L}}|_2^2 + 2\tau|\hat\theta^{\mathcal{L}}|_1 \le \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + 2\tau|\theta^*|_1\,.$$

Using Hölder's inequality, it implies

$$\begin{aligned}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 &\le 2\varepsilon^\top\mathbb{X}(\hat\theta^{\mathcal{L}} - \theta^*) + 2n\tau\big(|\theta^*|_1 - |\hat\theta^{\mathcal{L}}|_1\big)\\ &\le 2|\mathbb{X}^\top\varepsilon|_\infty|\hat\theta^{\mathcal{L}}|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1 + 2|\mathbb{X}^\top\varepsilon|_\infty|\theta^*|_1 + 2n\tau|\theta^*|_1\\ &= 2(|\mathbb{X}^\top\varepsilon|_\infty - n\tau)|\hat\theta^{\mathcal{L}}|_1 + 2(|\mathbb{X}^\top\varepsilon|_\infty + n\tau)|\theta^*|_1\end{aligned}$$

Observe now that for any $t>0$,

$$\mathbb{P}(|\mathbb{X}^\top\varepsilon|_\infty \ge t) = \mathbb{P}(\max_{1\le j\le d}|\mathbb{X}_j^\top\varepsilon| > t) \le 2de^{-\frac{t^2}{2n\sigma^2}}$$

Therefore, taking $t = \sigma\sqrt{2n\log(2d)} + \sigma\sqrt{2n\log(1/\delta)} = n\tau$, we get that with probability $1-\delta$,

$$|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \le 4n\tau|\theta^*|_1\,.$$

The bound in expectation follows using the same argument as in the proof of Corollary 2.9. □

Notice that the regularization parameter (2.16) depends on the confidence level δ. This not the case for the BIC estimator (see (2.14)).

The rate in Theorem 2.15 if of order $\sqrt{(\log d)/n}$ (**slow rate**), which is much slower than the rate of order $(\log d)/n$ (**fast rate**) for the BIC estimator. Hereafter, we show that fast rates can be achieved by the computationally efficient Lasso estimator but at the cost of a much stronger condition on the design matrix $\mathbb{X}$.

Incoherence

Assumption $\mathsf{INC}(k)$ We say that the design matrix $\mathbb{X}$ has incoherence k for some integer $k > 0$ if

$$\big|\frac{\mathbb{X}^\top \mathbb{X}}{n} - I_d\big|_\infty \le \frac{1}{14k}$$

where the $|A|_\infty$ denotes the largest element of A in absolute value. Equivalently,

1. For all $j = 1, \dots, d$,
$$\big|\frac{|\mathbb{X}_j|_2^2}{n} - 1\big| \le \frac{1}{14k}\,.$$

2. For all $1 \le i, j \le d$, $i \ne j$, we have
$$\big|\mathbb{X}_i^\top \mathbb{X}_j\big| \le \frac{1}{14k}\,.$$

Note that Assumption ORT arises as the limiting case of $\mathsf{INC}(k)$ as $k \to \infty$. However, while Assumption ORT requires $d \le n$, here we may have $d \gg n$ as illustrated in Proposition 2.16 below. To that end, we simply have to show that there exists a matrix that satisfies $\mathsf{INC}(k)$ even for $d > n$. We resort to the *probabilistic method* [AS08]. The idea of this method is that if we can find a probability measure that puts a positive probability of objects that satistify a certain property, then there must exist objects that satisfy said property. In our case, we consider the following probability distribution on random matrices with entries in $\{\pm 1\}$. Let the design matrix $\mathbb{X}$ have entries that are i.i.d Rademacher (± 1) random variables. We are going to show that most realizations of this random matrix satisfy Assumption $\mathsf{INC}(k)$ for large enough n.

Proposition 2.16. Let $\mathbb{X} \in \mathbb{R}^{n \times d}$ be a random matrix with entries $X_{ij}, i = 1, \dots, n, j = 1, \dots, d$ that are i.i.d Rademacher (± 1) random variables. Then, $\mathbb{X}$ has incoherence k with probability $1 - \delta$ as soon as

$$n \ge 392 k^2 \log(1/\delta) + 784 k^2 \log(d)\,.$$

It implies that there exists matrices that satisfy Assumption $\mathsf{INC}(k)$ for

$$n \gtrsim k^2 \log(d)\,,$$

for some numerical constant C.

Proof. Let $\varepsilon_{ij} \in \{-1, 1\}$ denote the Rademacher random variable that is on the ith row and jth column of $\mathbb{X}$.

Note first that the jth diagonal entries of $\mathbb{X}^\top \mathbb{X}/n$ is given by

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,j}^2 = 1$$

Moreover, for $j \neq k$, the (j,k)th entry of the $d \times d$ matrix $\frac{\mathbb{X}^\top \mathbb{X}}{n}$ is given by

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,j}\varepsilon_{i,k} = \frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)},$$

where for each pair, (j,k), $\xi_i^{(j,k)} = \varepsilon_{i,j}\varepsilon_{i,k}$ so that the random variables $\xi_1^{(j,k)}, \dots, \xi_n^{(j,k)}$ are iid Rademacher random variables.

Therefore, we get that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\big(\big|\frac{\mathbb{X}^\top \mathbb{X}}{n} - I_d\big|_\infty > t\big) &= \mathbb{P}\Big(\max_{j\neq k}\Big|\frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)}\Big| > t\Big)\\
&\le \sum_{j\neq k}\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)}\Big| > t\Big) && \text{(Union bound)}\\
&\le \sum_{j\neq k} 2e^{-\frac{nt^2}{2}} && \text{(Hoeffding: Theorem 1.9)}\\
&\le d^2 e^{-\frac{nt^2}{2}}
\end{aligned}$$

Taking now $t = 1/(14k)$ yields

$$\mathbb{P}\big(\big|\frac{\mathbb{X}^\top \mathbb{X}}{n} - I_d\big|_\infty > \frac{1}{14k}\big) \le d^2 e^{-\frac{n}{392k^2}} \le \delta$$

for

$$n \ge 392k^2\log(1/\delta) + 784k^2\log(d).$$

□

For any $\theta \in \mathbb{R}^d$, $S \subset \{1,\dots,d\}$ define θ_S to be the vector with coordinates

$$\theta_{S,j} = \begin{cases} \theta_j & \text{if } j \in S\,,\\ 0 & \text{otherwise}\,.\end{cases}$$

In particular $|\theta|_1 = |\theta_S|_1 + |\theta_{S^c}|_1$.

The following lemma holds

Lemma 2.17. *Fix a positive integer $k \le d$ and assume that $\mathbb{X}$ satisfies assumption* **INC**(k). *Then, for any $S \in \{1,\dots,d\}$ such that $|S| \le k$ and any $\theta \in \mathbb{R}^d$ that satisfies the* cone condition

$$|\theta_{S^c}|_1 \le 3|\theta_S|_1\,, \tag{2.17}$$

it holds

$$|\theta_S|_2^2 \le 2\frac{|\mathbb{X}\theta|_2^2}{n}$$

Proof. We have

$$\frac{|\mathbb{X}\theta|_2^2}{n} = \frac{1}{n}|\mathbb{X}\theta_S + \mathbb{X}\theta_{S^c}|_2^2 \ge \frac{|\mathbb{X}\theta_S|_2^2}{n} + 2\theta_S^\top \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta_{S^c}$$

If follows now from the incoherence condition that

$$\frac{|\mathbb{X}\theta_S|_2^2}{n} = \theta_S^\top \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta_S = |\theta_S|_2^2 + \theta_S^\top(\frac{\mathbb{X}^\top\mathbb{X}}{n} - I_d)\theta_S \ge |\theta_S|_2^2 - \frac{|\theta_S|_1^2}{14k}$$

and

$$\Big|\theta_S^\top \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta_{S^c}\Big| \le \frac{1}{14k}|\theta_S|_1|\theta_{S^c}|_1 \le \frac{3}{14k}|\theta_S|_1^2$$

Observe now that it follows from the Cauchy-Schwarz inequality that

$$|\theta_S|_1^2 \le |S||\theta_S|_2^2$$

Thus for $|S| \le k$,

$$\frac{|\mathbb{X}\theta|_2^2}{n} \ge \Big(1 - \frac{7|S|}{14k}\Big)|\theta_S|_2^2 \ge \frac{1}{2}|\theta_S|_2^2$$

□

Fast rate for the Lasso

Theorem 2.18. *Fix $n \ge 2$. Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume that $|\theta^*|_0 \le k$ and that $\mathbb{X}$ satisfies assumption $\mathsf{INC}(k)$. Then the Lasso estimator $\hat\theta^{\mathcal{L}}$ with regularization parameter defined by*

$$2\tau = 8\sigma\sqrt{\frac{\log(2d)}{n}} + 8\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

satisfies

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathcal{L}}) = \frac{1}{n}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}$$

and

$$|\hat\theta^{\mathcal{L}} - \theta^*|_1 \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}\,.$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathcal{L}})\big] \lesssim k\sigma^2\frac{\log(2d)}{n}\,, \quad \textit{and} \quad \mathbb{E}\big[|\hat\theta^{\mathcal{L}} - \theta^*|_1\big] \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}\,.$$

Proof. From the definition of $\hat\theta^{\mathcal{L}}$, it holds

$$\frac{1}{n}|Y - \mathbb{X}\hat\theta^{\mathcal{L}}|_2^2 \le \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + 2\tau|\theta^*|_1 - 2\tau|\hat\theta^{\mathcal{L}}|_1\,.$$

Adding $\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1$ on each side and multiplying by n, we get

$$|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 + n\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1 \le 2\varepsilon^\top\mathbb{X}(\hat\theta^{\mathcal{L}} - \theta^*) + n\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1\,.$$

Applying Hölder's inequality and using the same steps as in the proof of Theorem 2.15, we get that with probability $1-\delta$, we get

$$\begin{aligned}\varepsilon^\top \mathbb{X}(\hat\theta^{\mathcal{L}} - \theta^*) &\le |\varepsilon^\top \mathbb{X}|_\infty |\hat\theta^{\mathcal{L}} - \theta^*| \\ &\le \frac{n\tau}{2}|\hat\theta^{\mathcal{L}} - \theta^*|_1\,,\end{aligned}$$

where we used the fact that $|\mathbb{X}_j|_2^2 \le n + 1/(14k) \le 2n$. Therefore, taking $S = \mathrm{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 + n\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1 &\le 2n\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1 \\ &= 2n\tau|\hat\theta^{\mathcal{L}}_S - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta^{\mathcal{L}}_S|_1 \\ &\le 4n\tau|\hat\theta^{\mathcal{L}}_S - \theta^*|_1 \qquad (2.18)\end{aligned}$$

In particular, it implies that

$$|\hat\theta^{\mathcal{L}}_{S^c} - \theta^*_{S^c}|_1 \le 3|\hat\theta^{\mathcal{L}}_S - \theta^*_S|_1\,.$$

so that $\theta = \hat\theta^{\mathcal{L}} - \theta^*$ satisfies the cone condition (2.17). Using now the Cauchy-Schwarz inequality and Lemma 2.17 respectively, we get since $|S| \le k$,

$$|\hat\theta^{\mathcal{L}}_S - \theta^*|_1 \le \sqrt{|S|}|\hat\theta^{\mathcal{L}}_S - \theta^*|_2 \le \sqrt{\frac{2k}{n}}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2\,.$$

Combining this result with (2.18), we find

$$|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \le 32nk\tau^2\,.$$

Moreover, it yields

$$\begin{aligned}|\hat\theta^{\mathcal{L}} - \theta^*|_1 &\le 4\sqrt{\frac{2k}{n}}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2 \\ &\le 4\sqrt{\frac{2k}{n}}\sqrt{32nk\tau^2} \le 32k\tau\end{aligned}$$

The bound in expectation follows using the same argument as in the proof of Corollary 2.9. □

Note that all we required for the proof was not really incoherence but the conclusion of Lemma 2.17:

$$\inf_{|S|\le k}\inf_{\theta\in\mathcal{C}_S}\frac{|\mathbb{X}\theta|_2^2}{n|\theta_S|_2^2} \ge \kappa \qquad (2.19)$$

where $\kappa = 1/2$ and $\mathcal{C}_S$ is the cone defined by

$$\mathcal{C}_S = \big\{|\theta_{S^c}|_1 \le 3|\theta_S|_1\big\}\,.$$

Condition (2.19) is sometimes called *restricted eigenvalue (RE) condition*. Its name comes from the following observation. Note that all k-sparse vectors θ are in a cone $\mathcal{C}_S$ with $|S| \le k$ so that the RE condition implies that the smallest eigenvalue of $\mathbb{X}_S$ satisfies $\lambda_{\min}(\mathbb{X}_S) \ge n\kappa$ for all S such that $|S| \le k$. Clearly, the RE condition is weaker than incoherence and it can actually be shown that a design matrix $\mathbb{X}$ of i.i.d Rademacher random variables satisfies the RE conditions as soon as $n \ge Ck \log(d)$ with positive probability.

2.5 PROBLEM SET

Problem 2.1. Consider the linear regression model with fixed design with $d \le n$. The *ridge* regression estimator is employed when the $\mathrm{rank}(\mathbb{X}^\top\mathbb{X}) < d$ but we are interested in estimating θ^*. It is defined for a given parameter $\tau > 0$ by

$$\hat\theta^{\text{ridge}}_\tau = \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}|Y-\mathbb{X}\theta|_2^2 + \tau|\theta|_2^2\Big\}\,.$$

(a) Show that for any τ, $\hat\theta^{\text{ridge}}_\tau$ is uniquely defined and give its closed form expression.

(b) Compute the bias of $\hat\theta^{\text{ridge}}_\tau$ and show that it is bounded in absolute value by $|\theta^*|_2$.

Problem 2.2. Let $X = (1, Z, \dots, Z^{d-1})^\top \in \mathbb{R}^d$ be a random vector where Z is a random variable. Show that the matrix $\mathbb{E}(XX^\top)$ is positive definite if Z admits a probability density with respect to the Lebesgue measure on $\mathbb{R}$.

Problem 2.3. In the proof of Theorem 2.11, show that $4\min(|\theta^*_j|, \tau)$ can be replaced by $3\min(|\theta^*_j|, \tau)$, i.e., that on the event $\mathcal{A}$, it holds

$$|\hat\theta^{\text{HRD}}_j - \theta^*_j| \le 3\min(|\theta^*_j|, \tau)\,.$$

Problem 2.4. For any $q > 0$, a vector $\theta \in \mathbb{R}^d$ is said to be in a weak ℓ_q ball of radius R if the decreasing rearrangement $|\theta_{[1]}| \ge |\theta_{[2]}| \ge \dots$ satisfies

$$|\theta_{[j]}| \le Rj^{-1/q}\,.$$

Moreover, we define the weak ℓ_q norm of θ by

$$|\theta|_{w\ell_q} = \max_{1\le j\le d} j^{1/q}|\theta_{[j]}|$$

(a) Give examples of $\theta, \theta' \in \mathbb{R}^d$ such that

$$|\theta+\theta'|_{w\ell_1} > |\theta|_{w\ell_1} + |\theta'|_{w\ell_1}$$

What do you conclude?

(b) Show that $|\theta|_{w\ell_q} \le |\theta|_q$.

(c) Show that if $\lim_{d\to\infty}|\theta|_{w\ell_q} < \infty$, then $\lim_{d\to\infty}|\theta|_{q'} < \infty$ for all $q' > q$.

(d) Show that, for any $q \in (0,2)$ if $\lim_{d\to\infty}|\theta|_{w\ell_q} = C$, there exists a constant $C_q > 0$ that depends on q but not on d and such that under the assumptions of Theorem 2.11, it holds

$$|\hat\theta^{\text{HRD}} - \theta^*|_2^2 \le C_q\Big(\frac{\sigma^2\log 2d}{n}\Big)^{1-\frac{q}{2}}$$

with probability .99.

Problem 2.5. Show that

$$\begin{aligned}\hat{\theta}^{\text{HRD}} &= \mathop{\mathrm{argmin}}_{\theta \in \mathbb{R}^d} \Big\{ |y-\theta|_2^2 + 4\tau^2 |\theta|_0 \Big\} \\ \hat{\theta}^{\text{SFT}} &= \mathop{\mathrm{argmin}}_{\theta \in \mathbb{R}^d} \Big\{ |y-\theta|_2^2 + 4\tau |\theta|_1 \Big\}\end{aligned}$$

Problem 2.6. Assume that the linear model (2.2) with $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$ and $\theta^* \neq 0$. Show that the modified BIC estimator $\hat{\theta}$ defined by

$$\hat{\theta} \in \mathop{\mathrm{argmin}}_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{n}|Y - \mathbb{X}\theta|_2^2 + \lambda |\theta|_0 \log\Big(\frac{ed}{|\theta|_0}\Big) \Big\}$$

satisfies,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) \lesssim |\theta^*|_0 \sigma^2 \frac{\log\big(\frac{ed}{|\theta^*|_0}\big)}{n} \,.$$

with probability .99, for appropriately chosen λ. What do you conclude?

Problem 2.7. Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume the conditions of Theorem 2.2 and that the columns of X are normalized in such a way that $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$. Then the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter

$$2\tau = 8\sigma\sqrt{\frac{2\log(2d)}{n}}\,,$$

satisfies

$$|\hat{\theta}^{\mathcal{L}}|_1 \leq C|\theta^*|_1$$

with probability $1-(2d)^{-1}$ for some constant C to be specified.