Notes of high-dimensional probability and statistics

Ruihan Liu

Contents

1 Introduction

1.1 Curse of Dimensionality

In general, dealing with data is the core task of statistics, and with the development of information technology, more and more amount of data has been generated. It is common to analysis high-dimensional data, which mainly comes from biological data, images, marketing and business. Besides, high-dimensional data can indeed provide more hidden and potential information, the only trouble is how to deal with such large amount data by computers. In fact, high-dimensional data will cause many unbelieving phenomenon which hardly happens in low-dimensional statistics.

Example 1.1.1 (Fluctuations cumulate)**.** *Let $X \sim X^{(1)}, \cdots, X^{(n)} \in \mathbb{R}^p$ be i.i.d. samples with covariance of $\sigma^2 \mathbf{I}_p$, then estimation of $\mathbb{E}[X]$ is by*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}$$

then

$$\mathbb{E}\left[||\bar{X}_n - \mathbb{E}[X]||^2\right] = \sum_{j=1}^{p} \mathbb{E}\left[\left([\bar{X}_n]_j - \mathbb{E}[X_j]\right)^2\right] = \sum_{j=1}^{p} \operatorname{var}([\bar{X}_n]_j) = \frac{p}{n}\sigma^2$$

By the law of large number, the above expectation will converge to zero as $n \to \infty$, however, when the number of samples is fixed, $\sigma^2 p/n$ will be huge if dimension p is large.

Example 1.1.2 (Locality lost)**.** *Suppose the observation $(Y_i, X^{(i)}) \in \mathbb{R} \times [0,1]^p$ are i.i.d. for $i = 1, \cdots, n$ and model*

$$Y_i = f(X^{(i)}) + \epsilon_i$$

where f is smooth and ϵ_i are independent and centered. In addition, let $X^{(i)} \sim \mathcal{U}\left([0,1]^p\right)$ and define local average by

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i, \text{ for } X^{(i)} \to x \in [0,1]^p$$

The main dilemma of high-dimensional is the neighbor of one point, consider the pairwise distance $\{||X^{(i)} - X^{(j)}||^2 : i \neq j\}$. In detail, we have

$$M := \mathbb{E}\left[||X^{(i)} - X^{(j)}||^2\right] = \sum_{k=1}^{p} \mathbb{E}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right] = \frac{p}{6}$$

$$V := \sqrt{\sum_{k=1}^{p} \operatorname{var}\left[\left(X_k^{(i)} - X_k^{(j)}\right)^2\right]} = \frac{1}{6}\sqrt{\frac{7p}{5}}$$

Hence, we obtain $V/M \sim p^{-1/2}$, i.e. the square distances between two points generated by $\mathcal{U}\left([0,1]^p\right)$ grows linearly with p, while the scaled deviation V/M shrinks like $p^{-1/2}$. How *many observations do we need if there exists at least one $X^{(i)}$ such that $||X^{(i)} - x||^2 < 1$? Let $V_p(r)$ be the volume of p-dimensional ball with radius r and*

$$V_p(r) = \frac{\pi^{p/2}}{\Gamma(p/2+1)} r^p \sim \left(\frac{2\pi e r^2}{p}\right)^{p/2} (p\pi)^{-1/2} \text{ as } p \to \infty$$

In order to satisfy above requirement, we need

$$[0,1]^p \subset \bigcup_{i=1}^{n} \mathcal{B}(X^{(i)}, 1)$$

Or simply $1 \leq nV_p(1)$, that is,

$$n \geq \frac{\Gamma(p/2+1)}{\pi^{p/2}} \sim \left(\frac{p}{2\pi e}\right)^{p/2} \sqrt{p\pi} \text{ as } p \to \infty$$

Hence, as the dimension $p \to \infty$, the required number of samples also tends to infinity.

Example 1.1.3 (Thin tails concentration)**.** *For $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\epsilon > 0$ small,*

$$\begin{aligned}
\frac{1}{\epsilon}\mathbb{P}\left[R \leq ||X|| \leq R + \epsilon\right] &= \frac{1}{\epsilon}\int_{R \leq ||X|| \leq R+\epsilon} e^{-||\boldsymbol{x}||^2/2} \frac{d\boldsymbol{x}}{(2\pi)^{p/2}} \\
&= \frac{1}{\epsilon}\int_{R}^{R+\epsilon} e^{-r^2/2} r^{p-1} \frac{pV_p(1)dr}{(2\pi)^{p/2}} \\
&\approx \frac{p}{2^{p/2}\Gamma(p/2+1)} R^{p-1} e^{-R^2/2}
\end{aligned}$$

It is easy to see that the mass is concentrated around $R = \sqrt{p-1}$, hence Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ behaves like a uniform distribution on the sphere with radius of $\sqrt{p-1}$.

Although there are many dilemmas in high-dimensional statistics, some possible methods are worth considering.

Low-dimensional structures: high-dimensional data are usually concentrated around low-dimensional structures reflecting the (relatively) small complexity of the systems producing data. The simplest and most widely used technique for this purpose is the Principal Component Analysis (PCA), which is a ‘unsupervised’ method. For any data points $X^{(1)}, \cdots, X^{(n)} \in \mathbb{R}^p$ and any dimension $d \leq p$, PCA computes the linear span in $\mathbb{R}^p$

$$V_d \in \arg\min_{\dim(V) \leq d} \sum_{i=1}^{n} ||X^{(i)} - \text{Proj}_V X^{(i)}||^2 \tag{1}$$

where Proj_V is the orthogonal projection matrix onto V.

A Paradigm Shift: Classical statistics provide a very rich theory for analyzing data with the following characteristics:

- a small number p of parameters

- a large number n of observations
- we investigate the performances of estimators when $n \to \infty$ (central limit theorem)

Data in actual fields (high-dimensional)

- a huge number p of parameters
- a sample size n, which is either roughly of the same size as p, or sometimes much smaller than p

Hence the $n \to \infty$ asymptotic does not fit anymore in this case. Finally, in order to quantify non asymptotically the performances of an estimator, let's introduce some useful statistics tools in the next subsection.

1.2 Concentration Inequalities

Theorem 1.2.1 (Central limit theorem). *For $f : \mathbb{R} \to \mathbb{R}$ and $X_1, \cdots, X_n$ are i.i.d. such that* $\operatorname{var}(f(X_1)) < \infty$*, when $n \to \infty$*

$$\sqrt{\frac{n}{\operatorname{var}(f(X_1))}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_1)] \right) \to Z \sim \mathcal{N}(0,1) \text{ in distribution} \tag{2}$$

In addition, assume f is L-Lipschitz and $X_1, \cdots, X_n$ are i.i.d. with finite variance $\operatorname{var}(X_1) = \sigma^2$*, then*

$$\operatorname{var}(f(X_1)) = \frac{1}{2}\mathbb{E}\left[(f(X_1) - f(X_2))^2\right] \leq \frac{L^2}{2}\mathbb{E}\left[(X_1 - X_2)^2\right] = L^2\sigma^2 \tag{3}$$

Hence

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_1)] \geq \frac{L\sigma}{\sqrt{n}} x \right) \leq \mathbb{P}(|Z| \geq x) \leq e^{-x^2/2} \text{ for } x > 0 \tag{4}$$

Proof. For the last inequality in (4), define

$$\varphi(x) = e^{-x^2/2} - \sqrt{\frac{2}{\pi}} \int_x^\infty e^{-t^2/2} dt, \ \ x \geq 0$$

Then $\varphi'(x) = \left(\sqrt{2/\pi} - x\right) e^{-x^2/2}$ and $\varphi(0) = 0$, hence $\varphi(x) \geq 0$ in $[0, \sqrt{2/\pi}]$. For $x > \sqrt{2/\pi}$,

$$\sqrt{\frac{2}{\pi}} \int_x^\infty e^{-t^2/2} dt \leq \int_x^\infty t e^{-t^2/2} dt = e^{-x^2/2}$$

Therefore, $\varphi(x) \geq 0$ for $x \geq 0$. □

Theorem 1.2.2 (Gaussian concentration inequality). *If $X_1, \cdots, X_n$ are i.i.d. with $\mathcal{N}(0, \sigma^2)$ and $F : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz, then there exists ξ with exponential distribution of parameter 1 such that*

$$F(X_1, \cdots, X_n) \leq \mathbb{E}[F(X_1, \cdots, X_n)] + L\sigma\sqrt{2\xi} \tag{5}$$

Proof. Equivalently, we only need to show

$$\mathbb{P}\left(F(X_1, \cdots, X_n) \geq \mathbb{E}\left[F(X_1, \cdots, X_n)\right] + L\sigma\sqrt{2x}\right) \leq e^{-x} \tag{6}$$

Let $(W_t)_{t\geq 0}$ be the standard Wiener process in $\mathbb{R}^n$ and for $x \in \mathbb{R}^n$, $t \in [0, 1]$, set

$$G(t, x) = \mathbb{E}\left[F\left(x + W_{1-t}\right)\right]$$

The function G is continuous on $[0, 1] \times \mathbb{R}^n$, differentiable in t, and infinitely differentiable in x for any $t \in [0, 1)$. Then by Ito's formula, it concludes that

$$\begin{aligned} G(1, W_1) &= G(0, 0) + \int_0^1 \frac{\partial}{\partial t} G(s, W_s) ds + \int_0^1 \bigtriangledown_x G(s, W_s) \cdot dW_s + \frac{1}{2}\int_0^1 \triangle_x G(s, W_s) ds \\ &= G(0, 0) + \int_0^1 \bigtriangledown_x G(s, W_s) \cdot dW_s \end{aligned}$$

□

Definition 1.2.1 (Infinitesimal generator)**.** *Let $X = (X_t)_{t\geq 0}$ is a Markov process and $f : \mathbb{R}^n \to \mathbb{R}$ is bounded, the operator $\mathscr{L}$ is defined by*

$$\mathscr{L} f(x) = \lim_{t\to 0} \frac{\mathbb{E}\left[f(X_t)|X_0 = x\right] - f(x)}{t} \tag{7}$$

where $\mathscr{L}$ is called the infinitesimal generator of X_t.

Lemma 1.2.1. *Let $W = (W_t)_{t\geq 0}$ is a n-dimensional Wiener process with drift $\boldsymbol{m} = (m_1, \cdots, m_n)^T$ and covariance matrix Γ, then the infinitesimal generator $\mathscr{L}$ of W_t is*

$$\mathscr{L} f(x) = \boldsymbol{m} \cdot \bigtriangledown f(x) \tag{8}$$

2 Model Selection

2.1 Statistical settings

In this section, we will focus on the following regression model

$$y_i = f_i(x^{(i)}) + \epsilon_i \quad \text{for } i = 1, \cdots, n \tag{9}$$

where $\mathbb{E}[\epsilon_i] = 0$ and $x^{(i)} \in \mathbb{R}^p$. In general, the explicit form of f^* is unknown and our main goal is to give a proper estimation for this term, besides, the noise ϵ_i are often assumed to be $\mathcal{N}(0, \sigma^2)$. In particular, we often choose $f_i(x_i) = \langle \beta_i, x_i \rangle$, which is called the linear model. As a result, we will express our model in vector for the purpose of simplification, i.e.

$$Y = f^*(\boldsymbol{X}) + \boldsymbol{\epsilon} = \boldsymbol{X}\beta^* + \boldsymbol{\epsilon} \tag{10}$$

where $\boldsymbol{X} := [x_j^{(i)}]_{i,j=1}^{n,p}$ is a $n \times p$ matrix and $\boldsymbol{\epsilon} = (\epsilon_i)_{i=1}^n$ is a n-dimensional vector. Anyway, why do we consider the regression model? In fact, we have

$$Y = \mathbb{E}\left[Y|\boldsymbol{X}\right] + \left(Y - \mathbb{E}\left[Y|\boldsymbol{X}\right]\right)$$

where the first term is a function of $\boldsymbol{X}$ while the other has zero mean by the property of conditional expectation.

Sparsity patterns:

- **Coordinate sparsity:** Only a few coordinates of β^* are nonzero, i.e. $|\beta^*|_0 := \operatorname{card}\{i : \beta_i \neq 0\}$ is small.
- **Group sparsity:** The coordinates of β^* are clustered into groups, and only a few groups are nonzero. More precisely, we have a partition $\{1, \cdots, p\} = \bigcup_{k=1}^M G_k$, and only a few coordinates of $\beta^*_{G_k} = (\beta_i)_{i \in G_k}$ are nonzero

In general, the sparsity patterns are hidden statistical structure and we will give an estimation method in the next subsection which is called *Model Selection*.

2.2 To Select among a Collection of Models

Known structure:

- If we know $m^* := \operatorname{supp}(\beta^*)$, the regression model can be rewritten by

$$y_i = \sum_{j \in m^*} \beta_j^* x_j^{(i)} + \epsilon_j$$

- More generally, if we know f^* belongs to some subspace $S \subset \mathbb{R}^n$, we can maximize the likelihood function with the constraint $f^* \in S$. For example, in the coordinate-sparse setting where we know that the nonzero coordinates m^*, we may choose $S = \operatorname{span}\{x_j : j \in m^*\}$, under the Gaussian noises assumption, the likelihood is

$$\widehat{f} := \arg\min_{f \in S} -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|Y - f\|_2$$

Hence, the solution is $\widehat{f} = \operatorname{Proj}_S Y$.

However, S is usually unknown in practice, so we may try

1. consider a collection $\{S_m : m \in \mathscr{M}\}$ of linear subspaces of $\mathbb{R}^n$, called models;
2. associate to each subspace S_m the constrained maximum likelihood estimators $\widehat{f}_m = \mathrm{Proj}_{S_m} Y$; and
3. finally estimate $\widehat{f}$ by the *best* estimator among the collection $\left\{\widehat{f}_m : m \in \mathscr{M}\right\}$.

As a result, we define the L^2-risk as a measure to describe the 'distance' between $\widehat{f}$ and f^* by

$$R(\widehat{f}) = \mathbb{E}\left[\|\widehat{f} - f^*\|_2^2\right] \tag{11}$$

Hence, the *oracle estimator* is defined by $\widehat{f}_{m_0}$ where $m_0 := \arg\min_{m \in \mathscr{M}} R(\widehat{f}_m)$.

2.3 Model Selection Procedures

Computation of Risk:

Denote $\widehat{f}_{S_m} := \mathrm{Proj}_{S_m}(Y)$ and we have

$$\begin{aligned} R(\widehat{f}_{S_m}) &= \mathbb{E}\left[\|\widehat{f}_{S_m} - f^*\|_2^2\right] \\ &= \mathbb{E}\left[\|\mathrm{Proj}_{S_m}(f^* + \boldsymbol{\epsilon}) - f^*\|_2^2\right] \\ &= \mathbb{E}\left[\|\mathrm{Proj}_{S_m}(f^*) - f^*\|_2^2\right] + \mathbb{E}\left[\|\mathrm{Proj}_{S_m}(\boldsymbol{\epsilon})\|_2^2\right] \\ &= \underbrace{\|\mathrm{Proj}_{S_m}(f^*) - f^*\|_2^2}_{\text{Bias}} + \underbrace{\sigma^2 \dim(S_m)}_{\text{Variance}} \end{aligned}$$

Hence the oracle minimize $m_0 := \arg\min_{m \in \mathscr{M}} \left\{\|\mathrm{Proj}_{S_m}(f^*) - f^*\|_2^2 + \sigma^2 \dim(S_m)\right\}$. Moreover, the variance term is increasing as the dimension of S_m is increasing, while the bias term is decreasing. So we need to find a balance between these two terms. However, the bias term is usually unknown, which adds additional trouble for our estimation. One naive way is to consider the following method.

Unbiased estimator of the risk:

By the definition of $\widehat{f}_{S_m} = \mathrm{Proj}_{S_m}(f^* + \boldsymbol{\epsilon})$ and hence $Y - \widehat{f}_{S_m} = (I - \mathrm{Proj}_{S_m})(f^* + \boldsymbol{\epsilon})$

$$\begin{aligned} \mathbb{E}\left[\|\widehat{f}_{S_m} - Y\|_2^2\right] &= \mathbb{E}\left[\left\|\left(I - \mathrm{Proj}_{S_m}\right) f^*\right\|_2^2 + \left\|\left(I - \mathrm{Proj}_{S_m}\right) \boldsymbol{\epsilon}\right\|_2^2 + 2\langle\left(I - \mathrm{Proj}_{S_m}\right) f^*, \boldsymbol{\epsilon}\rangle\right] \\ &= \left\|\left(I - \mathrm{Proj}_{S_m}\right) f^*\right\|_2^2 + (n - \dim(S_m))\,\sigma^2 \\ &= R(\widehat{f}_{S_m}) + (n - 2\dim(S_m))\,\sigma^2 \end{aligned}$$

As a consequence, define $\widehat{R}(\widehat{f}_{S_m}) := \|Y - \widehat{f}_{S_m}\|_2^2 + (2\dim(S_m) - n)\,\sigma^2$ as the unbiased estimator of $R(\widehat{f}_{S_m})$, which is also called the *Akaike Information Criterion* (AIC) and

$$\widehat{m}_{\text{AIC}} := \arg\min_{m\in\mathscr{M}} \left\{ \|Y - \widehat{f}_{S_m}\|_2^2 + (2\dim(S_m) - n)\,\sigma^2 \right\} \tag{12}$$

Although this criterion is very natural and popular, it does not work when $|\mathscr{M}|$ is very large with an exponential number of models per dimension.

Penalized estimator of the risk:

How could we solve the above dilemma? One possible way is to consider a general *penalized estimator* which is determined by $m \in \mathscr{M}$, that is, we focus henceforth on a selection criterion of the form

$$\widehat{m} := \arg\min_{m\in\mathscr{M}} \left\{ \|Y - \widehat{f}_{S_m}\|_2^2 + \text{pen}(m)\sigma^2 \right\} \tag{13}$$

where pen : $\mathscr{M} \to \mathbb{R}^+$ is called the *penalty function*. Hence, our goal is to find the proper form of penalties. The idea is

- for a given pem(m), analyse $R(\widehat{f}_{S_{\widehat{m}}})$.
- design pen(m) in order to have a good estimation of $R(\widehat{f}_{S_{\widehat{m}}})$.
- the ideal inequality may have the following form

  $$R(\widehat{f}_{S_m}) \leq C \min_{m\in\mathscr{M}} R(\widehat{f}_{S_m}) + \text{Remaining Terms} \tag{14}$$

  where C is a constant closed to 1 and the remaining terms are hopefully small.

2.4 Risk Bound for Model Selection

Theorem 2.4.1 (Oracle risk bound for model selection). *Set*

$$B(d,\alpha) = \mathbb{E}\left[\left(\left(\sqrt{d} + \sqrt{2\xi}\right)^2 - \alpha\right)_+\right]$$

where $\xi \sim \text{Exp}(1)$*, then for any* $a > 1$*, it concludes that*

$$\frac{a}{a-1} R(\widehat{f}_{S_{\widehat{m}}}) \leq \min_{m\in\mathscr{M}} \left\{ R(\widehat{f}_{S_m}) + \text{pen}(m)\sigma^2 \right\} + a\sigma^2 \mathscr{P}(\mu) \tag{15}$$

where $\mathscr{P}(\mu) := 1 + \sum_{m\in\mathscr{M}} B(\dim(S_m), \text{pen}(m)/a)$.

Proof. First, according to the definition of $\widehat{m}$, it concludes that

$$\|Y - \widehat{f}_{S_{\widehat{m}}}\|_2^2 + \text{pen}(\widehat{m})\sigma^2 \leq \|Y - \widehat{f}_{S_m}\|_2^2 + \text{pen}(m)\sigma^2$$

Then replace Y by $f^* + \boldsymbol{\epsilon}$, we deduce

$$\|f^* - \widehat{f}_{S_{\widehat{m}}}\|_2^2 \le \|f^* - \widehat{f}_{S_m}\|_2^2 + \mathrm{pen}(m)\sigma^2 + 2\langle \boldsymbol{\epsilon}, f^* - \widehat{f}_{S_m}\rangle + 2\langle \boldsymbol{\epsilon}, \widehat{f}_{S_{\widehat{m}}} - f^*\rangle - \mathrm{pen}(\widehat{m})\sigma^2$$

Since

$$\mathbb{E}\left[\langle \boldsymbol{\epsilon}, f^* - \widehat{f}_{S_m}\rangle\right] = -\mathbb{E}\left[\|\mathrm{Proj}_{S_m}\boldsymbol{\epsilon}\|_2^2\right] \le 0$$

and let $\bar{S}_{\widehat{m}} := S_{\widehat{m}} + \langle f^*\rangle = \widetilde{S}_{\widehat{m}} \oplus \langle f^*\rangle$, we have

$$\begin{aligned}
2\langle \boldsymbol{\epsilon}, \widehat{f}_{S_{\widehat{m}}} - f^*\rangle &= 2\langle \mathrm{Proj}_{\bar{S}_{\widehat{m}}}\boldsymbol{\epsilon}, \widehat{f}_{S_{\widehat{m}}} - f^*\rangle \\
&\le a\|\mathrm{Proj}_{\bar{S}_{\widehat{m}}}\boldsymbol{\epsilon}\|_2^2 + \frac{1}{a}\|\widehat{f}_{S_{\widehat{m}}} - f^*\|_2^2 \\
&= a\|\mathrm{Proj}_{\widetilde{S}_{\widehat{m}}}\boldsymbol{\epsilon}\|_2^2 + a\|\mathrm{Proj}_{\langle f^*\rangle}\boldsymbol{\epsilon}\|_2^2 + \frac{1}{a}\|\widehat{f}_{S_{\widehat{m}}} - f^*\|_2^2
\end{aligned}$$

Hence for any $m \in \mathscr{M}$, we obtain

$$\frac{a}{a-1}R(\widehat{f}_{S_{\widehat{m}}}) \le R(\widehat{f}_{S_m}) + \mathrm{pen}(m)\sigma^2 + a\mathbb{E}\left[\|\mathrm{Proj}_{\langle f^*\rangle}\boldsymbol{\epsilon}\|_2^2\right] + a\mathbb{E}\left[\|\mathrm{Proj}_{\widetilde{S}_{\widehat{m}}}\boldsymbol{\epsilon}\|_2^2 - \frac{\mathrm{pen}(\widehat{m})}{a}\sigma^2\right]$$

In addition, since $\mathbb{E}\left[\|\mathrm{Proj}_{\langle f^*\rangle}\boldsymbol{\epsilon}\|_2^2\right] = \sigma^2$ and

$$\begin{aligned}
\mathbb{E}\left[\|\mathrm{Proj}_{\widetilde{S}_{\widehat{m}}}\boldsymbol{\epsilon}\|_2^2 - \frac{\mathrm{pen}(\widehat{m})}{a}\sigma^2\right] &\le \mathbb{E}\left[\sup_{m\in\mathscr{M}} \|\mathrm{Proj}_{\widetilde{S}_m}\boldsymbol{\epsilon}\|_2^2 - \frac{\mathrm{pen}(m)}{a}\sigma^2\right] \\
&\le \sum_{m\in\mathscr{M}} \mathbb{E}\left[\left(\|\mathrm{Proj}_{\widetilde{S}_m}\boldsymbol{\epsilon}\|_2^2 - \frac{\mathrm{pen}(m)}{a}\sigma^2\right)_+\right]
\end{aligned}$$

According to Theorem 1.2.2, there exists $\xi_m \sim \mathrm{Exp}(1)$ such that

$$\|\mathrm{Proj}_{\widetilde{S}_m}\boldsymbol{\epsilon}\|_2^2 \le \left(\sqrt{\mathbb{E}\left[\|\mathrm{Proj}_{\widetilde{S}_m}\boldsymbol{\epsilon}\|_2^2\right]} + \sigma\sqrt{2\xi_m}\right)^2 = \sigma^2\left(\sqrt{\dim(\bar{S}_m)} + \sqrt{2\xi_m}\right)$$

Finally, since $\dim(\bar{S}_m) \le \dim(S_m)$, we conclude that for any $m \in \mathscr{M}$

$$\frac{a}{a-1}R(\widehat{f}_{S_{\widehat{m}}}) \le R(\widehat{f}_{S_m}) + \mathrm{pen}(m)\sigma^2 + a\sigma^2\mathscr{P}(\mu)$$

and taking the minimum of $m \in \mathscr{M}$ which completes our proof. □

Now, we would like to choose some special forms of penalties. For instance, suppose the probability of choosing $m \in \mathscr{M}$ is π_m such that $\sum_{m\in\mathscr{M}}\pi_m = 1$. Moreover, based on the definition of $B(d,\alpha)$, we have the following estimation

$$\begin{aligned}
\mathbb{E}\left[\left(\left(\sqrt{d} + \sqrt{2\xi}\right)^2 - \alpha\right)_+\right] &= \int_{(\sqrt{\alpha}-\sqrt{d})^2/2}^{\infty} \left(\left(\sqrt{d} + \sqrt{2x}\right)^2 - \alpha\right) e^{-x}dx \\
&\asymp \exp\left(-\frac{1}{2}(\sqrt{\alpha} - \sqrt{d})^2\right)
\end{aligned}$$

So we may take the penalty pen(m) such that

$$\exp\left(-\frac{1}{2}\left(\sqrt{\dim(S_m)}-\sqrt{\frac{\text{pen}(m)}{2}}\right)^2\right)=\pi_m$$

i.e.

$$\text{pen}(m)=a\left(\sqrt{\dim(S_m)}+\sqrt{2\log\frac{1}{\pi_m}}\right)\tag{16}$$

where $a>1$.

Corollary 2.4.1. *For the above assumption with π_m, there exists a constant $C_a>1$ depending only on $a>1$, such that*

$$R(\widehat{f}_{S_{\widehat{m}}})\leq C_a\min_{m\in\mathscr{M}}\left\{R(\widehat{f}_m)+\left(1+\log\frac{1}{\pi_m}\right)\sigma^2\right\}\tag{17}$$

Choice of π_m:

- we want π_m such that the left side of Corollary 2.4.1 as small as enough.
- according to oracle inequality, we want $-\log\pi_m\leq C\dim(S_m)$ since

$$R(\widehat{f}_{S_m})=\|\text{Proj}_{S_m}f^*-f^*\|_2^2+\dim(S_m)\sigma^2\geq\dim(S_m)\sigma^2$$

However, it is not always possible when $|\mathscr{M}|$ is very large.

Example 2.4.1. *Let consider a case of coordinate sparsity, let $\mathscr{M}=\mathscr{P}(\{1,\cdots,p\})$, where $\mathscr{P}(\{1,\cdots,p\})$ denotes the collection of all subsets of $\{1,\cdots,p\}$, then we set the same probability to the C_p^d models of cardinality d. Next, taking $\pi_m\propto e^{-s|m|}$, where $s>0$, then*

$$\sum_{m\in\mathscr{M}}e^{-s|m|}=\sum_{d=0}^{p}C_p^d e^{-sd}=(1+e^{-s})^p$$

As a result, we obtain

$$\pi_m:=\frac{e^{-s|m|}}{(1+e^{-s})^p}$$

and

$$\log\frac{1}{\pi_m}=p\log(1+e^{-s})+s|m|$$

Suppose $s=\log p$, we conclude that $-\log\pi_m\leq 1+|m|\log p$. Another choice of π_m is

$$\frac{1}{C_p^{|m|}}\frac{e-1}{e-e^{-p}}e^{-|m|}$$

then we have

$$\log\frac{1}{\pi_m}\leq\log\left(\frac{e}{e-1}\right)+|m|\log\left(\frac{e^2p}{|m|}\right)$$

In fact, for both above cases, they all have $\log p$ term, which is actually unavoidable, see the next subsection. Hence, the ideal choice of π_m is not always possible as $|\mathscr{M}|$ (i.e. p) is very large.

2.5 Cases Study–Orthogonal Design

Consider the linear regression model with coordinate sparsity setting and assume the columns of $\boldsymbol{X}$ are orthogonal, the family $\mathscr{M}$ and the models $S_m := \mathrm{span}_{j\in m}\left\{\boldsymbol{X}_j^T\right\}$ of the coordinate-sparse setting and take the penalty

$$\mathrm{pen}(m) := K\left(1+\sqrt{2\log p}\right)^2 |m|$$

In fact, if we use assumption in Example 2.4.1, i.e. $\pi_m := p^{-|m|}(1+1/p)^{-p}$, then the penalty is actually equal to

$$K\left(\sqrt{\dim(S_m)}+\sqrt{2\log(1/\pi_m)}\right)^2$$

Hard thresholding:

If we denote

$$\widehat{m}_\lambda := \min_{m\in\mathscr{M}}\left\{\|Y-\mathrm{Proj}_{S_m}Y\|_2^2+\mathrm{pen}(m)\sigma^2\right\}$$

where $\mathrm{pen}(m)=\lambda|m|$ and $\lambda := K\left(1+\sqrt{2\log p}\right)^2>0$, then we have

$$\|Y-\mathrm{Proj}_{S_m}Y\|_2^2+\mathrm{pen}(m)\sigma^2 = \|Y\|_2^2+\sum_{j\in m}\left(\lambda\sigma^2-\left(\frac{\boldsymbol{X}_j^TY}{\|\boldsymbol{X}_j\|_2}\right)^2\right) \tag{18}$$

where $\boldsymbol{X}_j$ is the j-th column of $\boldsymbol{X}$.

Proof. In fact, we just need to show that

$$\|\mathrm{Proj}_{S_m}Y\|_2^2=\sum_{j\in m}\left(\frac{\boldsymbol{X}_j^TY}{\|\boldsymbol{X}_j\|_2}\right)^2$$

For the right side, it is equal to $\|\mathrm{Proj}_{S_m}\boldsymbol{X}Y\|_2^2=\|\mathrm{Proj}_{S_m}Y\|_2^2$ due to the definition of S_m. □

As a result, the minimizer $\widehat{m}_\lambda$ is given by

$$\widehat{m}_\lambda := \left\{j : \left(\boldsymbol{X}_j^TY\right)^2>\lambda\|\boldsymbol{X}_j\|_2^2\sigma^2\right\}$$

which is easy to obtain due to above result.

Minimal penalties:

In this part, we consider the penalty $\mathrm{pen}(m) := 2K|m|\log p$ and $f^*\equiv 0$, where $K<1$. Then we have for $j=1,\cdots,p$, set $Z_j := \boldsymbol{X}_j^T\boldsymbol{\epsilon}/\left(\|\sigma\boldsymbol{X}_j\|_2\right)$, then Z_j are i.d.d. with distribution of $\mathcal{N}(0,1)$. Hence, the minimizer $\widehat{m}_\lambda$ satisfies

$$\widehat{m}_\lambda := \left\{j : Z_j^2>2K\log p\right\}$$

where $|\widehat{m}_\lambda|$ follows a binomial distribution with parameters p and probability $\mathbb{P}\left(Z_j^2 > 2K\log p\right)$, which has an estimation of

$$\frac{p^{-K}}{\sqrt{K\pi\log p}}\left(1-\frac{1}{2K\log p}\right) \leq \mathbb{P}\left(Z_j^2 > 2K\log p\right) \leq \frac{p^{-K}}{\sqrt{K\pi\log p}}$$

Hence we have

$$\mathbb{P}\left(Z_j^2 > 2K\log p\right) \asymp \frac{p^{-K}}{\sqrt{K\pi\log p}} \quad \text{as } p\to\infty$$

As a result, we obtain

$$\mathbb{E}\left[|\widehat{m}_\lambda|\right] \asymp \frac{p^{1-K}}{\sqrt{K\pi\log p}}$$

The mean of selected models $\widehat{m}_\lambda$ grows like a fractional power of p as $K<1$, since an accurate value of λ must then be such that $\mathbb{E}\left[|\widehat{m}_\lambda|\right]$ is small.

Overfitting with $K<1$:

In this part, we do not assume that $f^*\equiv 0$ and let $\lambda := 2K\log p$ with $K<1$. As before, set $Z_j := \boldsymbol{X}_j^T\boldsymbol{\epsilon}/\left(\|\sigma\boldsymbol{X}_j\|_2\right)$ and $f^*\in S_{m^*}$ with $|m^*| = D^*$. Writing P_j for the projection on the line spanned by $\boldsymbol{X}_j$, then

$$\|\widehat{f}_{\widehat{m}_\lambda} - f^*\|_2^2 = \|f^* - \sum_{j\in\widehat{m}_\lambda} P_j f^*\|_2^2 + \sum_{j\in\widehat{m}_\lambda} Z_j^2\sigma^2$$

In fact, since

$$f^* - \widehat{f}_{\widehat{m}_\lambda} = f^* - \mathrm{Proj}_{S_{\widehat{m}_\lambda}}\left(f^*+\boldsymbol{\epsilon}\right) = \left(I - \mathrm{Proj}_{S_{\widehat{m}_\lambda}}\right)f^* - \mathrm{Proj}_{S_{\widehat{m}_\lambda}}\boldsymbol{\epsilon}$$

which can deduce above equation. As a result, we obtain

$$\|\widehat{f}_{\widehat{m}_\lambda} - f^*\|_2^2 \geq \sum_{j\in\widehat{m}_\lambda\setminus m^*} Z_j^2\sigma^2 \geq \left(|\widehat{m}_\lambda| - D^*\right)\lambda\sigma^2$$

The last inequality is due to the definition of $\widehat{m}_\lambda$, i.e. $\widehat{m}_\lambda = \left\{j : (Z_j + \boldsymbol{X}_j^T f^*/\|\sigma\boldsymbol{X}_j\|_2)^2 > \lambda\right\}$. In addition, for $a, x\in\mathbb{R}^+$, we have

$$\int_{x-a}^{x} e^{-z^2/2}dz \geq \int_{x}^{x+a} e^{-z^2/2}dz$$

As a consequence, we can deduce

$$\mathbb{P}\left(Z > x-a\right) + \mathbb{P}\left(Z > x+a\right) \geq 2\mathbb{P}\left(Z > x\right)$$

where $Z\sim\mathcal{N}(0,1)$, that is, $\mathbb{P}\left((Z+a)^2 > x^2\right) \geq \mathbb{P}\left(Z^2 > x^2\right)$. Since we have conclude that $|\widehat{m}_\lambda|$ follows a binomial distribution, then its mean is

$$\mathbb{E}\left[|\widehat{m}_\lambda|\right] = \sum_{j=1}^{p}\mathbb{P}\left((Z + \boldsymbol{X}_j^T f^*/\|\sigma\boldsymbol{X}_j\|_2)^2 > \lambda\right) \geq p\mathbb{P}\left(Z^2 > \lambda\right)$$

For $K < 1$ and $D^* \ll p^{1-K} (\log p)^{-1/2}$, we conclude that

$$\begin{aligned}\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_\lambda} - f^*\|_2^2\right] &\geq \lambda\sigma^2 \mathbb{E}\left[|\widehat{m}_\lambda|\right] - D^* \lambda\sigma^2 \\ &\geq \lambda\sigma^2 p \mathbb{P}\left(Z^2 > \lambda\right) - \lambda\sigma^2 D^* \\ &\asymp \lambda\sigma^2 \frac{p^{1-K}}{\sqrt{K\pi \log p}} = p^{1-k}\sigma^2 \sqrt{\frac{4K\log p}{\pi}}\end{aligned}$$

3 Minimax lower bounds

In last section, we have given an upper bound for the risk. It is natural and necessary to consider a lower bound for the best estimation, then we can compare if these two bounds match and hence guarantee that the proposed estimators are optimal.

3.1 Minimax risk

Statistical settings:

- Let $(\mathbb{P}_f)_{f\in\mathscr{F}}$ be a collection of distributions in measure space $(\mathscr{Y}, \mathscr{A})$.
- Define d to be the *distance* on $\mathscr{F}$.
- The risk related to distance is given by for any estimation $\widehat{f} : \mathscr{Y} \to \mathscr{F}$, denote

$$R(\widehat{f}) := \mathbb{E}_f \left[d(\widehat{f}, f)^q \right] \tag{19}$$

  where $q > 0$.

Since we want f to be good to approximate $\widehat{f}$ on the whole class of $\mathscr{F}$, it is meaningless to consider

$$\min_{f\in\mathscr{F}} \mathbb{E}_f \left[d(\widehat{f}, f)^q \right]$$

which can always be zero for some proper f. In other words, we will not focus on point-wise optimality. A popular notion of risk is the *minimax risk* which corresponds to best possible error uniformly over the class $\mathscr{F}$

$$\mathscr{R}^*(\mathscr{F}) := \min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{f\in\mathscr{F}} \mathbb{E}_f \left[d(\widehat{f}, f)^q \right] \tag{20}$$

where $\widehat{f} : \mathscr{Y} \to \mathscr{F}$ denotes all measurable functions satisfies this condition. Our goal is to derive a lower bound for $\mathscr{R}^*(\mathscr{F})$, which is useful if we can find such $\widehat{f}$ tends to this lower bound, then $\widehat{f}$ performs almost as well as the best possible estimator in terms on the max-risk over $\mathscr{F}$.

3.2 A recipe for proving lower bounds

In order to handle this lower bound on an infinite class $\mathscr{F}$, a standard recipe is to replace the maximum over $\mathscr{F}$ by a maximum over a finite set $\{f_1, \cdots, f_N\}$. Once we have discretized the problem, then it is possible to use lower bounds lifted from information theory, in order to get a lower bound on the minimax risk $\mathscr{R}^*(\mathscr{F})$.

Definition 3.2.1 (Kullback–Leibler divergence)**.** *For any two distributions* $\mathbb{P}, \mathbb{Q}$*, define*

$$KL\left(\mathbb{P}, \mathbb{Q}\right) = \begin{cases} \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P} & \text{when } \mathbb{P} \ll \mathbb{Q} \\ +\infty & \text{otherwise} \end{cases} \tag{21}$$

Proposition 3.2.1 (Kullback–Leibler divergence)**.** *The following properties are important.*

- *Non-negative: i.e.* $KL(\mathbb{P},\mathbb{Q}) \geq 0$ *for any two distributions* $\mathbb{P},\mathbb{Q}$*.*
- *Tensorisation: for* $\mathbb{P}_1 \ll \mathbb{Q}_1$ *and* $\mathbb{P}_2 \ll \mathbb{Q}_2$*, we have*

$$\begin{aligned} KL(\mathbb{P}_1 \otimes \mathbb{P}_2, \mathbb{Q}_1 \otimes \mathbb{Q}_2) &= \int \left(\log\left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}\right) + \log\left(\frac{d\mathbb{P}_2}{d\mathbb{Q}_2}\right)\right) d\mathbb{P}_1 d\mathbb{P}_2 \\ &= KL(\mathbb{P}_1,\mathbb{Q}_1) + KL(\mathbb{P}_2,\mathbb{Q}_2) \end{aligned} \tag{22}$$

Proof. In fact, by Jensen's inequality and $x \log x$ is convex, we conclude that

$$KL(\mathbb{P},\mathbb{Q}) = \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}\left[\log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\frac{d\mathbb{P}}{d\mathbb{Q}}\right] \geq \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\right]\log\left(\mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\right]\right) = 0$$

For the second term is just computational trivial. □

Lemma 3.2.1 (Fano)**.** *Let* $(\mathbb{P}_j)_{j=1,\cdots,N}$ *be a set of probability distributions on* Ω*. For any probability distribution* $\mathbb{Q}$ *such that* $\mathbb{P}_j \ll \mathbb{Q}$*, for* $j = 1, \cdots, N$*, we have*

$$\min_{\widehat{J}:\mathscr{Y}\to\{j=1,\cdots,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left(\widehat{J}(Y) \neq j\right) \geq 1 - \frac{1 + \frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_j,\mathbb{Q})}{\log N} \tag{23}$$

where $KL(\mathbb{P}_j,\mathbb{Q})$ *is the Kullback–Leibler divergence between* $\mathbb{P}$ *and* $\mathbb{Q}$*.*

Proof. First, since

$$\min_{\widehat{J}:\mathscr{Y}\to\{j=1,\cdots,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left(\widehat{J}(Y) \neq j\right) = 1 - \max_{\widehat{J}:\mathscr{Y}\to\{j=1,\cdots,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left(\widehat{J}(Y) = j\right)$$

Next, we compute

$$\begin{aligned} \sum_{j=1}^{N} \mathbb{P}_j\left(\widehat{J}(Y) = j\right) &= \int_{\mathscr{Y}} \sum_{j=1}^{N} 1_{\widehat{J}(Y)=j} \frac{d\mathbb{P}_j}{d\mathbb{Q}} d\mathbb{Q} \\ &\leq \int_{\mathscr{Y}} \sum_{j=1}^{N} 1_{\widehat{J}(Y)=j} \max_{k=1,\cdots,N} \frac{d\mathbb{P}_k}{d\mathbb{Q}} d\mathbb{Q} \\ &= \int_{\mathscr{Y}} \max_{k=1,\cdots,N} \frac{d\mathbb{P}_k}{d\mathbb{Q}} d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}\left[\max_{k=1,\cdots,N} \frac{d\mathbb{P}_k}{d\mathbb{Q}}\right] \end{aligned}$$

In addition, the inequality above is an equality for

$$\widehat{J}(y) \in \arg\max_{k=1,\cdots,N} \frac{d\mathbb{P}_k}{d\mathbb{Q}}(y)$$

As a result, we obtain

$$\max_{\widehat{J}:\mathscr{Y}\to\{j=1,\cdots,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left(\widehat{J}(Y) = j\right) \leq \mathbb{E}_{\mathbb{Q}}\left[\max_{k=1,\cdots,N} \frac{d\mathbb{P}_k}{d\mathbb{Q}}\right]$$

Moreover, denote $\varphi(x) = x\log x - x + 1$, which is convex for $x > 0$, then we have

$$\begin{aligned}
\varphi\left(\mathbb{E}_{\mathbb{Q}}\left[\max_{k=1,\cdots,N}\frac{d\mathbb{P}_k}{d\mathbb{Q}}\right]\right) &\leq \mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\max_{k=1,\cdots,N}\frac{d\mathbb{P}_k}{d\mathbb{Q}}\right)\right] \\
&\leq \mathbb{E}_{\mathbb{Q}}\left[\max_{k=1,\cdots,N}\varphi\left(\frac{d\mathbb{P}_k}{d\mathbb{Q}}\right)\right] \\
&\leq \sum_{j=1}^{N}\mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right)\right]
\end{aligned}$$

Besides, we have

$$\mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right)\right] = \mathbb{E}_{\mathbb{Q}}\left[\log\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right)\frac{d\mathbb{P}_j}{d\mathbb{Q}}\right] = KL(\mathbb{P}_j, \mathbb{Q})$$

Finally, for $u > 0$ we have

$$\begin{aligned}
\varphi(Nu) &= Nu\left(\log N + \log u\right) - Nu + 1 \\
&= Nu\log N + N\left(u\log u - u + 1\right) - (N-1) \\
&> Nu\log N - N
\end{aligned}$$

Let $Nu := \mathbb{E}_{\mathbb{Q}}\left[\max_{k=1,\cdots,N} d\mathbb{P}_k/d\mathbb{Q}\right]$, we conclude that

$$\log(N)\mathbb{E}_{\mathbb{Q}}\left[\max_{k=1,\cdots,N}\frac{d\mathbb{P}_k}{d\mathbb{Q}}\right] < N + \sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})$$

which completes our proof. □

Now, for a fixed measurable $\widehat{f} : \mathscr{Y} \to \mathscr{F}$, we define

$$\widehat{J}(y) \in \arg\min_{j=1,\cdots,N} d\left(\widehat{f}(y), f_j\right) \tag{24}$$

where we choose a finite subset $\{f_1, \cdots, f_N\}$ of class $\mathscr{F}$, then we have for $\forall j$:

$$\min_{i\neq k} d(f_i, f_k)1_{\widehat{J}(y)\neq j} \leq d(f_j, f_{\widehat{J}(y)}) \leq d(f_j, \widehat{f}(y)) + d(\widehat{f}(y), f_{\widehat{J}(y)}) \leq 2d(f_j, \widehat{f}(y))$$

As a result, we conclude that

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{f_j}\left[d(f_j, \widehat{f}(Y))^q\right] &\geq \frac{1}{2^q}\min_{i\neq k} d(f_i, f_k)^q \frac{1}{N}\sum_{j=1}^{N}\mathbb{P}_{f_j}\left(\widehat{J}(Y) \neq j\right) \\
&\geq \frac{1}{2^q}\min_{i\neq k} d(f_i, f_k)^q \min_{\widehat{J}:\mathscr{Y}\to\{j=1,\cdots,N\}}\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}_{f_j}\left(\widehat{J}(Y) \neq j\right)
\end{aligned}$$

Corollary 3.2.1 (Lower bound for discrete problem). *For any $\{f_1, \cdots, f_N\} \subset \mathscr{F}$ and for any probability distribution $\mathbb{Q}$ such that $\mathbb{P}_{f_j} \ll \mathbb{Q}$, for $j = 1, \cdots, N$, we have*

$$\begin{aligned} &\min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{f_j}\left[d(f_j, \widehat{f}(Y))^q\right] \\ &\qquad\qquad \geq 2^{-q}\left(1 - \frac{1+\frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})}{\log N}\right)\min_{i\neq k} d(f_i, f_k)^q \end{aligned} \tag{25}$$

where $KL(\mathbb{P}_j, \mathbb{Q})$ denotes the Kullback-Leibler divergence between $\mathbb{P}_j$ and $\mathbb{Q}$.

Until now, for any finite subsets of class $\mathscr{F}$, we have a lower bound for

$$\begin{aligned} \mathscr{R}^*(\mathscr{F}) &= \min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{f\in\mathscr{F}} \mathbb{E}_f\left[d(\widehat{f}, f)^q\right] \\ &\geq \min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{j=1,\cdots,N} \mathbb{E}_{f_j}\left[d(\widehat{f}, f_j)^q\right] \\ &\geq \min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{f_j}\left[d(f_j, \widehat{f}(Y))^q\right] \\ &\geq 2^{-q}\left(1 - \frac{1+\frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})}{\log N}\right)\min_{i\neq k} d(f_i, f_k)^q \end{aligned}$$

Hence, we need to find a good discretion subset such that

- $\min_{i\neq k} d(f_i, f_k)^q$ as large as possible.
- $(\log N)^{-1}\left(1 + N^{-1}\sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})\right)$ is always bounded from above to 1.

For the reason, we hope the lower bound of $\mathscr{R}^*(\mathscr{F})$ is as large as possible, it is meaningless if it is near to zero since $\mathscr{R}^*(\mathscr{F}) \geq 0$. All the art is to find a balance between above two points. First find $f_1, \cdots, f_N$ such that

$$\frac{1+\frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})}{\log N} \leq \frac{1}{2}$$

Lemma 3.2.2 (Birgé's Inequality). *Let us consider a family $(A_i)_{i=1,\cdots,N}$ of disjointed events, and a collection $(\mathbb{P}_i)_{i=1,\cdots,N}$ of probability measures. Then, we have*

$$\min_{i=1,\cdots,N} \mathbb{P}_i(A_i) \leq \max\left\{\frac{2e}{2e+1}, \frac{\max_{i\neq j} KL(\mathbb{P}_i, \mathbb{P}_j)}{\log N}\right\} \tag{26}$$

Proof. Assume that $\mathbb{P}_i \ll \mathbb{P}_j$ for $i \neq j$, or $KL(\mathbb{P}_i, \mathbb{P}_j) = +\infty$ which is a trivial case. In particular, we will show that

$$\mathbb{E}_{\mathbb{P}_2}[X] - \log \mathbb{E}_{\mathbb{P}_1}\left[e^X\right] \leq KL(\mathbb{P}_2, \mathbb{P}_1)$$

where X is any bounded random variable. By Jensen's inequality, we have

$$\begin{aligned}-\log \mathbb{E}_{\mathbb{P}_1}\left[e^X\right] &= -\log\left(\int e^X \frac{d\mathbb{P}_1}{d\mathbb{P}_2} d\mathbb{P}_2\right) \leq -\int \log\left(e^X \frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_2 \\ &= -\mathbb{E}_{\mathbb{P}_2}\left[X\right] + KL(\mathbb{P}_2, \mathbb{P}_1)\end{aligned}$$

Now, set $m := \min_{i=1,\cdots,N} \mathbb{P}_i(A_i)$ and for fixed $i \leq N-1$, define $X = 1_{A_i} \log\left(m/q\right)$ with $q = (1-m)/(N-1)$. Hence, we conclude that

$$\mathbb{P}_i(A_i)\log(m/q) - \log \mathbb{E}_{\mathbb{P}_N}\left[\left(\frac{m}{q}\right)^{1_{A_i}}\right] \leq KL(\mathbb{P}_i, \mathbb{P}_N)$$

In addition, we have

$$\log \mathbb{E}_{\mathbb{P}_N}\left[\left(\frac{m}{q}\right)^{1_{A_i}}\right] = \log\left(\mathbb{P}_N(A_i)\left(\frac{m}{q}-1\right)+1\right) \leq \mathbb{P}_N(A_i)\left(\frac{m}{q}-1\right) \leq \frac{m}{q}\mathbb{P}_N(A_i)$$

Since $\mathbb{P}_i(A_i) \geq m$ for $i = 1, \cdots, N-1$, averaging over $i \in \{1, \cdots, N-1\}$ and obtain

$$\frac{1}{N-1}\sum_{i=1}^{N-1} \mathbb{P}_i(A_i)\log\left(\frac{m}{q}\right) - \frac{m}{q(N-1)}\sum_{i=1}^{N-1}\mathbb{P}_N(A_i) \leq \frac{1}{N-1}\sum_{i=1}^{N-1} KL(\mathbb{P}_i, \mathbb{P}_N)$$

Because

$$\frac{1}{N-1}\sum_{i=1}^{N-1}\mathbb{P}_i(A_i) \geq m \quad \text{and} \quad \sum_{i=1}^{N-1}\mathbb{P}_N(A_i) \leq 1 - \mathbb{P}_N(A_N) \leq 1-m$$

We finally obtain

$$m\log\left(\frac{m(N-1)}{e(1-m)}\right) \leq \frac{1}{N-1}\sum_{i=1}^{N-1} KL(\mathbb{P}_i, \mathbb{P}_N)$$

If $m > 2e/(2e+1)$, we conclude that

$$\log\left(\frac{m(N-1)}{e(1-m)}\right) \geq \log(2(N-1)) \geq \log N$$

Hence,

$$m\log N \leq \frac{1}{N-1}\sum_{i=1}^{N-1} KL(\mathbb{P}_i, \mathbb{P}_N) \leq \max_{i\neq j} KL(\mathbb{P}_i, \mathbb{P}_j)$$

And for $m \leq 2e/(2e+1)$, it is trivial, which completes our proof. □

Corollary 3.2.2. *For any $\{f_1, \cdots, f_N\} \subset \mathscr{F}$ such that*

$$\max_{j\neq k} KL\left(\mathbb{P}_{f_j}, \mathbb{P}_{f_k}\right) \leq \frac{2e}{2e+1}\log N \tag{27}$$

we have

$$\min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{j=1,\cdots,N} \mathbb{E}_{f_j}\left[d(\widehat{f}, f_j)^q\right] \geq \frac{1}{2^q(2e+1)}\min_{j\neq k} d(f_j, f_k)^q \tag{28}$$

Proof. Denote $A_j := \left\{\widehat{J}(Y) = j\right\}$ which are disjoint and by Lemma 3.2.2, it concludes that

$$\min_{i=1,\cdots,N} \mathbb{P}_{f_i}(A_i) \leq \max\left\{\frac{2e}{2e+1}, \frac{\max_{i\neq j} KL(\mathbb{P}_{f_i}, \mathbb{P}_{f_j})}{\log N}\right\} \leq \frac{2e}{2e+1}$$

Next, by Lemma 3.2.1, it concludes that

$$\min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{j=1,\cdots,N} \mathbb{P}_{f_j}\left(\widehat{J}(Y) \neq j\right) \geq \frac{1}{2e+1}$$

Finally, we have

$$\min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{j=1,\cdots,N} \mathbb{E}_{f_j}\left[d(\widehat{f}, f_j)^q\right] \geq \frac{1}{2^q} \min_{i\neq k} d(f_i, f_k)^q \min_{\widehat{f}:\mathscr{Y}\to\mathscr{F}} \max_{j=1,\cdots,N} \mathbb{P}_{f_j}\left(\widehat{J}(Y) \neq j\right)$$

which completes our proof. □

3.3 Minimax risk for coordinate sparse regression

In this subsection, we will focus on the example of coordinate sparsity. For $D \in \{1, \cdots, p\}$, we define

$$V_D(\boldsymbol{X}) := \{\boldsymbol{X}\beta : \beta \in \mathbb{R}^p, |\beta|_0 = D\} \tag{29}$$

For the distributions class, we denote $\mathbb{P}_f \sim \mathcal{N}(f, \sigma^2 \boldsymbol{I}_n)$, the distance $d(f_1, f_2) = \|f_1 - f_2\|_2$ and $q = 2$. Then the Kullback–Leibler divergence of $\mathbb{P}_{f_1}$ and $\mathbb{P}_{f_2}$ is

$$KL(\mathbb{P}_{f_1}, \mathbb{P}_{f_2}) = \frac{\|f_1 - f_2\|_2^2}{2\sigma^2}$$

Define an integer $D_{\max} \leq p/2$, we introduce the restricted isometry constants

$$\underline{c}_{\boldsymbol{X}} := \inf_{\beta:|\beta|_0 \leq 2D_{\max}} \frac{\|\boldsymbol{X}\beta\|_2}{\|\beta\|_2} \leq \sup_{\beta:|\beta|_0 \leq 2D_{\max}} \frac{\|\boldsymbol{X}\beta\|_2}{\|\beta\|_2} := \bar{c}_{\boldsymbol{X}} \tag{30}$$

We consider the minimax risk on $V_D(\boldsymbol{X})$ is defined by

$$\boldsymbol{R}^*[\boldsymbol{X}, D] := \inf_{\widehat{f}} \sup_{f \in V_D(\boldsymbol{X})} \mathbb{E}_f\left[\|\widehat{f} - f\|_2^2\right] \tag{31}$$

Lemma 3.3.1. *For any $D \leq p/5$, there exists $\mathscr{C} \subset \{0,1\}^p_D := \{\beta \in \{0,1\}^p : |\beta|_0 = D\}$ such that*

$$|\beta_j - \beta_k|_0 \geq D \quad \text{for } \forall \beta_j \neq \beta_k \in \mathscr{C} \tag{32}$$

then we can conclude that

$$\log|\mathscr{C}| \geq \frac{D}{2} \log\left(\frac{p}{5D}\right) \tag{33}$$

Proof. First, we will show that

$$\{0,1\}^p_D = \bigcup_{\beta \in \mathscr{C}} \{x \in \{0,1\}^p_D : |x-\beta|_0 \le D\}$$

In fact, it is trivial to see the right side is a subset of left side. Now, suppose any $y \in \{0,1\}^p_D$, then we have

$$y \in \{x \in \{0,1\}^p_D : |x-y|_0 \le D\}$$

As a consequence, we deduce that

$$C^D_p \le \sum_{\beta \in \mathscr{C}} |\{x \in \{0,1\}^p_D : |x-\beta|_0 \le D\}| \le |\mathscr{C}| \max_{\beta \in \mathscr{C}} |B(\beta, D)|$$

where $B(\beta, D) := \{x \in \{0,1\}^p_D : |x-\beta|_0 \le D\}$. On the other hand, we have

$$|B(\beta, D)| = \sum_{i=0}^{[D/2]} C_D^{D-i} C^i_{p-D} \le C_p^{[D/2]} 2^D$$

Hence, we obtain

$$\frac{|B(\beta, D)|}{C^D_p} \le 2^D \frac{C_p^{[D/2]}}{C^D_p} \le 2^D \left(\frac{D}{p-D+1}\right)^{D-[D/2]} \le \left(\frac{5D}{p}\right)^{D/2}$$

which completes our proof. □

Theorem 3.3.1 (Minimax risk for coordinate-sparse regression)**.** *Let us fix some $D_{\max} \le p/5$. For any $D \le D_{\max}$, we have the lower bound*

$$\boldsymbol{R}^*[\boldsymbol{X}, D] \ge \frac{e}{4(2e+1)^2} \left(\frac{\underline{c}_{\boldsymbol{X}}}{\bar{c}_{\boldsymbol{X}}}\right)^2 D \log\left(\frac{p}{5D}\right) \sigma^2 \tag{34}$$

Proof. By Lemma 3.3.1, we can obtain $\mathscr{C} = \{\beta_1, \cdots, \beta_N\}$, and denote $f_j := c\boldsymbol{X}\beta_j$, where c is a parameters which will be determined later. Then

$$\max_{j \ne k} KL(\mathbb{P}_j, \mathbb{P}_k) = \max_{j \ne k} \frac{c^2 \|\boldsymbol{X}(\beta_j - \beta_k)\|_2^2}{2\sigma^2} \le \frac{c^2}{2\sigma^2} \bar{c}^2_{\boldsymbol{X}} \max_{j \ne k} \|\beta_j - \beta_k\|^2_2 \underbrace{\le}_{\text{hope}} \frac{2e}{2e+1} \log N$$

Since $\|\beta_j - \beta_k\|^2_2 = |\beta_j - \beta_k|_0 \le 2D$ and we let

$$c^2 := \frac{\sigma^2}{\bar{c}^2_{\boldsymbol{X}} D} \frac{2e}{2e+1} \log N$$

In addition, we have

$$\begin{aligned} \|f_j - f_k\|^2_2 &= c^2 \|\boldsymbol{X}(\beta_j - \beta_k)\|^2_2 \\ &\ge \frac{\sigma^2}{\bar{c}^2_{\boldsymbol{X}} D} \frac{2e}{2e+1} (\log N) \underline{c}^2_{\boldsymbol{X}} \|\beta_j - \beta_k\|^2_2 \\ &\ge \left(\frac{\underline{c}_{\boldsymbol{X}}}{\bar{c}_{\boldsymbol{X}}}\right)^2 \sigma^2 \frac{2e}{2e+1} D \log\left(\frac{p}{5D}\right) \end{aligned}$$

The lower bound for $\|\beta_j - \beta_k\|_2^2$ and $\log N$ is according to Lemma 3.3.1. Finally, by Corollary 3.2.2, we conclude that

$$\boldsymbol{R}^*[\boldsymbol{X}, D] \geq \min_{\widehat{f}} \max_{j=1,\cdots,N} \mathbb{E}_{f_i}\left[\|\widehat{m} - f_i\|_2^2\right] \geq \frac{1}{4(2e+1)} \min_{j \neq k} \|f_j - f_k\|_2^2$$

which completes our proof. □

3.4 Some other minimax lower bounds

Let d be a distance on $\mathbb{R}^p$ and fix $q \geq 1$. We write $\mathbb{P}_f$ for the Gaussian distribution $\mathcal{N}(f, \sigma^2 \boldsymbol{I}_n)$ and the expectation $\mathbb{E}_f$ will refer to the expectation when the vector of observations Y is distributed according to $\mathbb{P}_f$.

Let $\mathscr{C}$ be any finite subset of $\mathbb{R}^p$, then we can show that if the following condition is valid

$$\max_{\beta \neq \beta' \in \mathscr{C}} \|\boldsymbol{X}(\beta - \beta')\|_2^2 \leq \frac{4e}{2e+1} \sigma^2 \log|\mathscr{C}|$$

We can conclude that

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathscr{C}} \mathbb{E}_{\boldsymbol{X}\beta}\left[d(\beta, \widehat{\beta})^q\right] \geq \frac{1}{2^q} \frac{1}{2e+1} \min_{\beta \neq \beta' \in \mathscr{C}} d(\beta, \beta')^q$$

Proof. First, as in Corollary 3.2.1, we have

$$2d(\beta_j, \widehat{\beta}) \geq \min_{\beta_i \neq \beta_k \in \mathscr{C}} d(\beta_i, \beta_k) 1_{\widehat{J}(y) \neq j}$$

where

$$\widehat{J}(y) \in \arg \min_{j=1,\cdots,|\mathscr{C}|} d(\widehat{\beta}, \beta_j)$$

Hence, we have

$$\begin{aligned} \inf_{\widehat{\beta}} \sup_{\beta_j \in \mathscr{C}} \mathbb{E}_{\boldsymbol{X}\beta_j}\left[d(\beta_j, \widehat{\beta})^q\right] &\geq \inf_{\widehat{\beta}} \frac{1}{|\mathscr{C}|} \sum_{\beta \in \mathscr{C}} \mathbb{E}_{\boldsymbol{X}\beta}\left[d(\beta, \widehat{\beta})^q\right] \\ &\geq \inf_{\widehat{\beta}} \frac{1}{2^q |\mathscr{C}|} \min_{\beta \neq \beta' \in \mathscr{C}} d(\beta, \beta')^q \sum_{j=1}^{|\mathscr{C}|} \mathbb{P}_{\beta_j}\left(\widehat{J}(Y) \neq j\right) \end{aligned}$$

In addition, since

$$\sum_{j=1}^{|\mathscr{C}|} \mathbb{P}_{\beta_j}\left(\widehat{J}(Y) \neq j\right) = |\mathscr{C}| - \sum_{j=1}^{|\mathscr{C}|} \mathbb{P}_{\beta_j}\left(\widehat{J}(Y) = j\right) \geq |\mathscr{C}| \left(1 - \min_{\beta_j \in \mathscr{C}} \mathbb{P}_{\beta_j}\left(\widehat{J}(Y) = j\right)\right)$$

By Lemma 3.2.2, it concludes that

$$\min_{\beta_j \in \mathscr{C}} \mathbb{P}_{\beta_j}\left(\widehat{J}(Y) = j\right) \leq \max\left\{\frac{2e}{2e+1}, \frac{\max_{\beta_i \neq \beta_j \in \mathscr{C}} \|\boldsymbol{X}(\beta_i - \beta_j)\|_2^2}{2\sigma^2 \log|\mathscr{C}|}\right\} \leq \frac{2e}{2e+1}$$

The last inequality is valid due to the given assumption. Finally, we obtain

$$\inf_{\widehat{\beta}} \sup_{\beta_j \in \mathscr{C}} \mathbb{E}_{\boldsymbol{X}\beta_j} \left[d(\beta_j, \widehat{\beta})^q \right] \geq \frac{1}{2^q} \frac{1}{2e+1} \min_{\beta \neq \beta' \in \mathscr{C}} d(\beta, \beta')^q$$

which completes our proof. □

Next, we fix $D_{\max} \leq p/5$. For $D \leq D_{\max}$, we set

$$r^2 := \frac{e}{2e+1} \frac{\sigma^2}{\bar{c}_{\boldsymbol{X}}^2} \log \left(\frac{p}{5D} \right)$$

consider below the set $\mathscr{C}_r = \{ r\beta : \beta \in \mathscr{C} \}$, where $\mathscr{C}$ satisfies properties in Lemma 3.3.1, and the distance d induced by the q-norm. Then it concludes that

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathscr{C}_r} \mathbb{E}_{\boldsymbol{X}\beta} \left[\|\beta - \widehat{\beta}\|_q^q \right] \geq \frac{r^q D}{2^q (2e+1)}$$

Proof. By Lemma 3.3.1, for $\beta_i \neq \beta_j \in \mathscr{C}_r$, we have

$$\|\beta_i - \beta_j\|_q^q = r^q |\beta_i - \beta_j|_0 \geq r^q D$$

Since we have already concluded that

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathscr{C}_r} \mathbb{E}_{\boldsymbol{X}\beta} \left[\|\beta - \widehat{\beta}\|_q^q \right] \geq \frac{1}{2^q (2e+1)} \min_{\beta \neq \beta' \in \mathscr{C}} \|\beta - \beta'\|_q^q \geq \frac{r^q D}{2^q (2e+1)}$$

which completes our proof. □

Finally, for all $q \geq 1$, we have the following lower bound.

$$\inf_{\widehat{\beta}} \sup_{\beta : |\beta|_0 = D} \mathbb{E}_{\boldsymbol{X}\beta} \left[\|\beta - \widehat{\beta}\|_q^q \right] \geq \frac{e^{q/2}}{2^q (2e+1)^{1+q/2}} \left(\frac{\sigma}{\bar{c}_{\boldsymbol{X}}} \right)^q D \left(\log \left(\frac{p}{5D} \right) \right)^{p/2}$$

4 Convex Relaxation

In this section, we will try to solve the model selection problem with minimization criterion.

$$\widehat{m} := \inf_{m \in \mathscr{M}} \left\{ \|Y - f_m\|_2^2 + \mathrm{pen}(m)\sigma^2 \right\} \tag{35}$$

which can be impossible to realize in practice when $|\mathscr{M}|$ is very large. For example, if we consider the coordinate-sparsity regression of $\mathscr{M} := \mathscr{P}(\{1, \cdots, p\})$, we must evaluate $|\mathscr{M}| = 2^p$ quantities. Before giving a useful method to deal with such dilemma, let's introduce some backgrounds about convex optimization.

4.1 Reminder on Convex Multivariate Functions

For a function $F : \mathbb{R}^n \to \mathbb{R}$, which is convex if its domain $\mathrm{dom}(F) \subset \mathbb{R}^n$ is a convex set and for $\forall x, y \in \mathrm{dom}(F)$

$$F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)F(y) \quad \forall \alpha \in [0, 1] \tag{36}$$

In addition, if F is differentiable, we have the following property

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle \quad \forall x, y \in \mathrm{dom}(F) \tag{37}$$

Hence, we can introduce the general form of gradient, that is,

Definition 4.1.1 (Sub-differential)**.** *For a convex function F, the sub-differential of F is defined by*

$$\partial F(x) := \{\omega \in \mathbb{R}^n : F(y) \geq F(x) + \langle \omega, y - x \rangle, \forall y \in \mathrm{dom}(F)\} \tag{38}$$

Lemma 4.1.1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex function, then*

- *$\partial F(x) \neq \emptyset$ for $\forall x \in \mathrm{dom}(F)$.*
- *if F is differentiable at x, $\partial F(x) = \{\nabla F(x)\}$.*

Proof. First, define the epigraph of F by

$$\{(x, y) \in \mathrm{dom}(F) \times \mathbb{R} : y \in [F(x), \infty)\}$$

which is convex Suppose $\exists x_0 \in \mathrm{dom}(F)$ such that $\partial F(x_0) = \emptyset$, i.e. for $\forall \omega \in \mathbb{R}^n$, there exists $y_\omega \in \mathrm{dom}(F)$ such that

$$F(y_\omega) < F(x_0) + \langle \omega, y_\omega - x_0 \rangle$$

□

Proposition 4.1.1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex function, then*

- **monotonicity:** $\forall \omega_x \in \partial F(x)$ *and* $\forall \omega_y \in \partial F(y)$

$$\langle \omega_x - \omega_y, x - y \rangle \geq 0 \tag{39}$$

- **minimum:**

$$x^* \in \arg \min_{x \in \mathrm{dom}(F)} F(x) \quad \Leftrightarrow \quad \mathbf{0} \in \partial F(x) \tag{40}$$

Proof. First, by Definition 4.1.1, we conclude that

$$\begin{aligned} F(y) &\geq F(x) + \langle \omega_x, y - x \rangle \\ F(x) &\geq F(y) + \langle \omega_y, x - y \rangle \end{aligned}$$

Adding these two equations and we obtain the first term. For the second,

$$x^* \in \arg \min_{x \in \mathrm{dom}(F)} F(x) \quad \Leftrightarrow \quad F(y) \geq F(x^*) + \langle \mathbf{0}, y - x^* \rangle$$

which completes our proof. □

Example 4.1.1. *We introduce sub-differentials of some special norms. For L^1-norm, which is $|x|_1 = \sum_j |x_j|$, its sub-differential is given by*

$$\partial |x|_1 := \{\omega \in \mathbb{R}^n : \omega_j = \mathrm{sgn}(x_j) \text{ for } x_j \neq 0, \omega_j \in [-1, 1] \text{ otherwise}\} \tag{41}$$

or equivalently,

$$\partial |x|_1 = \{\phi : \langle \phi, x \rangle = |x|_1 \text{ and } |\phi|_\infty \leq 1\} \tag{42}$$

Since the L^∞-norm is the dual of L^1-norm, it concludes that

$$\partial |x|_\infty := \{\varphi : \langle \varphi, x \rangle = |x|_\infty \text{ and } |\varphi|_1 \leq 1\} \tag{43}$$

Proof. First, since for $\dim(x) = 1$, $|x|_1 = |x|$ and

$$\partial |x| = \begin{cases} \mathrm{sgn}(x) & x \neq 0 \\ [-1, 1] & x = 0 \end{cases}$$

which concludes that (41). In addition, if ϕ satisfies (42), then for $\forall y \in \mathbb{R}^n$, it has

$$|y|_1 \geq \langle \phi, y \rangle = \langle \phi, x \rangle + \langle \phi, y - x \rangle = |x|_1 + \langle \phi, y - x \rangle$$

i.e. $\phi \in \partial |x|_1$. On the other hand, if $\omega \in \partial |x|_1$, then it concludes that

$$\begin{cases} 2|x|_1 \geq |x|_1 + \langle \omega, x \rangle \\ 0 \geq |x|_1 + \langle \omega, -x \rangle \end{cases}$$

which concludes that $\langle \omega, x \rangle = |x|_1$ and since $\exists \xi \in \mathbb{R}^n$ such that $|\omega|_\infty = \langle \omega, \xi \rangle$ and $|\xi|_1 = 1$, then we have

$$|x|_1 + |\xi|_1 \geq |x + \xi|_1 \geq |x|_1 + \langle \phi, \xi \rangle = |x|_1 + |\omega|_\infty$$

which concludes that $|\omega|_\infty \leq 1$. Finally, suppose $\varphi \in \partial |x|_\infty$, by similar argument, it can also conclude that $\langle \varphi, x \rangle = |x|_\infty$ and $|\varphi|_1 \leq 1$. □

4.2 Lasso Estimator

In order to solve model selection problem, we will modify it into a convex criterion, which can be amenable to numerical computation. For sparse linear regression,

$$Y = \boldsymbol{X}\beta^* + \boldsymbol{\epsilon}$$

where $|\beta^*|_0$ is small. In this section, we assume that the columns of $\boldsymbol{X}$ are **normalized**, i.e. $\|\boldsymbol{X}_j\|_2 = 1$ for $j = 1, \cdots, p$. In addition, for $m \in \mathscr{M}$, denote $\pi_m \propto \exp(-|m| \log p)$, then choose the penalty $\text{pen}(m)\sigma^2 = \lambda|m|$, where $\lambda := K\left(1 + \sqrt{2\log p}\right)^2 \sigma^2$, so the model selection criterion is

$$\widehat{m} := \arg\min_{m \in \mathscr{M}} \left\{ \|Y - \widehat{f}_m\|_2^2 + \lambda|m| \right\}$$

where $\widehat{f}_m := \text{Proj}_{S_m} Y$ and $S_m := \text{span}\{\boldsymbol{X}_j : j \in m\} = \{\boldsymbol{X}\beta : \text{supp}(\beta) = m\}$. Next, denote

$$\widehat{\beta}_m := \arg\min_{\text{supp}(\beta) = m} \|Y - \boldsymbol{X}\beta\|_2^2$$

Then we conclude that $\widehat{f}_m \equiv \boldsymbol{X}\widehat{\beta}_m$ and hence deduce that

$$\widehat{m} = \arg\min_{\text{supp}(\beta) = m} \left\{ \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda|\beta|_0 \right\}$$

and

$$\widehat{\beta}_{\widehat{m}} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda|\beta|_0 \right\}$$

Now, for the term of $\|Y - \boldsymbol{X}\beta\|_2^2$, which is convex, so it is easy to deal with. The trouble appears on the term of $\lambda|\beta|_0$, which is highly non-smooth and non-convex. The main ideal for deriving from above optimization project is to replace $|\beta|_0$ by $|\beta|_1$, i.e.

$$\widehat{\beta}_\lambda := \arg\min_{\beta \in \mathbb{R}^p} \mathscr{L}(\beta) \quad \text{where } \mathscr{L}(\beta) := \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda|\beta|_1 \tag{44}$$

The term $\widehat{\beta}_\lambda$ is called the *Lasso* estimator, which may not be unique. However, the deriving estimator $\widehat{f}_\lambda := \boldsymbol{X}\widehat{\beta}_\lambda$ can be always unique.

Remark 4.2.1. *The reason of using L^1-norm instead of L^p-norm, where $1 < p < \infty$ will be shown in the last subsection of this part. In fact, L^p-norm would not lead to variable selection.*

Consider the solution of (44), whose sub-differential has the following form.

$$\partial\mathscr{L}_\lambda(\beta) := \left\{ -2\boldsymbol{X}^T(Y - \boldsymbol{X}\beta) + \lambda z : z \in \partial|\beta|_1 \right\}$$

By the minimum property of sub-differential, we conclude that $\mathbf{0} \in \partial\mathscr{L}_\lambda(\widehat{\beta}_\lambda)$, i.e. these exists $z_\lambda \in \partial|\beta_\lambda|_1$ such that

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\beta}_\lambda = \boldsymbol{X}^T Y - \frac{\lambda}{2} z_\lambda \tag{45}$$

In general, above equation does not have explicit solution, but if $\boldsymbol{X}$ is orthogonal, we have the following result.

$$\widehat{\beta}_\lambda = \boldsymbol{X}^T Y - \frac{\lambda}{2} z_\lambda$$

Moreover, since $z_\lambda \in \partial|\widehat{\beta}_\lambda|_1$, which has been discussed in (41), hence the solution should be

$$\begin{cases} \widehat{\beta}_\lambda^{(j)} + \frac{\lambda}{2}\mathrm{sgn}\left(\widehat{\beta}_\lambda^{(j)}\right) = \boldsymbol{X}_j^T Y & \text{for } \widehat{\beta}_\lambda^{(j)} \neq 0 \\ z_\lambda^{(j)} = \frac{2}{\lambda}\boldsymbol{X}_j^T Y & \text{otherwise} \end{cases} \tag{46}$$

To solve $\widehat{\beta}_\lambda^{(j)}$ when it is non-zero, consider the inverse function of $x + \lambda \mathrm{sgn}(x)/2$, it concludes that

$$\widehat{\beta}_\lambda^{(j)} = \left[\boldsymbol{X}_j^T Y - \frac{\lambda}{2}\mathrm{sgn}\left(\boldsymbol{X}_j^T Y\right)\right] 1_{|\boldsymbol{X}_j^T Y| \geq \lambda/2}$$

As a result, we conclude that

$$\begin{cases} \widehat{\beta}_\lambda^{(j)} = \boldsymbol{X}_j^T Y - \frac{\lambda}{2}\mathrm{sgn}\left(\boldsymbol{X}_j^T Y\right) & \text{if } |\boldsymbol{X}_j^T Y| \geq \frac{\lambda}{2} \\ z_\lambda^{(j)} = \frac{2}{\lambda}\boldsymbol{X}_j^T Y & \text{if } |\boldsymbol{X}_j^T Y| < \frac{\lambda}{2} \end{cases} \tag{47}$$

For the compact form, it has

$$\widehat{\beta}_\lambda^{(j)} = \boldsymbol{X}_j^T Y \left(1 - \frac{2}{\lambda|\boldsymbol{X}_j^T Y|}\right)_+ \tag{48}$$

4.3 Statistical Analysis of Lasso Estimation

As preliminary, we will the following concept.

Definition 4.3.1 (Compatibility constant)**.** *Account for (local) orthogonality, define*

$$\kappa(\beta) := \min\left\{\frac{\sqrt{|\beta|_0}\|\boldsymbol{X}v\|_2}{|v_s|_1} : v \in \mathscr{C}(\beta)\right\} \tag{49}$$

where $s = \mathrm{supp}(\beta)$ *and* $\mathscr{C}(\beta) := \{v \in \mathbb{R}^p : 5|v_s|_1 > |v_s^c|_1\}$.

Theorem 4.3.1 (Deterministic bound)**.** *For any* $\lambda > 3|\boldsymbol{X}^T\boldsymbol{\epsilon}|_\infty$*, we have*

$$\|\boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^*\|_2^2 \leq \inf_{\beta \in \mathbb{R}^p \backslash \{\mathbf{0}\}} \left\{\|\boldsymbol{X}\beta - \boldsymbol{X}\beta^*\|_2^2 + \frac{\lambda^2}{\kappa^2(\beta)}|\beta|_0\right\} \tag{50}$$

Proof. Since $\mathbf{0} \in \partial\mathscr{L}(\widehat{\beta}_\lambda)$, i.e. $\exists z_\lambda \in \partial|\widehat{\beta}_\lambda|$ such that

$$2\boldsymbol{X}^T\left(\boldsymbol{X}\widehat{\beta}_\lambda - Y\right) + \lambda z_\lambda = \mathbf{0}$$

where $Y = \boldsymbol{X}\beta^* + \boldsymbol{\epsilon}$, then we obtain for $\forall \beta \in \mathbb{R}^p \backslash \{\mathbf{0}\}$

$$2\left\langle \boldsymbol{X}^T\left(\boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^* - \boldsymbol{\epsilon}\right), \widehat{\beta}_\lambda - \beta\right\rangle + \lambda\left\langle z_\lambda, \widehat{\beta}_\lambda - \beta\right\rangle = 0$$

i.e.

$$2\left\langle \boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^*, \boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta \right\rangle - 2\left\langle \boldsymbol{X}^T\boldsymbol{\epsilon}, \widehat{\beta}_\lambda - \beta \right\rangle + \lambda\left\langle z_\lambda, \widehat{\beta}_\lambda - \beta \right\rangle = 0$$

By monotonicity property of sub-differential, it concludes that, for $\forall z \in \partial|\beta|_1$

$$\langle z - z_\lambda, \beta - \widehat{\beta}_\lambda \rangle \geq 0$$

i.e. $\langle z, \widehat{\beta}_\lambda - \beta \rangle \leq \langle z_\lambda, \widehat{\beta}_\lambda - \beta \rangle$ and hence

$$\mathcal{A}(\widehat{\beta}_\lambda) := 2\left\langle \boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^*, \boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta \right\rangle \leq 2\left\langle \boldsymbol{X}^T\boldsymbol{\epsilon}, \widehat{\beta}_\lambda - \beta \right\rangle - \lambda\langle z, \widehat{\beta}_\lambda - \beta \rangle$$

Next, we will show that

$$\mathcal{A}(\widehat{\beta}_\lambda) \leq \frac{\lambda}{3}\left(5\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1 - \left|\left(\widehat{\beta}_\lambda - \beta\right)_{s^c}\right|_1\right) \leq 2\lambda\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1$$

where $s = \text{supp}(\beta)$. First, define z by

$$\begin{cases} z_j := \text{sgn}(\beta^{(j)}) & j \in s \\ z_j := \text{sgn}(\widehat{\beta}_\lambda^{(j)} - \beta^{(j)}) & j \in s^c \end{cases}$$

By (41), it concludes that $z \in \partial|\beta|_1$, hence

$$\begin{aligned} \mathcal{A}(\widehat{\beta}_\lambda) &\leq |\boldsymbol{X}^T\boldsymbol{\epsilon}|_\infty \left|\widehat{\beta}_\lambda - \beta\right|_1 - \lambda\left\langle z, \left(\widehat{\beta}_\lambda - \beta\right)_s \right\rangle - \lambda\left\langle z, \left(\widehat{\beta}_\lambda - \beta\right)_{s^c} \right\rangle \\ &\leq \frac{2\lambda}{3}\left|\widehat{\beta}_\lambda - \beta\right|_1 + \lambda\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1 - \lambda\left|\left(\widehat{\beta}_\lambda - \beta\right)_{s^c}\right|_1 \\ &\leq \frac{\lambda}{3}\left(5\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1 - \left|\left(\widehat{\beta}_\lambda - \beta\right)_{s^c}\right|_1\right) \\ &\leq 2\lambda\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1 \end{aligned}$$

Finally, if

$$\mathcal{A}(\widehat{\beta}_\lambda) = \left\|\boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^*\right\|_2^2 + \left\|\boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta\right\|_2^2 - \|\boldsymbol{X}\beta - \boldsymbol{X}\beta^*\|_2^2 \leq 0$$

if concludes that

$$\left\|\boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^*\right\|_2^2 \leq \|\boldsymbol{X}\beta - \boldsymbol{X}\beta^*\|_2^2$$

which already concludes our conclusion. On the other hand, if $\mathcal{A}(\widehat{\beta}_\lambda) > 0$, we have known that

$$0 < \mathcal{A}(\widehat{\beta}_\lambda) \leq \frac{\lambda}{3}\left(5\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1 - \left|\left(\widehat{\beta}_\lambda - \beta\right)_{s^c}\right|_1\right)$$

i.e. $\widehat{\beta}_\lambda - \beta \in \mathscr{C}(\beta)$, then it concludes that

$$\mathcal{A}(\widehat{\beta}_\lambda) \leq 2\lambda\left|\left(\widehat{\beta}_\lambda - \beta\right)_s\right|_1 \leq 2\frac{\sqrt{|\beta|_0}}{\kappa(\beta)}\|\boldsymbol{X}(\widehat{\beta}_\lambda - \beta)\|_2 \leq \|\boldsymbol{X}(\widehat{\beta}_\lambda - \beta)\|_2^2 + \frac{|\beta|_0}{\kappa^2(\beta)}$$

which completes our proof. □

Corollary 4.3.1 (Risk bound for the Lasso)**.** *Assume that all the columns of $\boldsymbol{X}$ have norm 1 and that the noise $(\epsilon_i)_{i=1,\cdots,n}$ is i.i.d. with $\mathcal{N}(0,\sigma^2)$ distribution. Then, for any $L>0$, the Lasso estimator with tuning parameter*

$$\lambda := 3\sigma\sqrt{2\log p + 2L}$$

fulfills with probability at least $1-e^{-L}$ the risk bound

$$\|\boldsymbol{X}\widehat{\beta}_\lambda - \boldsymbol{X}\beta^*\|_2^2 \leq \inf_{\beta\in\mathbb{R}^p\setminus\{\boldsymbol{0}\}} \left\{ \|\boldsymbol{X}\beta - \boldsymbol{X}\beta^*\|_2^2 + \frac{18\sigma^2(L+\log p)}{\kappa^2(\beta)}|\beta|_0 \right\} \tag{51}$$

Proof. We only need to show that

$$\mathbb{P}\left(\lambda > 3|\boldsymbol{X}^T\boldsymbol{\epsilon}|_\infty\right) \geq 1-e^{-L}$$

Since $\boldsymbol{X}_j^T\boldsymbol{\epsilon} \sim \mathcal{N}(0,\sigma^2)$ for $j=1,\cdots,p$, we obtain that

$$\mathbb{P}\left(|\boldsymbol{X}^T\boldsymbol{\epsilon}|_\infty \geq \sigma\sqrt{2\log p+2L}\right) \leq p\mathbb{P}\left(|\mathcal{N}(0,1)| \geq \sqrt{2\log p+2L}\right) \leq \frac{e^{-L}}{\sqrt{\pi(\log p+L)}}$$

which completes our proof according to Theorem 4.3.1. □

4.4 Computing Algorithm

Algebraic Computation: (LARS algorithm)

Let $S_\lambda := \mathrm{supp}(\widehat{\beta}_\lambda)$ and the optimality condition is: $\exists z_\lambda$ such that

$$\begin{cases} (z_\lambda)_{S_\lambda} = \mathrm{sgn}\left((\widehat{\beta}_\lambda)_{S_\lambda}\right) \\ |z_\lambda|_\infty \leq 1 \end{cases}$$

and

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\beta}_\lambda = \boldsymbol{X}^TY - \frac{\lambda}{2}z_\lambda$$

As a result, we conclude that

$$\begin{cases} \boldsymbol{X}_{S_\lambda}^T\boldsymbol{X}\widehat{\beta}_\lambda = \boldsymbol{X}_{S_\lambda}^TY - \mathrm{sgn}\left((\widehat{\beta}_\lambda)_{S_\lambda}\right) & \text{on } S_\lambda \\ \left|\boldsymbol{X}_{S_\lambda^c}^TY - \boldsymbol{X}_{S_\lambda^c}^T\boldsymbol{X}\widehat{\beta}_\lambda\right|_\infty \leq \lambda/2 & \text{on } S_\lambda^c \end{cases}$$

Since $\lambda \to \widehat{\beta}_\lambda$ is continuous, and it concludes that $\lambda \to S_\lambda$ is piecewise constant, i.e. $\lambda \to \widehat{\beta}_\lambda$ is piecewise linear. In detail, LARS algorithm computes algebraically the breakpoints $\{\widehat{\beta}_{\lambda_1}, \widehat{\beta}_{\lambda_2}, \cdots\}$ of the path $\lambda \to \widehat{\beta}_\lambda$ as λ decreases from ∞ to 0. Then for $\lambda \in (\lambda_{k+1}, \lambda_k)$, $\widehat{\beta}$ is computed by linear interpolation. See Figure 1. However, we notice that

- All the computations are of algebraic nature, we do not know the explicit solution of $\widehat{\beta}_\lambda$.

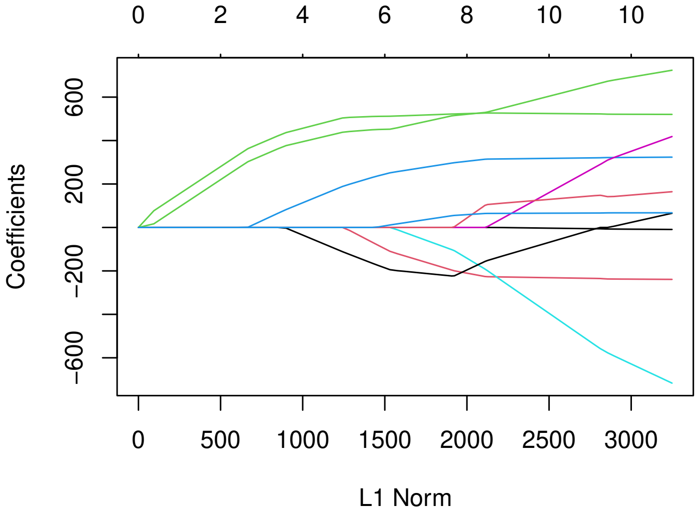

Figure 1: The line j represents the value of the j-th coordinate $\widehat{\beta}_{\lambda}^{(j)}$ of the Lasso estimator $\widehat{\beta}_{\lambda}$, when λ decreases from $+\infty$ to 0 on the diabetes data set.

- Matrix inverse are not easy to computed precisely.

Due to these weaknesses, we may consider

$$\min_{\beta\in\mathbb{R}^p}\{F(\beta)+\lambda|\beta|_1\}$$

where F is convex and smooth. Then by Taylor expansion, it concludes that

$$F(\beta)=F(\beta^t)+\langle\nabla F(\beta^t),\beta-\beta^t\rangle+\mathcal{O}\left(\|\beta-\beta^t\|_2^2\right)$$

And we try to solve

$$\beta^{t+1}\in\arg\min_{\beta\in\mathbb{R}^p}\left\{F(\beta^t)+\langle\nabla F(\beta^t),\beta-\beta^t\rangle+\frac{1}{2\eta}\|\beta-\beta^t\|_2^2+\lambda|\beta|_1\right\}$$

where η is a small positive constant. Notice that

$$\begin{aligned}\beta^{t+1}\in\arg\min_{\beta\in\mathbb{R}^p}&\left\{F(\beta^t)+\langle\nabla F(\beta^t),\beta-\beta^t\rangle+\frac{1}{2\eta}\|\beta-\beta^t\|_2^2+\lambda|\beta|_1\right\}\\=\arg\min_{\beta\in\mathbb{R}^p}&\left\{\frac{1}{2\eta}\|\beta-\beta^t+\eta\nabla F(\beta^t)\|_2^2+F(\beta^t)-\frac{\eta}{2}\|\nabla F(\beta^t)\|_2^2+\lambda|\beta|_1\right\}\\=\arg\min_{\beta\in\mathbb{R}^p}&\left\{\frac{1}{2}\|\beta-\beta^t+\eta\nabla F(\beta^t)\|_2^2+\lambda\eta|\beta|_1\right\}\end{aligned}$$

Now, we have known the explicit solution of

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|\beta-\alpha\|_2^2+\lambda|\beta|_1\right\}$$

is

$$S_\lambda(\alpha) = \left[\begin{array}{c} \alpha_1 \left(1 - \lambda/|\alpha_1|\right)_+ \\ \vdots \\ \alpha_p \left(1 - \lambda/|\alpha_p|\right)_+ \end{array} \right]$$

As a result, we conclude that $\beta^{t+1} = S_{\lambda\eta}\left(\beta^t - \eta\nabla F(\beta^t)\right)$. When $\lambda = 0$, since $S_0(\alpha) = \alpha$, the above algorithm simply amounts to a minimization of F by gradient descent with step size η. In addition, the convergence of this algorithm is ensured if η is small enough. Finally, it can be accelerated by using Nesterov's acceleration trick leading to FISTA algorithm described below.

Ridge regression:

In this part, we focus on one simple algorithm, which is called the coordinate decent method. Consider the linear model $Y = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$, where $Y, \boldsymbol{\epsilon} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, and $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$. For $\lambda > 0$, define the ridge estimator

$$\widehat{\beta}_\lambda := \arg\min_{\beta \in \mathbb{R}^p} \mathscr{L}_1(\beta)$$

where $\mathscr{L}_1(\beta) := \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$. It is easy to see that $\mathscr{L}_1(\beta)$ is strictly convex and hence only has unique minimum. Then consider the solution of $\nabla\mathscr{L}_1(\beta) = 0$, it concludes that

$$-2\boldsymbol{X}^T\left(Y - \boldsymbol{X}\beta\right) + 2\lambda\beta = \mathbf{0}$$

i.e.

$$\widehat{\beta}_\lambda = \left(\lambda \boldsymbol{I}_n + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T Y$$

Next, consider the singular value decomposition of $\boldsymbol{X} = \sum_{k=1}^r s_k u_k v_k^T$ and denote $A_\lambda = \left(\lambda \boldsymbol{I}_n + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$, then it concludes that

$$A_\lambda = \sum_{k=1}^r \frac{s_k}{s_k^2 + \lambda} v_k u_k^T$$

Proof. In fact, consider the eigenvalues of $A_\lambda^T A_\lambda = \boldsymbol{X}\left(\lambda \boldsymbol{I}_n + \boldsymbol{X}^T\boldsymbol{X}\right)^{-2} \boldsymbol{X}^T$, suppose

$$\boldsymbol{X}^T\boldsymbol{X} = Q^T \begin{pmatrix} s_1^2 & & & \\ & \ddots & & \\ & & s_r^2 & \\ & & & 0 \end{pmatrix} Q$$

Then we obtain

$$\begin{aligned} A_\lambda^T A_\lambda &= \boldsymbol{X} Q^T \begin{pmatrix} (\lambda + s_1^2)^{-2} & & & \\ & \ddots & & \\ & & (\lambda + s_r^2)^{-2} & \\ & & & 0 \end{pmatrix} Q \boldsymbol{X}^T \\ &= P \begin{pmatrix} s_1^2(\lambda + s_1^2)^{-2} & & & \\ & \ddots & & \\ & & s_r^2(\lambda + s_r^2)^{-2} & \\ & & & 0 \end{pmatrix} P^T := P\Sigma' P^T \end{aligned}$$

where P is orthogonal such that

$$P\boldsymbol{X}Q = \begin{pmatrix} s_1 & & & \\ & \ddots & & \\ & & s_r & \\ & & & 0 \end{pmatrix} := \Sigma$$

Moreover, we can show that $A_\lambda A_\lambda^T = Q^T \Sigma' Q$, which completes our proof. □

As a result, we obtain that

$$\lim_{\lambda \to 0^+} A_\lambda = \sum_{k=1}^r \frac{1}{s_k} v_k u_k^T = Q^T \Sigma^{-1} P^T := A^+$$

where A^+ is called the Moore–Penrose pseudo-inverse of A. In fact, we can conclude that

$$A^+ A = \boldsymbol{I}_p \quad AA^+ = \boldsymbol{I}_n$$

Hence,

$$\boldsymbol{X}\widehat{\beta}_\lambda = \boldsymbol{X} A_\lambda Y = \sum_{k=1}^r \frac{s_k}{s_k^2 + \lambda} \boldsymbol{X} v_k u_k^T Y = \sum_{k=1}^r \frac{s_k^2}{s_k^2 + \lambda} \langle Y, u_k \rangle u_k$$

and

$$\mathbb{E}\left[\widehat{\beta}_\lambda\right] = \sum_{k=1}^r \frac{s_k}{s_k^2 + \lambda} v_k u_k^T \boldsymbol{X} \beta = \sum_{k=1}^r \frac{s_k^2}{s_k^2 + \lambda} \langle \beta, v_k \rangle v_k$$

Consequently, it deduces that

$$\left\|\beta - \mathbb{E}\left[\widehat{\beta}_\lambda\right]\right\|_2^2 = \left\|\beta - \sum_{k=1}^r \langle \beta, v_k \rangle v_k\right\|_2^2 + \sum_{k=1}^r \left(\frac{\lambda}{s_k^2 + \lambda}\right)^2 \langle \beta, v_k \rangle^2$$

and the covariance of ridge estimator is

$$\mathrm{Cov}\left(\widehat{\beta}_\lambda\right) = \mathbb{E}\left[\left(\widehat{\beta}_\lambda - \mathbb{E}\left[\widehat{\beta}_\lambda\right]\right)^T \left(\widehat{\beta}_\lambda - \mathbb{E}\left[\widehat{\beta}_\lambda\right]\right)\right]$$

In addition, since

$$\begin{aligned}\widehat{\beta}_\lambda - \mathbb{E}\left[\widehat{\beta}_\lambda\right] &= \sum_{k=1}^r \frac{s_k}{s_k^2+\lambda}\left(\langle Y, u_k\rangle - s_k\langle \beta, v_k\rangle\right) v_k \\ &= \sum_{k=1}^r \frac{s_k}{s_k^2+\lambda}\langle \boldsymbol{\epsilon}, u_k\rangle v_k\end{aligned}$$

and $\langle \boldsymbol{\epsilon}, u_k\rangle \sim \mathcal{N}(0,\sigma^2)$. Hence, we conclude that

$$\mathrm{Cov}\left(\widehat{\beta}_\lambda\right) = \sigma^2 \sum_{k=1}^r \left(\frac{s_k}{s_k^2+\lambda}\right)^2 = \sigma^2 \mathrm{Tr}\left(A_\lambda^T A_\lambda\right)$$

Elastic-Net:

In ridge regression model, we use L^2-norm as a replacement of L^1-norm, but we notice that it does not select variables. Therefore, in order to make balance, we will use both these two norms, that is,

$$\widetilde{\beta}_{\lambda,\mu} := \min_{\beta\in\mathbb{R}^p} \mathscr{L}_2(\beta)$$

where $\mathscr{L}_2(\beta) := \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 + \mu|\beta|_1$, as before, we still assume that the columns of $\boldsymbol{X}$ have norm of 1, then its sub-differential can be computed as

$$-2\boldsymbol{X}^T\left(Y - \boldsymbol{X}\widetilde{\beta}_{\lambda,\mu}\right) + 2\lambda\widetilde{\beta}_{\lambda,\mu} + \mu z_{\lambda,\mu} = \mathbf{0}$$

where $z_{\lambda,\mu} \in \partial\mathscr{L}_2(\widetilde{\beta}_{\lambda,\mu})$, i.e. for $\widetilde{\beta}_{\lambda,\mu}^{(j)} \neq 0$, it has

$$-2\boldsymbol{X}_j^T\left(Y - \boldsymbol{X}\widetilde{\beta}_{\lambda,\mu}\right) + 2\lambda\widetilde{\beta}_{\lambda,\mu}^{(j)} + \mu\mathrm{sgn}\left(\widetilde{\beta}_{\lambda,\mu}^{(j)}\right) = 0$$

In addition, since

$$\boldsymbol{X}\widetilde{\beta}_{\lambda,\mu} = \sum_{k=1}^p \widetilde{\beta}_{\lambda,\mu}^{(k)}\boldsymbol{X}_k$$

it concludes that

$$(2+2\lambda)\widetilde{\beta}_{\lambda,\mu}^{(j)} - 2\boldsymbol{X}_j^T R_j + \mu\mathrm{sgn}\left(\widetilde{\beta}_{\lambda,\mu}^{(j)}\right) = 0$$

where $R_j := \boldsymbol{X}_j^T\left(Y - \sum_{k\neq j}\widetilde{\beta}_{\lambda,\mu}^{(k)}\boldsymbol{X}_k\right)$, and hence

$$\widetilde{\beta}_{\lambda,\mu}^{(j)} = \frac{R_j}{1+\lambda}\left(1 - \frac{\mu}{2|R_j|_1}\right)_+$$

4.5 Bias of Lasso Estimator

Remember the explicit solution of Lasso estimation when $\boldsymbol{X}$ is orthogonal, i.e.

$$\widehat{\beta}_\lambda^{(j)} = \boldsymbol{X}_j^T Y\left(1 - \frac{\lambda}{2|\boldsymbol{X}_j^T Y|}\right)_+$$

Hence the L^1-norm actually select the j such that $|\boldsymbol{X}_j^T Y| \geq \lambda/2$, but it does also shrink the coordinates $\boldsymbol{X}_j^T Y$ by a factor of $(1 - \lambda/(2|\boldsymbol{X}_j^T|))_+$, which will cause that the estimation results are shrunk toward zero. A common trick to remove this shrinkage is to use as final estimator the so-called Gauss-Lasso estimator.

$$\widehat{f}_\lambda^{\text{Gauss}} = \text{Proj}_{\widehat{S}_\lambda} Y \quad \text{where } \widehat{S}_\lambda := \text{span}\left\{\boldsymbol{X}_j : j \in \widehat{m}_\lambda\right\}$$

In other words, $\widehat{f}_\lambda^{\text{Gauss}} = \widehat{f}_{\widehat{m}_\lambda}$ where $\widehat{m}_\lambda := \text{supp}\left(\widehat{\beta}_\lambda\right)$. Another trick for reducing the shrinkage is to compute first the Gauss-Lasso estimator $\widehat{f}_\lambda^{\text{Gauss}} = \boldsymbol{X}\widehat{\beta}_\lambda^{\text{Gauss}}$

5 Iterative Algorithms

So far, we have discussed the model selection and convex relaxation, they both have some weaknesses. As we can see, the model selection provides statistically optimal estimators, but with a prohibitive computational cost. While the convex relaxation can be computationally efficient, it suffers from shrinkage bias. In this section, we will focus on iterative algorithms, which largely avoids these weaknesses.

5.1 Iterative Hard Thresholding

First, we focus on coordinate sparsity linear regression model, i.e.

$$Y = \boldsymbol{X}\beta^* + \boldsymbol{\epsilon}$$

where $|\beta^*|_0$ is small and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Reminder on Lasso estimation, we relax the following optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda |\beta|_0 \right\}$$

by

$$\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - \boldsymbol{X}\beta\|_2^2 + \lambda |\beta|_1 \right\}$$

which is a convex optimization problem. And the proximal recipe is to consider the following iterative procedure.

$$\begin{aligned} \beta^{t+1} &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|_2^2 + \lambda |\beta|_1 \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|_2^2 + \eta\lambda |\beta|_1 \right\} \\ &= S_{\eta\lambda}\left(\beta^t - \eta \nabla F(\beta^t)\right) \end{aligned}$$

where $S_\lambda(\alpha) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|\beta - \alpha\|_2^2 + \lambda |\beta|_1 \right\}$ and $F(\beta) = \|Y - \boldsymbol{X}\beta\|_2^2$, which has been introduced in last section. Now, we will replace L^1-norm by L^0-norm, i.e.

$$\begin{aligned} \beta^{t+1} &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|_2^2 + \lambda^2 |\beta|_0 \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|_2^2 + 2\eta\lambda^2 |\beta|_0 \right\} \\ &= H_{\lambda\sqrt{2\eta}}\left(\beta^t - \eta \nabla F(\beta^t)\right) \end{aligned}$$

where $H_\lambda(\alpha) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|\beta - \alpha\|_2^2 + \lambda |\beta|_0 \right\}$, and we can show that

$$H_\lambda(\alpha) = \begin{pmatrix} \alpha_1 1_{|\alpha_1| > \lambda} \\ \vdots \\ \alpha_p 1_{|\alpha_p| > \lambda} \end{pmatrix}$$

In fact, we have

$$\begin{aligned}\min_{\beta\in\mathbb{R}^p}\left\{\|\beta-\alpha\|_2^2+\lambda|\beta|_0\right\} &= \min_{m\in\mathscr{M}}\left\{\|\alpha-\mathrm{Proj}_{S_m}\alpha\|_2^2+\lambda|m|\right\}\\ &= \min_{m\in\mathscr{M}}\left\{\lambda|m|-\|\mathrm{Proj}_{S_m}\alpha\|_2^2\right\}\\ &= \min_{m\in\mathscr{M}}\left\{\sum_{j\in m}\left(\lambda-\alpha_j^2\right)\right\}\end{aligned}$$

which concludes this claim. And we call this method is hard thresholding, it is clear to see that there is no shrinkage, but since $|\beta|_0$ is non-convex, so the convergence can not be guaranteed.

How to do next? In fact, we can give a stronger conclusion as a variant of above optimization problem. Let $\eta = 0.5$ and $\beta^0 = \mathbf{0}$, λ_t is no longer a constant, but decreasing from a high-value to the target value $c\sigma\sqrt{\log p}$, which has been discussed in section of model selection. This method is called the Iterative Hard Thresholding (IHT) algorithm. In detail, for the choice of λ_t, let $A, B > 0$ and $a > 1$, denote

$$\lambda_t = Aa^{-t} + B \tag{52}$$

In fact, the sequence λ_t satisfies the equation of $\lambda_{t+1} = a^{-1}\lambda_t + (a-1)B/a$ with initial $\lambda_0 = A + B$. Moreover,

$$\widehat{\beta}^{t+1} = H_{\lambda_{t+1}}\left(\widehat{\beta}^t - \frac{1}{2}\nabla F(\widehat{\beta}^t)\right) = H_{\lambda_{t+1}}\left(\left(\boldsymbol{I}_p - \boldsymbol{X}^T\boldsymbol{X}\right)\widehat{\beta}^t + \boldsymbol{X}^TY\right) \tag{53}$$

Next, denote $\Lambda := \boldsymbol{I}_p - \boldsymbol{X}^T\boldsymbol{X}$ and we will analyses the risk bound of IHT.

Reconstruction error:

Here, let β^t_{Lasso} be the approximation of $\widehat{\beta}^{\text{Lasso}}$ after t steps of soft-thresholding iterations. The classic error decomposition is

$$\left\|\beta^t_{\text{Lasso}} - \beta^*\right\|_2 \leq \underbrace{\left\|\beta^t_{\text{Lasso}} - \widehat{\beta}^{\text{Lasso}}\right\|_2}_{\text{optim.error}} + \underbrace{\left\|\widehat{\beta}^{\text{Lasso}} - \beta^*\right\|_2}_{\text{stat.error}}$$

The two errors can be analyzed apart, using tools from convex optimization for the optimization error and tools from statistics for the second term. However, the classical approach does not make sense for the analysis of IHT estimator, since we do not guarantee the convergence of $\widehat{\beta}^t$. Our recipe is to decompose

$$\widehat{\beta}^{t+1} := H_{\lambda_{t+1}}\left(\Lambda\widehat{\beta}^t + \boldsymbol{X}^TY\right) = H_{\lambda_{t+1}}\left(\beta^* + \Lambda\left(\widehat{\beta}^t - \beta^*\right) + \boldsymbol{X}^T\boldsymbol{\epsilon}\right)$$

where β^* is our target and $\boldsymbol{X}^T\boldsymbol{\epsilon}$ is the statistical noise. And we hope that $\Lambda\left(\widehat{\beta}^t - \beta^*\right)$ will be a contraction term, in fact, with suitable λ_t, $\widehat{\beta}^t - \beta^*$ will be sparse and then Λ will act as a contraction, hence for $t \to \infty$,

$$\widehat{\beta}^t \asymp H_{\lambda_\infty}\left(\beta^* + \boldsymbol{X}^T\boldsymbol{\epsilon}\right)$$

Let's see the following theorem.

Theorem 5.1.1 (Deterministic error bound for IHT)**.** *Set* $\Lambda = \boldsymbol{I}_p - \boldsymbol{X}^T\boldsymbol{X}$ *and* $m^* = \operatorname{span}(\beta^*)$. *Assume that for* $0 < \delta < 1$ *and* $c \geq 1$ *such that* $c^2|\beta^*|_0 \in \mathbb{N}$,

$$\max_{\substack{S \subset \{1,\cdots,p\} \\ |S| \leq \bar{k}}} \sup_{\|x\|_2 \leq 1} |\Lambda_{SS}x| \leq \delta \quad \text{with} \quad \bar{k} = (1+2c^2)|\beta^*|_0 \tag{54}$$

Assume also that,

$$1 < a \leq \frac{c}{\delta(1+2c)}, \qquad A \geq \frac{\|\beta^*\|_2}{(1+2c)\sqrt{|\beta^*|_0}}, \qquad B > \frac{a}{a-1}|\boldsymbol{X}^T\boldsymbol{\epsilon}|_\infty \tag{55}$$

Then, for all $t \geq 0$, *the estimator* $\widehat{\beta}^t$ *defined by (53) with threshold levels given fulfills (52) and*

$$\left|\widehat{\beta}^t_{\bar{m}^*}\right|_0 \leq c^2\,|\beta^*|_0 \tag{56}$$

where $\bar{m}^* := \{1,\cdots,p\}\backslash m^*$ *and*

$$\|\widehat{\beta}^t - \beta^*\|_2 \leq (1+2c)\lambda_t\sqrt{|\beta^*|_0} \tag{57}$$

Before giving the proof of Theorem 5.1.1, let's see how IHT algorithm works, suppose

$$\left|\widehat{\beta}^t_{\bar{m}^*}\right|_0 \leq c^2\,|\beta^*|_0$$

we conclude that

$$|\widehat{\beta}^t - \beta^*|_0 \leq |\beta^*|_0 + |\widehat{\beta}^t|_0 \leq \left|\widehat{\beta}^t_{\bar{m}^*}\right|_0 + \left|\widehat{\beta}^t_{m^*}\right|_0 + |\beta^*|_0 \leq (2+c^2)|\beta^*|_0 \leq \bar{k}$$

As a result, we conclude that

$$\left\|\Lambda\left(\widehat{\beta}^t - \beta^*\right)\right\|_2 \leq \delta\|\widehat{\beta}^t - \beta^*\|_2$$

i.e. Λ acts as a contraction.

Proof of Theorem 5.1.1. Let's make some notations first, denote $Z := \boldsymbol{X}^T\boldsymbol{\epsilon}$, $b^{t+1} := \beta^* + \Lambda(\widehat{\beta}^t - \beta^*) + Z$ and $k^* := |\beta^*|_0$, then it concludes that $\widehat{\beta}^{t+1} = H_{\lambda_{t+1}}(b^{t+1})$. Then for $t = 0$, it is trivial to obtain our conclusions. Next, by induction, suppose they have been hold at step t, for step $t+1$, we first show that

$$\max_{\substack{S \subset \{1,\cdots,p\} \\ |S| \leq c^2k^*}} \left(\left\|\left(b^{t+1} - \beta^*\right)_S\right\|_2 - \sqrt{|S|}|Z|_\infty\right) \leq \delta\|\beta^* - \widehat{\beta}^t\|_2$$

Denote $\widetilde{S}^t := m^* \cup \operatorname{supp}(\widehat{\beta}^t_{\bar{m}^*})$ and $S' := \widetilde{S}^t \cup S$, then by step t, it has that $|\widetilde{S}^t| \le (1+c^2)k^*$ and hence $|S'| \le (1+2c^2)k^* = \bar{k}$. As a result, it concludes that

$$\begin{aligned}
\left\|\left(b^{t+1}-\beta^*\right)_S\right\|_2 &\le \left\|\left(\Lambda\left(\widehat{\beta}^t-\beta^*\right)\right)_S\right\|_2 + \|Z_S\|_2 \\
&\le \left\|\left(\Lambda\left(\widehat{\beta}^t-\beta^*\right)\right)_{S'}\right\|_2 + \sqrt{|S|}|Z|_\infty \\
&\le \left\|\Lambda_{S'S'}\left(\widehat{\beta}^t-\beta^*\right)_{S'}\right\|_2 + \sqrt{|S|}|Z|_\infty \\
&\le \delta\left\|\left(\widehat{\beta}^t-\beta^*\right)_{S'}\right\|_2 + \sqrt{|S|}|Z|_\infty \\
&\le \delta\left\|\widehat{\beta}^t-\beta^*\right\|_2 + \sqrt{|S|}|Z|_\infty
\end{aligned}$$

which concludes this claim. In addition, we have

$$\|\widehat{\beta}^{t+1}-\beta^*\|_2 \le \|\widehat{\beta}^{t+1}_{m^*}-\beta^*_{m^*}\|_2 + \|\widehat{\beta}^{t+1}_{\bar{m}^*}\|_2$$

And the first term of the right side is bounded by

$$\begin{aligned}
\|\widehat{\beta}^{t+1}_{m^*}-\beta^*_{m^*}\|_2 &\le \|b^{t+1}_{m^*}-H_{\lambda_{t+1}}(b^{t+1}_{m^*})\|_2 + \|b^{t+1}_{m^*}-\beta^*_{m^*}\|_2 \\
&\le \lambda_{t+1}\sqrt{|m^*|} + \delta\left\|\widehat{\beta}^t-\beta^*\right\|_2 + \sqrt{|S|}|Z|_\infty \\
&\le \lambda_{t+1}\sqrt{|m^*|} + (1+2c)\delta\lambda_t\sqrt{|m^*|} + c|Z|_\infty\sqrt{|m^*|} \\
&\le \sqrt{|m^*|}\left(\lambda_{t+1} + ca^{-1}\lambda_t + c|Z|_\infty\right) \\
&\le (1+c)\,\lambda_{t+1}\sqrt{|m^*|}
\end{aligned}$$

The last inequality is valid due to $\lambda_{t+1} = a^{-1}\lambda_t + (a-1)B/a$ and $|Z|_\infty < (a-1)B/a$. For the term of $\|\widehat{\beta}^{t+1}_{\bar{m}^*}\|_2$, for any $S \subset \operatorname{supp}(\widehat{\beta}^{t+1}_{\bar{m}^*})$ with $|S| \le c^2k^*$, it has that

$$\lambda_{k+1}\sqrt{|S|} \le \|\widehat{\beta}^{t+1}_S\|_2 = \left\|\left(H_{\lambda_{t+1}}\left(b^{t+1}\right)\right)_S\right\|_2 = \left\|b^{t+1}_S\right\|_2 = \left\|b^{t+1}_S - \beta^*_S\right\|_2$$

Moreover, we have $\left\|b^{t+1}_S - \beta^*_S\right\|_2 \le \delta\|\widehat{\beta}^t-\beta^*\|_2 + \sqrt{|S|}|Z|_\infty$, since $\lambda_{k+1} > a^{-1}\lambda_t + |Z|_\infty$, it concludes that

$$\sqrt{|S|} \le \frac{\delta\|\widehat{\beta}^t-\beta^*\|_2}{\lambda_{t+1}-|Z|_\infty} \le \frac{a\delta\|\widehat{\beta}^t-\beta^*\|_2}{\lambda_t} \le c\sqrt{|m^*|}$$

Since S is any subset of $\operatorname{supp}(\widehat{\beta}^{t+1}_{\bar{m}^*})$, it concludes that $|\widehat{\beta}^{t+1}_{\bar{m}^*}|_0 \le c^2|\beta^*|_0$, which is the first conclusion of step $t+1$. Finally, since

$$\|\widehat{\beta}^{t+1}_S\|_2 \le \delta\|\widehat{\beta}^t-\beta^*\|_2 + \sqrt{|S|}|Z|_\infty \le c\sqrt{|\beta^*|_0}\left(a^{-1}\lambda_t + |Z|_\infty\right) \le \lambda_{t+1}c\sqrt{|\beta^*|_0}$$

which concludes the second conclusion of step $t+1$. □

For practical usage, let's choose special values of parameters A and B, e.g.

$$A := \frac{\|\boldsymbol{X}^TY\|_2 + \sigma|\boldsymbol{X}|_2\sqrt{2}(1+\sqrt{L})}{3(1-\delta)} \quad \text{and} \quad B := \frac{a\sigma}{a-1}\sqrt{2\log p + 2L} \tag{58}$$

where $L > 0$ is a constant and the noise $\boldsymbol{\epsilon}$ is chosen as $\mathcal{N}(\mathbf{0}, \sigma \boldsymbol{I}_n)$. Since we need to stop our iterative process at some fixed step, let's choose

$$\widehat{t} := \min\{k \in \mathbb{N} : k \geq \log_a(A/B)\} \tag{59}$$

and we will obtain the following result.

Corollary 5.1.1 (Error bound in the Gaussian setting for IHT)**.** *Assume that the columns of $\boldsymbol{X}$ have unit L^2-norm, i.e. $\|\boldsymbol{X}_j\|_2 = 1$ for $j = 1, \cdots, p$, and that the noise $\boldsymbol{\epsilon}$ follows a $\mathcal{N}(\mathbf{0}, \sigma \boldsymbol{I}_n)$ Gaussian distribution. Assume also that Assumption (54) holds and that*

$$1 < a \leq \frac{c}{\delta(1+2c)}$$

Then, for any iteration t larger than $\widehat{t}$ defined by (59), with probability larger than $1 - 2e^{-L}$, the estimator $\widehat{\beta}^t$, with A, B given by (58), fulfills $|\widehat{\beta}^t_{\bar{m}^*}|_0 \leq c^2 |\beta^*|_0$ and

$$\|\widehat{\beta}^t - \beta^*\|_2^2 \leq C_{a,c} |\beta^*|_0 \sigma^2 (\log p + L) \tag{60}$$

with

$$C_{a,c} = 2\left(\frac{2a(1+2c)}{a-1}\right)^2$$

Proof. Only we need to do is to show that (55) holds with probability $1 - 2e^{-L}$, in fact, $\boldsymbol{X}^T \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$ and it concludes that

$$\mathbb{P}\left(B < \frac{a}{a-1} |\boldsymbol{X}^T \boldsymbol{\epsilon}|\right) \leq e^{-L}$$

Moreover, we also need to show that

$$\mathbb{P}\left(A \geq \frac{\|\beta^*\|_2}{(1+2c)\sqrt{|\beta^*|_0}}\right) \geq 1 - e^{-L}$$

First, notice that $\boldsymbol{\epsilon} \to \|\boldsymbol{X}^T(\boldsymbol{X}\beta^* + \boldsymbol{\epsilon})\|_2$ is $|\boldsymbol{X}|_2$-Lipschitz. By Theorem 1.2.2, there exists $\xi \sim \mathrm{Exp}(1)$ such that

$$\left| \|\boldsymbol{X}^T Y\|_2 - \mathbb{E}\left[\|\boldsymbol{X}^T Y\|_2\right] \right| \leq \sigma |\boldsymbol{X}|_2 \sqrt{2\xi}$$

Hence,

$$\mathbb{E}\left[\|\boldsymbol{X}^T Y\|_2^2\right] \leq \mathbb{E}\left[\left(\mathbb{E}\left[\|\boldsymbol{X}^T Y\|_2\right] + \sigma |\boldsymbol{X}|_2 \sqrt{2\xi}\right)^2\right] \leq \left(\mathbb{E}\left[\|\boldsymbol{X}^T Y\|_2\right] + \sqrt{2}\sigma |\boldsymbol{X}|_2 \sqrt{\mathbb{E}[\xi]}\right)^2$$

The last inequality is valid due to Jensen's inequality and $\mathbb{E}[\xi] = 1$, hence

$$\begin{aligned}
\|\boldsymbol{X}^T Y\|_2 + \sigma |\boldsymbol{X}|_2 (\sqrt{2} + \sqrt{2\xi}) &\geq \mathbb{E}\left[\|\boldsymbol{X}^T Y\|_2\right] + \sqrt{2}\sigma |\boldsymbol{X}|_2 \\
&\geq \mathbb{E}\left[\|\boldsymbol{X}^T Y\|_2^2\right]^{1/2} \\
&= \sqrt{\|\boldsymbol{X}^T \boldsymbol{X} \beta^*\|_2^2 + \mathbb{E}\left[\|\boldsymbol{X}\boldsymbol{\epsilon}\|_2^2\right]} \\
&\geq \|(\boldsymbol{X}^T \boldsymbol{X} \beta^*)_{m^*}\|_2 \\
&\geq \|\beta^*_{m^*}\|_2 - \|\Lambda_{m^* m^*} \beta^*_{m^*}\|_2 \\
&\geq (1-\delta)\|\beta^*_{m^*}\|_2 = (1-\delta)\|\beta^*\|_2
\end{aligned}$$

Now, replace ξ by L and the probability of $\mathbb{P}(\xi \leq L) = 1 - e^{-L}$, hence we conclude that

$$A := \frac{\|\boldsymbol{X}^T Y\|_2 + \sigma|\boldsymbol{X}|_2\sqrt{2}(1+\sqrt{L})}{3(1-\delta)} \geq \frac{\|\beta^*\|_2}{3} \geq \frac{\|\beta^*\|_2}{(1+2c)\sqrt{|\beta^*|_0}}$$

with probability $1 - e^{-L}$ since $c \geq 1$. As a result, the probability of A, B which satisfies Assumption (55) with probability larger than $1 - 2e^{-L}$, and so Theorem 5.1.1 is valid. In addition, for any $t \geq \widehat{t}$, it concludes that

$$\lambda_t = a^{-t}A + B \leq 2B$$

Hence, by Theorem 5.1.1, it concludes that

$$\|\widehat{\beta}^t - \beta^*\|_2^2 \leq (1+2c)^2\lambda_t^2|\beta^*|_0 \leq 4(1+2c)^2B^2|\beta^*|_0 = C_{a,c}|\beta^*|_0\sigma^2\left(\log p + L\right)$$

which completes our proof. □

5.2 Iterative Group Thresholding

In this subsection, we will extend the methodology developed above to the group-sparse setting, i.e. $\{1, \cdots, p\} = \cup_{j=1}^M G_j$, and the minimization problem is

$$\min_{\beta\in\mathbb{R}^p}\left\{\|Y - \boldsymbol{X}\beta\|_2^2 + \sum_{j=1}^M \left(\lambda^{(j)}\right)^2 1_{\beta_{G_j}\neq 0}\right\} \tag{61}$$

For the iterative algorithm, consider the following process, denote $\widehat{\beta}^t$ as the current estimation and $F(\beta) = \|Y - \boldsymbol{X}\beta\|_2^2$, the updating $\widehat{\beta}^{t+1}$ should be

$$\begin{aligned}\widehat{\beta}^{t+1} &\in \arg\min_{\beta\in\mathbb{R}^p}\left\{F(\widehat{\beta}^t) + \langle\beta - \widehat{\beta}^t, \nabla F(\widehat{\beta}^t)\rangle + \frac{1}{2\eta}\|\beta - \widehat{\beta}^t\|_2^2 + \sum_{j=1}^M \left(\lambda^{(j)}\right)^2 1_{\beta_{G_j}\neq 0}\right\} \\ &= \arg\min_{\beta\in\mathbb{R}^p}\left\{\|\beta - \beta^t + \eta\nabla F(\beta^t)\|_2^2 + 2\eta\sum_{j=1}^M \left(\lambda^{(j)}\right)^2 1_{\beta_{G_j}\neq 0}\right\}\end{aligned}$$

For the solution of this minimization problem, we have

$$\begin{aligned}\min_{\beta\in\mathbb{R}^p}\left\{\|\beta - \alpha\|_2^2 + \sum_{j=1}^M \left(\lambda^{(j)}\right)^2 1_{\beta_{G_j}\neq 0}\right\} &= \min_{m\in\mathscr{M}}\left\{\|\alpha - \mathrm{Proj}_{S_m}(\alpha)\|_2^2 + \sum_{j\in m}\left(\lambda^{(j)}\right)^2\right\} \\ &= \min_{m\in\mathscr{M}}\left\{\sum_{j\in m}\left(\lambda^{(j)}\right)^2 - \|\mathrm{Proj}_{S_m}(\alpha)\|_2^2\right\} \\ &= \min_{m\in\mathscr{M}}\left\{\sum_{j\in m}\left(\left(\lambda^{(j)}\right)^2 - \|\alpha_{G_j}\|_2^2\right)\right\}\end{aligned}$$

where $\mathscr{M} := \mathscr{P}(1, \cdots, M)$ and hence the solution of above minimization problem is

$$\widehat{m} := \big\{ j : \|\alpha_{G_j}\|_2 \geq \lambda^{(j)}, j = 1, \cdots, M \big\}$$

And define the function H^G_λ such that

$$\big[H^G_\lambda(\alpha)\big]_{G_j} = \alpha_{G_j} 1_{\|\alpha_{G_j}\|_2 \geq \lambda^{(j)}}$$

As a result, we conclude that $\beta^{t+1} := H^G_{\lambda\sqrt{2\eta}}(\beta^t - \eta \nabla F(\beta^t))$. Similar as before, choose $\eta = 1/2$ and λ_t be a function of t, i.e.

$$\widehat{\beta}^{t+1} = H^G_{\lambda_{t+1}}\left(\widehat{\beta}^t - \frac{1}{2}\nabla F(\widehat{\beta}^t)\right) = H^G_{\lambda_{t+1}}\left((\boldsymbol{I}_n - \boldsymbol{X}^T\boldsymbol{X})\widehat{\beta}^t + \boldsymbol{X}^T Y\right) \tag{62}$$

For the choice of λ_t, consider the threshold levels of the form

$$\lambda_t^{(j)} := \sqrt{|G_j|}\gamma_t \quad \text{where} \quad \gamma_t = Aa^{-t} + B \tag{63}$$

with $a > 1$ and $A, B > 0$. Before talking about the Iterative Group Thresholding (IGT) algorithm, let's introduce some notations. First, define the group support

$$\operatorname{supp}_G(\beta) = \bigcup_{j : \beta_{G_j} \neq \mathbf{0}} G_j$$

and denote $|\beta|_0^G := \operatorname{card}(\operatorname{supp}_G(\beta))$, then set the norm

$$|Z|^G_\infty := \max_{j=1,\cdots,M} \frac{\|Z_{G_j}\|_2}{\sqrt{|G_j|}}$$

Theorem 5.2.1 (Deterministic error bound for IGT)**.** *Let's set $\Lambda = \boldsymbol{I}_n - \boldsymbol{X}^T\boldsymbol{X}$ and $m^*_G = \operatorname{supp}_G(\beta^*)$. Let's assume that all the groups have the same cardinality q. Let us also assume that for some $0 < \delta < 1$ and some $c \geq 1$ such that $c^2|\beta^*_G|_0 \in q\mathbb{N}$,*

$$\max_{\substack{S = \cup_{j \in J} G_j \\ |J| \leq J_G}} |\Lambda_{SS}|_2 \leq \delta \quad \text{with} \quad \bar{J}_G = (1 + 2c^2)|\beta^*|^G_0 / q \tag{64}$$

Assume also that,

$$1 < a \leq \frac{c}{\delta(1+2c)}, \quad A \geq \frac{\|\beta^*\|_2}{(1+2c)\sqrt{|\beta^*|^G_0}}, \quad B > \frac{a}{a-1}|\boldsymbol{X}^T\boldsymbol{\epsilon}|^G_\infty \tag{65}$$

Then, for all $t \geq 0$, the estimator β_t defined by (62) with threshold levels given by (63) fulfills

$$|\widehat{\beta}_{\bar{m}^*_G}|^G_0 \leq c^2|\beta^*|^G_0 \quad \text{where } \bar{m}^*_G = \{1, \cdots, p\} \backslash m^*_G \tag{66}$$

and

$$\|\widehat{\beta}^t - \beta^*\|_2 \leq (1+2c)\gamma_t\sqrt{|\beta^*|^G_0} \tag{67}$$

Proof. As the proof of Theorem 5.1.1, we still use induction to conclude these results, since it is trivial at step 0 for $\beta^0 = \mathbf{0}$, suppose they have been hold at step t, denote that $k_G^* := |\beta^*|_0^G$, $Z := \boldsymbol{X}^T \boldsymbol{\epsilon}$ and $b^{t+1} := \beta^* + \Lambda(\widehat{\beta}^t - \beta^*) + Z$, hence $\beta^{t+1} = H_{\lambda_{t+1}}(b^{t+1})$ and we claim that

$$\max_{\substack{S=\cup_{j\in J} G_j \\ |J|\le c^2 k_G^*/q}} \left(\left\| \left(b^{t+1} - \beta^*\right)_S \right\|_2 - \sqrt{|S|} |Z|_\infty^G \right) \le \delta \|\beta^* - \beta^t\|_2$$

Denote $\widetilde{S}_G^t := m_G^* \cup \mathrm{supp}_G(\beta_{\bar{m}_G^*}^t)$ and $S' := \widetilde{S}_G^t \cup S$, then by step t, it has that $|\widetilde{S}_G^t| \le (1+c^2)k_G^*$ and hence $|S'| \le (1+2c^2)k_G^*$. As a result, it concludes that

$$\begin{aligned}
\left\| \left(b^{t+1} - \beta^*\right)_S \right\|_2 &\le \left\| \left(\Lambda\left(\beta^t - \beta^*\right)\right)_S \right\|_2 + \|Z_S\|_2 \\
&\le \left\| \left(\Lambda\left(\beta^t - \beta^*\right)\right)_{S'} \right\|_2 + \sqrt{|S|} |Z|_\infty^G \\
&\le \left\| \Lambda_{S'S'} \left(\beta^t - \beta^*\right)_{S'} \right\|_2 + \sqrt{|S|} |Z|_\infty^G \\
&\le \delta \left\| \left(\beta^t - \beta^*\right)_{S'} \right\|_2 + \sqrt{|S|} |Z|_\infty^G \\
&\le \delta \left\| \beta^t - \beta^* \right\|_2 + \sqrt{|S|} |Z|_\infty^G
\end{aligned}$$

which concludes this claim. In addition, we have

$$\|\beta^{t+1} - \beta^*\|_2 \le \|\beta_{m_G^*}^{t+1} - \beta_{m_G^*}^*\|_2 + \|\beta_{\bar{m}^*}^{t+1}\|_2$$

And the first term of the right side is bounded by

$$\begin{aligned}
\|\beta_{m_G^*}^{t+1} - \beta_{m_G^*}^*\|_2 &\le \|b_{m_G^*}^{t+1} - [H_{\lambda_{t+1}}^G(b^{t+1})]_{m_G^*}\|_2 + \|b_{m_G^*}^{t+1} - \beta_{m_G^*}^*\|_2 \\
&\le \gamma_{t+1}\sqrt{|\beta^*|_0^G} + \delta\|\widehat{\beta}^t - \beta^*\|_2 + \sqrt{|\beta^*|_0^G}|Z|_\infty^G \\
&\le \gamma_{t+1}\sqrt{|\beta^*|_0^G} + (1+2c)\delta\gamma_t\sqrt{|\beta^*|_0^G} + \sqrt{|\beta^*|_0^G}|Z|_\infty^G \\
&\le \sqrt{|\beta^*|_0^G}\left(\gamma_{t+1} + ca^{-1}\gamma_t + |Z|_\infty^G\right) \\
&\le \sqrt{|\beta^*|_0^G}\left(\gamma_{t+1} + ca^{-1}\gamma_t + c|Z|_\infty^G\right)
\end{aligned}$$

Next, for any $S = \cup_{j\in J} G_j \subset \mathrm{supp}_G(\beta_{\bar{m}_G^*}^{t+1})$ with $|S| \le c^2 k_G^*$, it has that

$$\gamma_{t+1}\sqrt{|S|} \le \|\beta_S^{t+1}\|_2 = \left\| \left[H_{\lambda_{t+1}}^G\left(b^{t+1}\right)\right]_S \right\|_2 = \left\| b_S^{t+1} \right\|_2 = \left\| b_S^{t+1} - \beta_S^* \right\|_2$$

Moreover, we have $\left\| b_S^{t+1} - \beta_S^* \right\|_2 \le \delta\|\widehat{\beta}^t - \beta^*\|_2 + \sqrt{|S|}|Z|_\infty^G$. Besides, we can not obtain $S = c^2 k_G^*$ since $\gamma_{t+1} = a^{-1}\gamma_t + B > a^{-1}\gamma_t + (a-1)B/a > a^{-1}\gamma_t + |Z|_\infty^G$ and

$$\sqrt{|S|} \le \frac{\delta\|\beta^t - \beta^*\|_2}{\gamma_{t+1} - |Z|_\infty^G} < \frac{a\delta\|\widehat{\beta}^t - \beta^*\|_2}{\gamma_t} \le c\sqrt{|\beta^*|_0^G}$$

Since S is any subset of $\mathrm{supp}_G(\beta_{\bar{m}_G^*}^{t+1})$, it concludes that $|\beta_{\bar{m}_G^*}^{t+1}|_0 \le c^2|\beta^*|_0^G$, which is the first conclusion of step $t+1$. Finally, since

$$\|\beta_{\bar{m}_G^*}^{t+1}\|_2 \le \delta\|\widehat{\beta}^t - \beta^*\|_2 + \sqrt{|\bar{m}_G^*|}|Z|_\infty^G \le c\sqrt{|\beta^*|_0^G}\left(a^{-1}\gamma_t + |Z|_\infty^G\right) \le \gamma_{t+1}c\sqrt{|\beta^*|_0^G}$$

which concludes the second conclusion of step $t+1$. □

As the end of this section, we will finally give the following result. First, let's make some notations.

$$A := \frac{\|\boldsymbol{X}^T Y\|_2 + \sigma|\boldsymbol{X}|_2\sqrt{2}(1+\sqrt{L})}{3(1-\delta)} \quad \text{and} \quad B := \frac{a\sigma}{a-1}\left(1+\phi_G\sqrt{2\log M + 2L}\right) \tag{68}$$

with

$$\phi_G := \max_{j=1,\cdots,M} \frac{|\boldsymbol{X}_{G_j}|_2}{\sqrt{|G_j|}}$$

where $L > 0$ is a constant and the noise $\boldsymbol{\epsilon}$ is chosen as $\mathcal{N}(\mathbf{0}, \sigma\boldsymbol{I}_n)$. Since we need to stop our iterative process at some fixed step, let's choose

$$\widehat{t} := \min\{k \in \mathbb{N} : k \geq \log_a(A/B)\} \tag{69}$$

Corollary 5.2.1 (Error bound in the Gaussian setting for IGT)**.** *Assume that the columns of $\boldsymbol{X}$ have unit L^2-norm, i.e. $\|\boldsymbol{X}_j\|_2 = 1$ for $j = 1, \cdots, p$, and that the noise $\boldsymbol{\epsilon}$ follows a $\mathcal{N}(\mathbf{0}, \sigma\boldsymbol{I}_n)$ Gaussian distribution. Assume also that Assumption (64) holds and that*

$$1 < a \leq \frac{c}{\delta(1+2c)}$$

Then, for any iteration t larger than $\widehat{t}$ defined by (69), with probability larger than $1-2e^{-L}$, the estimator $\widehat{\beta}^t$, with A, B given by (68), fulfills $|\widehat{\beta}^t_{\bar{m}^*}|_0^G \leq c^2|\beta^*|_0^G$ and

$$\|\widehat{\beta}^t - \beta^*\|_2^2 \leq C_{a,c}|\beta^*|_0^G\sigma^2\left(1+2\phi_G^2(\log M + L)\right) \tag{70}$$

with

$$C_{a,c} = 2\left(\frac{2a(1+2c)}{a-1}\right)^2$$

Proof. Only we need to do is to show that (65) holds with probability at least $1-2e^{-L}$, we first check that $\|\boldsymbol{X}_{G_j}\|_F^2 = |G_j|$. In fact, since the columns of $\boldsymbol{X}$ have unit L^2-norm

$$\|\boldsymbol{X}_{G_j}\|_F^2 = \sum_{k\in G_j}\|X_k\|_2^2 = |G_j|$$

In addition, $x \to \|\boldsymbol{X}_{G_j}x\|_2$ is $|\boldsymbol{X}_{G_j}|_2$-Lipschitz. Then define $Z_{G_j} := \boldsymbol{X}_{G_j}^T\boldsymbol{\epsilon}$ and by Theorem 1.2.2, it concludes that

$$\|Z_{G_j}\|_2 \leq \mathbb{E}\left[\|Z_{G_j}\|_2\right] + \frac{\sigma}{|\boldsymbol{X}_{G_j}|_2}\sqrt{2\xi}$$

where $\xi \sim \text{Exp}(1)$. As a result, for any $u \geq 0$, it has

$$\mathbb{P}\left(\|Z_{G_j}\|_2 \geq \sigma\left(\sqrt{|G_j|} + |\boldsymbol{X}_{G_j}|_2\sqrt{2u}\right)\right) \leq e^{-u}$$

Hence, choose $u = \log M + L$, it concludes that

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\cdots,M} \frac{\|Z_{G_j}\|_2}{\sqrt{|G_j|}} \geq \sigma\left(1+\phi_G\sqrt{2\log M + 2L}\right)\right) \\
&\leq \sum_{j=1}^{M} \mathbb{P}\left(\|Z_{G_j}\|_2 \geq \sigma\left(\sqrt{|G_j|} + \sqrt{|G_j|}\phi_G\sqrt{2\log M + 2L}\right)\right) \\
&\leq \sum_{j=1}^{M} \mathbb{P}\left(\|Z_{G_j}\|_2 \geq \sigma\left(\sqrt{|G_j|} + |\boldsymbol{X}_{G_j}|_2\sqrt{2\log M + 2L}\right)\right) \\
&\leq e^{-L}
\end{aligned}$$

Now, we have shown that

$$\mathbb{P}\left(B < \frac{a}{a-1}|\boldsymbol{X}^T\boldsymbol{\epsilon}|\right) \leq e^{-L}$$

Moreover, we also need to show that

$$\mathbb{P}\left(B > \frac{a}{a-1}|\boldsymbol{X}^T\boldsymbol{\epsilon}|_\infty^G\right) \geq e^{-L}$$

And we have show that

$$\mathbb{P}\left(A \geq \frac{\|\beta^*\|_2}{(1+2c)\sqrt{|\beta^*|_0}}\right) \geq 1 - e^{-L}$$

in Corollary 5.1.1, which completes our proof. □

6 Multivariate Regression

So far, we discuss a lot about the simple linear regression model and its relative computation methods such as convex relaxation and iterative algorithm. In this section, we will focus on multivariate linear model, i.e.

$$y^{(i)} = A^T x^{(i)} + \epsilon^{(i)} \quad \text{for } i = 1, \cdots, m \tag{71}$$

where $y^{(i)}, \epsilon^{(i)} \in \mathbb{R}^n$, $x \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times n}$. Hence, rewrite it into matrix form by

$$\underbrace{\begin{pmatrix} (y^{(1)})^T \\ \vdots \\ (y^{(m)})^T \end{pmatrix}}_{Y\in\mathbb{R}^{m\times n}} = \underbrace{\begin{pmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{pmatrix}}_{X\in\mathbb{R}^{m\times p}} A + \underbrace{\begin{pmatrix} (\epsilon^{(1)})^T \\ \vdots \\ (\epsilon^{(m)})^T \end{pmatrix}}_{E\in\mathbb{R}^{m\times n}} \tag{72}$$

As a result, the statistical model is $Y = XA + E$ with i.i.d. $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$, in order to find the best estimation A^{MLE} of A, then by maximum likelihood estimation, it concludes that

$$A^{\text{MLE}} = \arg\min_{A\in\mathbb{R}^{p\times n}} \left\{ \prod_{i=1}^{m} \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \|y^{(i)} - A^T x^{(i)}\|_2^2 \right) \right\} \tag{73}$$

Take the log-likelihood function, we obtain that

$$A^{\text{MLE}} = \arg\min_{A\in\mathbb{R}^{p\times n}} \left\{ \sum_{i=1}^{n} \|y^{(i)} - A^T x^{(i)}\|_2^2 \right\} \tag{74}$$

Since $\|M\|_F = \sum_{i,j} M_{i,j}^2 = \text{tr}\left(M^T M\right)$, we conclude that

$$A^*_{\text{MLE}} = \arg\min_{A\in\mathbb{R}^{p\times n}} \left\{ \|Y - XA\|_F^2 \right\}$$

Remark 6.0.1. *Denote A_k to be the k-th column of A, so*

$$A_k^{\text{MLE}} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y_k - X\beta\|_2^2 \right\} \quad \text{for } k = 1, \cdots, n$$

which is a simple liner regression.

6.1 Sparsity Estimation

Coordinate sparsity:

Assume that $|A|_0 = \text{card}\{(i,j) : A_{i,j} \neq 0\}$, which is quite small. Then similar as the model selection in §2, we add a L^1 penalty as the following form.

$$\widehat{A}^{L^1} := \arg\min_{A\in\mathbb{R}^{p\times n}} \left\{ \|Y - XA\|_F^2 + \lambda |A|_1 \right\}$$

where $|A|_1 := \sum_k |A_k|_1$, which can be separated into n following lasso problems.

$$\widehat{A}_k^{L^2} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y_k - X\beta\|_F^2 + \lambda |\beta|_1 \right\}$$

Row sparsity:

Assume that $\operatorname{card}\{j : A_j \neq \mathbf{0}\}$ is small, hence

$$y^{(i)} = \sum_{j=1}^{p} A_{j,i}^T x_j^{(i)} + \epsilon^{(i)}$$

and the estimation result is

$$\widehat{A}^{\mathrm{RS}} := \arg\min_{A \in \mathbb{R}^{p \times n}} \left\{ \|Y - XA\|_F^2 + \lambda \sum_{j=1}^{p} |A_j|_1 \right\}$$

In fact, you may find above minimization problem is quite similar as group lasso discussed in §4. Before giving estimation method, some backgrounds about matrix algebra is necessary. For $M \in \mathbb{R}^{d \times n}$, define

$$\operatorname{vect}(M) := \begin{pmatrix} M_1 \\ \dots \\ M_n \end{pmatrix} \in \mathbb{R}^{dn}$$

Then, it has

$$\operatorname{vect}(Y) = \underbrace{\begin{pmatrix} X & & \\ & \ddots & \\ & & X \end{pmatrix}}_{:= \widetilde{X} \in \mathbb{R}^{mn \times np}} \operatorname{vect}(A) + \operatorname{vect}(E)$$

Next, setting $G_j := \{k : k \equiv j \mod p\}$ and it concludes that the group lasso estimation is

$$\operatorname{vect}(\widehat{A}^{RS}) := \arg\min_{\beta \in \mathbb{R}^{np}} \left\{ \|\operatorname{vect}(Y) - \widetilde{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta_{G_j}\|_2 \right\}$$

6.2 Low Rank Regression

Reminder on SVD:

First, let's remind the singular value decomposition (SVD) and its relative properties. For $A \in \mathbb{R}^{n \times p}$, suppose its SVD is $A = \sum_{j=1}^{r} s_j(A) u_j v_j^T$ with $s_1(A) \geq \cdots \geq s_r(A)$. Hence the Frobenius of A is defined by

$$\|A\|_F^2 := \sum_{i,j=1}^{n,p} A_{i,j}^2 = \operatorname{tr}(A^T A) = \sum_{k=1}^{r} s_k(A)^2$$

Moreover, for any integer $q \geq 1$, the Ky–Fan $(2, q)$-norm is defined by

$$\|A\|_{(2,q)}^2 := \sum_{k=1}^{q} s_k(A)^2$$

Since $\|A\|_{(2,q)} \leq \|A\|_F$ for any q, it concludes that $A \to \|A\|_{(2,q)}$ is Lipschitz with respect to the Frobenius norm. Specially, for $q = 1$, we have

$$\|A\|_{(2,1)} = s_1(A) = \sup_{\|x\|_2=1} \|Ax\|_2 = |A|_2$$

In fact, we can show the following properties.

(1.) For any matrices $A, B \in \mathbb{R}^{n\times p}$, we have

$$\langle A, B\rangle_F \leq \|A\|_{(2,r)}\|B\|_{(2,r)} \tag{75}$$

where $r := \min\{\text{rank}(A), \text{rank}(B)\}$ and $\langle A, B\rangle = B^T A$.

(2.) For any $q \geq 1$, we have

$$\min_{\text{rank}(B)\leq q} \|A - B\|_F^2 = \sum_{k=q+1}^{r} s_k(A)^2 \quad \text{for } q < \text{rank}(A) \tag{76}$$

And for $q \geq r$, the left side is equal to 0.

Proof. For the first term, suppose $\text{rank}(B) \leq \text{rank}(A)$, it is easy to see that

$$\langle A, B\rangle_F = \langle \text{Proj}_B A, B\rangle_F \leq \|\text{Proj}_B A\|_F \|B\|_F \leq \|A\|_{(2,r)}\|B\|_{(2,r)}$$

Since $|\text{Proj}_B|_2 \leq 1$, we can obtain the last inequality of above by

$$s_k(\text{Proj}_B A) = \max_{\dim(S)=k} \min_{\substack{\|x\|_2=1 \\ x\in S}} \|\text{Proj}_B Ax\|_2 \leq \max_{\dim(S)=k} \min_{\|x\|_2=1} |\text{Proj}_B|_2 \|Ax\|_2 \leq s_k(A)$$

For the second term, it has that

$$\|A - B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 - 2\langle A, B\rangle_F \geq \|A\|_F^2 + \|B\|_F^2 - 2\|A\|_{(2,r)}\|B\|_F$$

In addition, we have

$$\|B\|_F^2 - 2\|A\|_{(2,r)}\|B\|_F = (\|B\|_F - \|A\|_{(2,r)})^2 - \|A\|_{(2,r)}^2 \geq -\|A\|_{(2,r)}^2$$

The equality is attended when $\|B\|_F = \|A\|_{(2,r)}$ and hence we obtain the second term. □

Theorem 6.2.1 (Weyl inequality)**.** *For two $n \times p$ matrices A and B, we have for any $k \leq \min(n, p)$*

$$|s_k(A) - s_k(B)| \leq s_1(A - B) = |A - B|_2 \tag{77}$$

Proof. From the definition of s_k, it deduces that

$$s_k(A) = \max_{\dim(S)=k} \min_{\substack{\|x\|_2=1 \\ x\in S}} \|Ax\|_2 \leq \max_{\dim(S)=k} \min_{\substack{\|x\|_2=1 \\ x\in S}} (\|Ax\|_2 + \|(A-B)x\|_2) = s_k(B) + s_k(A-B)$$

Exchange A and B, then complete our proof. □

Reminder on Random Matrix:

Suppose $W \in \mathbb{R}^{m\times n}$ such that $W_{i,j}$ be i.i.d. $\mathcal{N}(0,1)$, then we have

Proposition 6.2.1 (Classical asymptotic property)**.** *For fixed n, let $m \to \infty$ and then it has*

$$\left[\frac{1}{m}W^TW\right]_{i,j} = \frac{1}{m}\sum_{k=1}^{m} W_{k,i}W_{k,j} \to 1_{i\neq j} \quad \text{a.s.} \tag{78}$$

i.e.

$$\lim_{m\to\infty} \frac{1}{m}W^TW = \boldsymbol{I}_n \quad \text{a.s.} \tag{79}$$

and for $k = 1, \cdots, n$, it has

$$\lim_{m\to\infty} s_k\left(\frac{1}{\sqrt{m}}W_{m\times n}\right) = 1 \quad \text{a.s.} \tag{80}$$

Proof. By the Law of Large Number, it is easy to obtain this proposition. □

Furthermore, we will introduce another forms of asymptotic properties.

Proposition 6.2.2 (Marchenko-Pastur Law)**.** *Suppose $W_{i,j}$ are independent identically distributed random variables with mean 0 and variance $\sigma^2 < \infty$, let*

$$Y_n := \frac{1}{m}W^TW$$

and $\lambda_1 > \cdots > \lambda_m$ be the eigenvalues of Y_n, consider the random measure

$$\mu_m(A) := \frac{1}{m}\#\{\lambda_j : \lambda_j \in A\} \quad A \subset \mathbb{R} \tag{81}$$

Then let $m, n \to \infty$ with $m/n \to \beta \in \mathbb{R}^+$, then $\mu_m \to \mu$ in distribution where

$$\mu(A) := \begin{cases} (1-1/\beta)1_{0\in A} + v(A) & \text{if } \beta > 1 \\ v(A) & \text{if } \beta \in (0,1] \end{cases} \tag{82}$$

and

$$dv(x) := \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\beta_+ - x)(\beta_- - x)}}{\beta x}1_{x\in[\beta_-,\beta_+]} \tag{83}$$

with $\beta_\pm := \sigma^2(1 \pm \sqrt{\beta})^2$.

Before giving the proof of Proposition 6.2.2, the following lemmas are necessary.

Lemma 6.2.1 (Stieltjes continuity theorem)**.** *Let μ_n be a sequence of random probability measures on the real line, and let μ be a deterministic probability measure. Then μ_n converges almost surely to μ in the vague topology if and only if $s_{\mu_n}(z)$ converges almost surely to $s_\mu(z)$ for every z in the upper half-plane of $\mathbb{C}$, where*

$$s_\mu(z) := \int_{\mathbb{R}} \frac{1}{x-z} dx$$

Proof. In fact, if $\mu_n \to \mu$ in distribution as $n \to \infty$, the only if part is easy to obtain. For the if part, take $\phi \in C_c(\mathbb{R})$ be any test function □