BST235: Advanced Regression and Statistical Learning

Lecture Notes

(This version: August 29, 2021)

Junwei Lu

Spring 2020

HARVARD
UNIVERSITY

Contents

*

Lecture 1

Introduction

1.1 What is Big Data?

The main theme of the lecture is the modern methods and theory of big data analysis. In general, our course aim to (ambitiously) solve two major questions:

1. *How to analyze big data? (Method)*
2. *Why it works? (Theory)*

To clarify the questions above, we need to define what is "big data". In our notes, big data is almost a synonym of "high dimensional data". The dataset, usually denoted as $\mathbb{X}$, is a $n \times d$ matrix, where n is the sample size and d is the number of features (or feature dimension). People summarize the following properties of big data.

Big data has three important features. (1)**Volume**: both the sample size n and feature dimension d are both very large. (2) **Velocity**: the scale of the data is so large such that the computation efficiency is crucial. (3)**Variety**: the datasets are heterogenous and of various types.

When talking about **high dimensional statistics**, we typically emphasize the first feature and the feature dimension d is usually even (much) larger than the sample size n.

There are many tasks when analyzing big data. A typical protocol of data analysis is as follows

$$\begin{array}{ccccccc} \text{Data} & \Longrightarrow & \text{Models} & \Longrightarrow & \text{Algorithms} & \Longrightarrow & \text{Decision Making} \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \text{Probability} & & \text{Statistical Learning} & & \text{Optimization} & & \text{Inference.} \end{array}$$

When analyzing data, of course we will start with datasets. Every statistician, as long as most people when viewing ratings on Amazon, have the intuition that the more samples we have in the dataset, the better information we will have. This intuition is supported by the probability theory by assuming the data is generated from specific statistical or machine learning models. In the big data analysis, both the sample size and feature dimensions are much larger than the classical analysis. We need to build high dimensional statistical models. And when the model is specified, we need concrete algorithms to solve the model. Since the datasets are of large scale, optimization theory on how to develop fast algorithms is needed. Finally, we need make decisions like rejecting the hypotheses and assessing the quality of our estimation. Statistical inference theory on the confidence interval and hypotheses testing is needed.

1.2 Three Principles

Seeing the typical protocol of data analysis above, we can see the the four major cornerstones of modern big data analysis is: probability, statistical learning, optimization and inference, which therefore will be the four main topics of our course. Before diving into the details of the four topics, we would like to first talk about some philosophy. There are three major principles we will use for multiple times in our lecture. They may seem vague the first you see them but we promise that we will make it rigorous when we implement these principles in different scenarios of big data analysis.

Concentration Principle: Random observations will converge to the population truth when the sample size becomes larger.

Immediate examples of concentration principle are the law of large numbers and the central limit theory. They are usually used in the theory of estimation and inference. However, to handle high dimensional data, there are much richer probability theory developed beyond the law of large numbers and the central limit theory, which will be the major topic in the next chapter.

Parsimonious Principle: Although the feature dimension d is large, only a small proportion of features really play a role.

This principle plays a key role in high dimensional statistics. It states that the high dimensional datasets are only seemingly to be high dimensional. There is an essentially low-dimensional structured inside! It is a statistical version of the "Occam's razor": if two explanations are equally good, we prefer the simpler one. If two statistical models have a similar prediction accuracy, we always prefer the simpler and more interpretable one. This principle is also useful to avoid overfitting in data analysis. The word "simple" in the

parsimonious principle has different meanings under different statistical models. We list two examples below for what do we mean by simpler models.

1.1 Example (Sparsity). Consider the linear model

$$Y = \sum_{j=1}^{d} X_j \beta_j + \varepsilon.$$

We impose the sparsity assumption that only s of β_j's are nonzero and $s \ll d$. □

Sparse linear model above is the most straightforward example of the parsimonious principle. Instead of considering a d-dimensional model, the sparse linear model only consider $\binom{d}{s}$ number of s-dimensional models. However, the parsimonious principle is not restricted to sparsity. It includes all regularization structures which make the model simple.

1.2 Example (Additive Model). The additive model assumes that the general nonparametric model $Y = f(X_1, \ldots, X_d) + \varepsilon$ has the additive structure:

$$Y = \sum_{j=1}^{d} f_j(X_j) + \varepsilon.$$

The sparse additive model further assumes that only s of these d nonlinear functions $f_j(x)$ is not zero. □

Taylor Principle: All functions (we are interested in) are "almost" quadratic.

The last principle gets its name by Taylor expansion. Although there are a variety of functions, in our lecture, we can find that most of the analysis is almost same as the quadratic functions. In the statistical analysis, the Taylor principle implies that all (log)-likelihood function is almost least square. In the optimization, we will introduce the concept of convexity and smoothness, which also states that the objective function is almost quadratic.

1.3 Basics of Probability

When we roll a die, the small cube with six different numbers on its six faces will never give us a determinist answer before it comes to rest. Einstein thought even "God does not throw dice". However, probabilists have to (as well as statisticians!). Actually, they have developed a dedicated language to describe the world of uncertainty. We will first begin with reviewing several fundamental terminologies in probability theory.

1.3 Definition (Statistical Model). *A statistical model $\mathcal{P}$ is a set of probability distributions indexed by a parameter θ. We denote $\mathcal{P} := \{p_\theta(x) \,|\, \theta \in \Theta\}$. The set Θ is called a parameter space.*

- *When the parameter space Θ is finite-dimensional, we call the statistical model as a **parametric model**.*
- *When the parameter space Θ is infinite-dimensional, we call the statistical model as a **nonparametric model**.*

1.4 Definition (Random Sample). *The random variables $X_1, \ldots, X_n$ are called a random sample from $p_\theta(x)$ if $X_1, \ldots, X_n$ are independent and $X_i \sim p_\theta(x)$ for every $i = 1, 2, \ldots, n$. In other words, $X_1, \ldots, X_n$ are independent and identically distributed random variables. (We usually use the abbreviation i.i.d.)*

1.5 Definition (Data). *In statistics, data $x_1, \ldots, x_n$ is the realization of the random sample $X_1, \ldots, X_n$.*

People always get confused about these two concepts: random sample and data. A random variable is like a random number generator. For example, a die is a machine to generate the number from 1 to 6 with equal probability. Suppose you want to roll a die for n times, as the results of the n rollings are uncertain, you have a random sample $X_1, \ldots, X_n$. You can study its probability properties like distribution functions, expectation, variance, etc. When you do the experiment to roll a die for n times and get n numbers $x_1, \ldots, x_n$, this n numbers is a realization of the random sample, namely the data. Therefore, data are concrete numbers. It is meaningless to discuss the distribution of data as they are not random anymore.

1.6 Definition (Distribution Functions). *The cumulative distribution function (cdf) of a random variable X is defined as*

$$F(x) := \mathbb{P}(X \leq x).$$

The density function or probability density function (pdf) $p(x)$ of a random variable X satisfies

$$p(x) = \frac{d}{dx} F(x).$$

For discrete random variables, the cumulative distribution functions are not differentiable. We define the probability mass function (pmf) $p(x)$ of a discrete random variable X as

$$p(x) = \mathbb{P}(X = x), \quad \textit{for all } x \in X(\Omega).$$

A note on notation: We use $p_\theta(x)$ as a generic representation to represent a distribution function. It could be a cdf, pdf or pmf. $X \sim p_\theta(x)$ means the random variable X has the distribution as $p_\theta(x)$.

1.7 Definition (Expectation and Variance). *If a continuous random variable X has the*

density function $p(x)$, the expectation and variance of X is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xp(x)dx;$$

$$\textit{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 p(x)dx.$$

When we have a random vector X, the covariance matrix of X is defined as

$$\Sigma := \textit{Cov}(X) = \mathbb{E}(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T.$$

1.8 Definition (Statistic). *Let $X_1, \ldots, X_n$ be a random sample and $T(x_1, \ldots, x_n)$ is a function from $\mathbb{R}^n$ to $\mathbb{R}$, we call the random variable $T(X_1, \ldots, X_n)$ a statistic. The distribution of the statistic $T(X_1, \ldots, X_n)$ is called the sampling distribution. When a statistic is an estimator of a parameter θ, we usually denote it as $\widehat{\theta}$*

1.9 Example (Sample Mean and Covariance). The average of the random sample $X_1, \ldots, X_n$ is called sample mean, and we denote it by

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

The sample covarianc of the random sample is defined by

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T.$$

□

1.3.1 Asymptotic Theory

We will never know which number it will turn out to be when we rolling a die, but if we roll a fair die for sufficiently many times, we will find the average of all the results approaches $3.5 = (1+2+3+4+5+6)/6$. In general, when we get more and more data, the random sample will have some asymptotic properties.

In order to describe the asymptotic behavior of random variables, we should first define what is the limitation of random variables.

1.10 Definition (Converge in Probability). *A sequence of random variables $X_1, X_2, \ldots$ converges to μ in probability if*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mu| > \epsilon) = 0, \quad \textit{for all } \epsilon > 0.$$

We denote it as $X_n \xrightarrow{P} X$.

Another important asymptotic property of random variable is to describe the limit distribution of random variables. We first define another kind of convergence of random variables.

1.11 Definition (Converge in Distribution). *A sequence of random variables $X_1, X_2, \ldots$ have cdf $F_1, F_2, \ldots$. $X_1, X_2, \ldots$ converges in distribution to a random variable X with cdf $F(x)$ if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all x such that $F(x)$ is continuous. We denote it as

$$X_n \xrightarrow{d} F(x).$$

1.12 Definition (Consistent Estimator). *A statistic $\widehat{\theta}_n$ is a consistent estimator of θ if $\widehat{\theta}_n \xrightarrow{P} \theta$.*

The consistency of an estimator is an example of how to describe the concentration principle by the probability language. However, it is not accurate enough as it does not have the information of how fast the estimator converges. Therefore, we introduce the O_P notation below.

1.13 Definition (O_P notations). *Given a fixed sequence a_n and a sequence of random variables $X_1, X_2, \ldots$, we denote $X_n = o_P(a_n)$ if $X_n/a_n \xrightarrow{P} 0$.*

We denote $X_n = O_P(a_n)$ if for any $\epsilon > 0$, there exist $C > 0$ and $N > 0$ such that for all $n > N$,

$$\mathbb{P}(|X_n/a_n| > C) < \epsilon.$$

Notice that the definition of O_P is weaker than the argument that there exists a constant $C > 0$ such that $\lim_{n\to\infty} \mathbb{P}(|X_n/a_n| > C) = 0$ (Think about why). However, the later definition is easier to use in practice.

1.14 Definition (Statistical Rate). *We say an estimator $\widehat{\theta}$ has the statistical rate $O(a_n)$ if $|\widehat{\theta} - \theta| = O_P(a_n)$.*

1.4 Basics of Linear Algebra

Following the parsimonious principle, we prefer the simpler explanations. The simplest structure in mathematics is linearity. Now we review the basic notations and terminologies in linear algebra.

A is an $m \times n$ real matrix, written $A \in \mathbb{R}^{m \times n}$, if

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

where $a_{ij} \in \mathbb{R}$. The (i, j)th entry of A is $A_{ij} = a_{ij}$.

The *transpose* of $A \in \mathbb{R}^{m \times n}$ is defined as

$$A^T = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{mn} \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

In other words, $(A^T)_{ij} = A_{ji}$. Note: $\boldsymbol{x} \in \mathbb{R}^n$ is considered to be a column vector in $\mathbb{R}^{n \times 1}$.

1.15 Definition (Sums and products of matrices). *The sum of matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times n}$ is the matrix $A + B \in \mathbb{R}^{m \times n}$ such that*

$$(A + B)_{ij} = A_{ij} + B_{ij}.$$

The product of matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times \ell}$ is the matrix $AB \in \mathbb{R}^{m \times \ell}$ such that

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

The $n \times n$ *identity matrix*, denoted I_n or I for short, is

$$I = I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

We have $IA = A = AI$.

If it exists, the *inverse* of A, denoted A^{-1}, is a matrix such that $A^{-1}A = I$ and $AA^{-1} = I$. If A^{-1} exists, we say that A is *invertible*. We have $(A^{-1})^T = (A^T)^{-1}$ and $(AB)^{-1} = B^{-1}A^{-1}$. The *trace* of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(A)$, is defined as $\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$. We have $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ if AB is a square matrix. We say A is *symmetric* if $A = A^T$. The rank of a matrix A is the dimension of A's column space.

1.4.1 Eigenvalues and Eigenvectors

Given two vectors $u, v \in \mathbb{R}^d$, the *inner product* is $\langle u, v\rangle := u^T v = \sum_{j=1}^d u_j v_j$. We define the ℓ_2-norm of $\mathbf{u}$ as $\|u\| = \sqrt{|\langle u, u\rangle|}$. The cosine of the angle between u and v is $\cos\alpha = \frac{\langle u,v\rangle}{\|u\|\|v\|}$. We say u is orthogonal to v, denoted as $u \perp v$ if $\langle u, v\rangle = 0$.

1.16 Definition (Eigenvalues and Eigenvectors). *We say λ and v are the eigenvalue and eigenvector of a matrix A if $Av = \lambda v$.*

1.17 Theorem (Eigenvalue Decomposition). *A symmetric matrix $A \in \mathbb{R}^{d\times d}$ has*

- *Real eigenvalues: $\lambda_1 \geq \ldots \geq \lambda_d$;*
- *Orthonormal eigenvectors: $u_1, \ldots, u_d$ such that $\|u_i\| = 1$ and $u_i \perp u_j$ for all $1 \leq i < j \leq d$,*

such that we have the eigenvalue decomposition $A = \sum_{i=1}^d \lambda_i u_i u_i^T$.

Denote $U = [u_1, \ldots, u_d] \in \mathbb{R}^{d\times d}$ and

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d) := \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix}.$$

The eigenvalue decomposition can also be written as $A = U\Lambda U^T$. We call U is an *orthogonal matrix* as $U^T U = I_d$.

As a cornerstone of our lecture, we prefer to interpret the concepts as a solution of an optimization problem. Such interpretation is usually called variational form. The following theorem gives a variational form of eigenvalues.

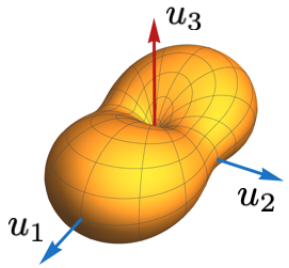


Figure 1.1. *Variational form of eigenvalues*

1.18 Theorem (Variational Form of Eigenvalues). *Given a symmetric matrix A same as*

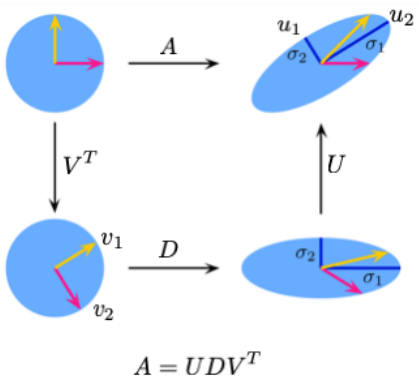


Figure 1.2. *Visualization of singular value decomposition.*

Theorem 1.17, its maximum eigenvalue has

$$\lambda_{\max}(A) := \lambda_1 = \max_{\|x\|\leq 1} x^T Ax = \max_x \frac{x^T Ax}{x^T x} \text{ and } u_1 = \arg\max_{\|x\|\leq 1} x^T Ax$$

and its minimum eigenvalue has

$$\lambda_{\min}(A) := \lambda_d = \min_{\|x\|\leq 1} x^T Ax = \min_x \frac{x^T Ax}{x^T x} \text{ and } u_d = \arg\min_{\|x\|\leq 1} x^T Ax$$

Question: What is the variational form of other eigenvalues? See the visualization in Figure 1.1.

The concept of eigenvalue decomposition can be generalized to non-symmetric or even non-square matrices. We have the following theorem on singular value decomposition.

1.19 Theorem (Singular Value Decomposition). *Given a rank r matrix $A \in \mathbb{R}^{n\times d}$, there exist*

- *Singular values: $\sigma_1 \geq \ldots \geq \sigma_r > 0$ and denote $D = diag(\sigma_1, \ldots, \sigma_r)$;*
- *Orthogonal matrices: $U = [u_1, \ldots, u_r] \in \mathbb{R}^{n\times r}$ and $V = [v_1, \ldots, v_r] \in \mathbb{R}^{d\times r}$ satisfying $U^T U = V^T V = I_r$,*

such that we have the singular value decomposition (SVD)

$$A = UDV^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

We can see that the eigenvalue decomposition is a special SVD. The visualization of the SVD is shown in Figure 1.2[1] The matrix A as a linear map transforms the two canonical

[1]Credit to the wikipedia page `https://en.wikipedia.org/wiki/Singular_value_decomposition`

unit vectors (yellow and red vectors)in the top left plot to the top right plot. This map can be decomposed into three steps: (1) V^T: rotating yellow and red vectors to v_1 and v_2; (2) D: scaling the disc by σ_1 horizontally and σ_2 vertically; (3) U: rotating two canonical unit vectors (blue lines) to u_1 and u_2.

Lecture 2

Sub-Gaussian Random Variables

2.1 Asymptotic versus Non-Asymptotic

In the previous lecture, we discussed the concentration principle. It states that the more samples we have, the random observations converge to the population truth. In particular, we have the two important theorems in the probability theory describing this phenomenon.

Let $X_1, \ldots, X_n, \ldots$ be i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$ for $i = 1, 2, \ldots$. Define the sample mean estimator $\bar{X}_n = (X_1 + \ldots + X_n)/n$. The first result describes the phenomenon that the average of random variables will converge to the expected value.

2.1 Theorem (Law of Large Numbers). *The average $\bar{X}_n$ converges to mean, μ, in probability, i.e.,*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

Another important concentration property of random variable is to describe the limit distribution of the sample mean.

2.2 Theorem (Central Limit Theorem). *The average $\bar{X}_n$ converges in distribution to the normal distribution. In particular, we have*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0, 1).$$

Notice that both the law of large numbers and central limit theorem are the theorems characterizing the property of sample mean when the sample size n goes to infinity. We usually call this type of result the asymptotic property of random variables.

Although asymptotic properties like law of large numbers and central limit theorem tell us that some estimators converge to the truth, it has two major problems.

1. Asymptotic properties do not have the information of convergence rate;
2. Asymptotic properties may fail in high dimension when both n and d go to infinity.

The definition of convergence in probability or in distribution does not tell us how fast it converges. The law of large numbers implies that $\bar{X}_n$ is a consistent estimator while we do not know the statistical rate. The central limit theorem actually tells us the rate $|\bar{X}_n - \mu| = O_P(1/\sqrt{n})$. (Think about why) However, this is only true when n goes to infinity. The second problem of the asymptotic properties are more severe. We would like to show why it may fail in high dimension by the following example.

2.3 Example (Gaussian Mean Model). Let $X_1, \ldots, X_n$ be i.i.d. d-dimensional normal random vectors with the distribution $N(\mu, \sigma^2 I_d)$. We estimate the mean vector $\mu = (\mu_1, \ldots, \mu_d)^T$ by the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. It is easy to check the sample mean is also normal, i.e., $\bar{X}_n \sim N(\mu, \sigma^2/n \cdot I_d)$. We measure the quality of the sample mean estimator by the mean square error MSE $:= \mathbb{E}\|\bar{X}_n - \mu\|^2$. Using the covariance structure and the definition of ℓ_2-norm, we have

$$\mathbb{E}\|\bar{X}_n - \mu\|^2 = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^d \mathbb{E}(X_{ij} - \mu_j)^2 = \sigma^2 \frac{d}{n}.$$

Since we have the closed form of the mean squared error, we can discuss it under different settings.

- When $d = 1$ or d is finite, we have MSE $= O(1/n)$ and the sample mean is a consistent estimator.
- When $d = \sqrt{n}$, we have MSE $= O(1/\sqrt{n})$ and the sample mean is still consistent but the rate becomes slower.
- When $d = n$, we have MSE = constant and no longer converges as n goes to infinity.
- When $d = n^2$, MSE is diverging.

□

The example above shows that the asymptotic results may not be strong enough when we want to study the behaviour of estimators when both d and n diverge. An idea to solve this problem is to first fix d and n as constants and see how the rate depends on them. We are interested in looking into the upper bound of tail probability like

$$\mathbb{P}(\|\bar{X}_n - \mu\|^2 > t) \leq 2de^{-nt^2}$$

for any $t > 0$ and fixed d and n. We call this type of inequalities **non-asymptotic concentration inequalities** as it is true for all finite n and d.

2.2 Sub-Gaussian Random Variables

Let us start with the simplest tail probability inequality.

2.4 Theorem (Markov Inequality). *Let X be a non-negative random variable. For any $t > 0$, we have*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}X}{t}.$$

Proof. We prove the inequality for X when its pdf exists and it is not hard to use a similar idea to prove the general case.

$$\mathbb{P}(X > t) = \int_t^\infty p(x)dx \leq \int_t^\infty \frac{x}{t}p(x)dx \leq \frac{1}{t}\int_{-\infty}^\infty xp(x)dx = \frac{\mathbb{E}X}{t}.$$

□

We can see that the Markove inequality shows that the tail probability decays at the rate $1/t$. Actually, we can achieve faster decaying rate by applying Markov inequality to $|X - \mathbb{E}[X]|^2$ and get the following **Chebyshev Inequality**:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}(|X - \mathbb{E}[X]|^2 > t^2) \leq \frac{\mathbb{E}|X - \mathbb{E}[X]|^2}{t^2} = \frac{\text{Var}(X)}{t^2}.$$

This gives us a $1/t^2$ decaying rate! You can probably imagine that we can actually further improve the rate by considering the k-th moment:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}(|X - \mathbb{E}[X]|^k > t^k) \leq \frac{\mathbb{E}|X - \mathbb{E}[X]|^k}{t^k}.$$

We now have a $1/t^k$ decaying rate for arbitrarily large k! Before we get too excited, we need to notice that there is a price to pay to get a faster and faster rate: the tail probability has the $1/t^k$ rate only if the kth moment of X is finite. Although it is not exactly a free lunch, the lunch is cheap enough: most random variables in reality is not that heavy-tailed.

So let us be aggressive again. Why only polynomial? Why not use the exponential function? This leads us to the so-called **Chernoff bound**:

$$\mathbb{P}(X - \mathbb{E}[X] > t) = \mathbb{P}(e^{\lambda(X-\mathbb{E}[X])} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]}{e^{\lambda t}}, \text{for any } t \geq 0, \lambda \geq 0.$$

We now have an exponential decaying rate. Notice that we can also choose an optimal $\lambda > 0$ such that the right hand side of the tail probability inequality is minimized. Let us formalize this idea in the following theorem.

2.5 Theorem (Chernoff Bound). *Define the log-moment generating function of X and its Legendre dual as*

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}], \quad \psi^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi(\lambda)\}.$$

We have $\mathbb{P}(X - \mathbb{E}[X] > t) \leq e^{-\psi^*(t)}$ *for all* $t \geq 0$.

Let's look at a concrete example for the Gaussian distribution.

2.6 Example (Gaussian). Suppose $X \sim N(\mu, \sigma^2)$. We know that the moment-generating function for X is $\mathbb{E}[e^{\lambda(X-\mu)}] = e^{\frac{\lambda^2\sigma^2}{2}}$ and the log-moment generating function is $\psi(\lambda) = \lambda^2\sigma^2/2$. We can get the closed form of its Legendre dual as

$$\psi^*(t) = \sup_{\lambda>0}\left\{\lambda t - \frac{\lambda^2\sigma^2}{2}\right\} = \frac{t^2}{2\sigma^2} \text{ by choosing } \lambda = \frac{t}{\sigma^2}$$

Then Chernoff bound gives us $\mathbb{P}(X - \mu > t) \leq \exp(-t^2/(2\sigma^2))$. The tail probability of Gaussian distribution is similar to its density function. □

The Gaussian example above motivates us to consider the random variables with the tail probability similar to the Gaussian distribution.

2.7 Definition (Sub-Gaussian). *A random variable* X *is sub-Gaussian with the variance proxy* σ^2*, if*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\frac{\lambda^2\sigma^2}{2}}, \quad \textit{for all } \lambda \in \mathbb{R}.$$

Notice that the range of λ becomes the real line instead of only being positive in Theorem 2.5. This is because $\lambda \geq 0$ only gives us right-sided tail probability and we will get the the other side by considering $\lambda < 0$.

2.8 Theorem (Two-Sided Tail Probability). *If* X *is sub-Gaussian with the variance proxy* σ^2*, we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Proof. We get the right-sided tail probability $\mathbb{P}(X - \mathbb{E}[X] > t) \leq e^{-\frac{t^2}{2\sigma^2}}$ by directly using the Chernoff bound. We can bound the left-side tail probability by considering $-X$:

$$\mathbb{P}(X-\mathbb{E}[X] < -t) = \mathbb{P}(-X-\mathbb{E}[-X] > t) \leq \frac{\mathbb{E}[e^{-\lambda(X-\mathbb{E}[X])}]}{e^{\lambda t}} \leq e^{-\lambda t+\frac{\lambda^2\sigma^2}{2}}, \text{ for all } \lambda > 0.$$

So we choose the optimal $\lambda = t/\sigma^2$ to minimize the tail probability and get $\mathbb{P}(X - \mathbb{E}[X] < -t) \leq e^{-\frac{t^2}{2\sigma^2}}$. We then have the two-sided tail probability bound by using union bound. □

From the proof of two-sided tail probability above, we find that the positive and negative domains of the moment generating function imply the right and left-sided tail probability respectively. In the following of our lecture notes, we sometimes only prove the one-sided

tail probability of a random variable X and the other side can be proved similarly by considering $-X$.

Besides Gaussian random variables, our next example of sub-Gaussian random variable is the bounded random variable.

2.9 Theorem (Bounded Random Variables). *If X is a bounded random variable, such that $a \leq X \leq b$ almost surely, for some $a, b \in \mathbb{R}$, then X is a sub-Gaussian with variance proxy $\frac{(b-a)^2}{4}$.*

Proof. Intuitively, this theorem is obvious as the density of X goes directly to 0 when $t > b$, so it must has sub-Gaussian tail when t is large enough. However, the spirit of non-asymptotic concentration inequality is to prove the result for all $t > 0$ instead of letting t goes to infinity. So let us prove it by the definition of the sub-Gaussian.

Without loss of generality, we assume $\mathbb{E}[X] = 0$. Our goal is to find an upper bound of the log-moment generating function $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$. Our idea is to study its first and second derivatives, as the Taylor principle tells us that we should consider its quadratic approximation. We have

$$\psi'(\lambda) = \frac{d}{d\lambda} \log \mathbb{E}[e^{\lambda X}] = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]},$$
$$\psi''(\lambda) = \frac{d^2}{d\lambda^2} \log \mathbb{E}[e^{\lambda X}] = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2.$$

Define a new measure $dQ = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]} dP$ and we find that the derivatives of ψ become the expectation and variance under the new measure:

$$\psi'(\lambda) = \mathbb{E}_Q[X] \text{ and } \psi''(\lambda) = \text{Var}_Q(X).$$

We can bound the variance by

$$\psi''(\lambda) = \text{Var}_Q(X) = \text{Var}_Q\Big(X - \frac{a+b}{2}\Big) \leq \mathbb{E}_Q\Big[X - \frac{a+b}{2}\Big]^2 \leq \frac{(b-a)^2}{4},$$

where we use $a \leq X \leq b$ in the last inequality.

Notice that $\psi(0) = \log 1 = 0$ and $\psi'(0) = \mathbb{E}[X] = 0$. Using the fundamental theorem of calculus, we bound the log-moment generating function by

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(p) dp d\mu \leq \frac{\lambda^2 (b-a)^2}{8}.$$

□

Notice that the fact that $\psi''(\lambda) = \text{Var}_Q(X)$ in the proof above is generally true for all random variables. This gives us an interesting insight: while Chebyshev inequality can only give a $1/t^2$-tail with a bound in variance, the Chernoff bound can give us a sub-Gaussian

tail by also bounding the variance (everything is almost quadratic!) only under a different measure.

We show the concentration of sample mean of sub-Gaussian random variables.

2.10 Theorem (Hoeffding Inequality). *Let $X_1, \dots, X_n$ be independent random variables such that X_i is sub-Gaussian with variance proxy σ^2 for all $i = 1, \dots, n$. We have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X > t\right) \le e^{-\frac{nt^2}{2\sigma^2}}.$$

Proof. Again without loss of generality we assume $\mathbb{E}[X] = 0$. By Chernoff bound, we have

$$\mathbb{P}(\bar{X}_n > t) \le \mathbb{E}(e^{\lambda \sum_{i=1}^n X_i})e^{-n\lambda t} = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}]e^{-n\lambda t} \le e^{n\left(\frac{\sigma^2\lambda^2}{2} - \lambda t\right)},$$

where first equality is due to independence of X_i's and the last inequality is due to sub-Gaussian. So we finish the proof by choosing $\lambda = \frac{t}{\sigma^2}$. □

Similar to Theorem 2.8, we also have the two-sided result

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X\right| > t\right) \le 2e^{-\frac{nt^2}{2\sigma^2}}.$$

We can write it in a high probability statement: with probability at least $1 - \delta$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X\right| \le \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}, \text{ or } \left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X\right| = O_P(\sqrt{\sigma^2/n})$$

when we want to go back to the asymptotic statement.

Lecture 3

Sub-Exponential Random Variables

3.1 Concentration Beyond Average

In the previous lecture, we showed the concentration of sample average in the Hoeffding inequality. The asymptotic results like law of large numbers and central limiting theorem is also about the sample average. If we look into the proof of these results, we can find that these results rely on the additive formality of the sample mean. So we have the impression that the concentration principle works for the average, but does it cover other statistics? In fact, we have many nonlinear estimators in statistics and machine learning. Can we expect that a general statistic $f(X_1, \ldots, X_n)$ concentrates to its expectation? The answer is positive. Sample mean is not special. We have the general concentration principle as follows.

General Concentration Principle: A random variable $f(X_1, \ldots, X_n)$ concentrates to its mean $\mathbb{E}f(X_1, \ldots, X_n)$ if

1. $X_1, \ldots, X_n$ are independent,
2. The function $f(x_1, \ldots, x_n)$ is not too "sensitive" to any of its coordinate x_i.

To make the above statement precise, we need to specify what do we mean by "sensitive"? One possible way is to measure it by coordinate derivatives. For example, for the sample mean, $f(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. So the sample mean follows the general concentration principle because the derivative $\frac{\partial f}{\partial x_i} = \frac{1}{n}$, which goes to zero as n goes to infinity. Therefore, we can say the sample mean is not sensitive to any one of its observation. Can we use the derivatives in general to measure the sensitivity? The following theorem shows

that the answer is positive but we need to modify the definition of derivatives.

3.1 Theorem (McDiarmid Inequality). *Define the i-th discrete derivative of $f(x_1, \ldots, x_n)$*

$$D_i f = \sup_{x_1,\ldots,x_n,x_i^{'}} \Big| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i^{'}, x_{i+1}, \ldots, x_n) \Big|.$$

For $X_1, \ldots, X_n$ independent, we have for any $t > 0$,

$$\mathbb{P}\big(f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n) > t\big) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (D_i f)^2}}.$$

Proof. The main idea is to decompose the $f - \mathbb{E}f$ into additive form via the martingale method. Define the martingale difference

$$\Delta_k = \mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_k] - \mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots X_{k-1}].$$

We then have the decomposition $f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n) = \sum_{k=1}^n \Delta_k$. We call Δ_k as the martingale difference because $\mathbb{E}[\Delta_k|X_1, \ldots, X_{k-1}] = 0$ by the tower property of conditional expectation.

Next, we will control the upper and lower bound of Δ_k. Define

$$U_k = \mathbb{E}\Big[\sup_z f(X_1, \ldots, X_{k-1}, z, X_{k+1}, \ldots X_n) - f(X_1, \ldots, X_n)|X_1, \ldots, X_{k-1}\Big],$$
$$L_k = \mathbb{E}\Big[\inf_z f(X_1, \ldots, X_{k-1}, z, X_{k+1}, \ldots X_n) - f(X_1, \ldots, X_n)|X_1, \ldots, X_{k-1}\Big]$$

and we have $L_k \leq \Delta_k \leq U_k$ almost surely. Notice that this is the only place we use the the independence of $X_1, \ldots, X_n$ throughout the proof.

As $|U_k - L_k| \leq D_k f$, using Theorem 2.9 of Lecture 2, we have

$$\mathbb{E}[e^{\lambda \Delta_k}|X_1, \ldots, X_{k-1}] \leq e^{\lambda^2 (D_k f)^2/8}.$$

Notice that although we did not use $\mathbb{E}[\Delta_k|X_1, \ldots, X_{k-1}] = 0$ explicitly. The above inequality is true only when Δ_k is a martingale difference.

We then compute

$$\mathbb{E}\big[e^{\lambda \sum_{k=1}^n \Delta_k}\big] = \mathbb{E}\Big[e^{\lambda \sum_{k=1}^{n-1} \Delta_k} \mathbb{E}\big[e^{\lambda \Delta_n}|X_1, \ldots, X_{n-1}\big]\Big] \leq \mathbb{E}\Big[e^{\lambda \sum_{k=1}^{n-1} \Delta_k} e^{\lambda^2 (\Delta_n f)^2/8}\Big].$$

Repeating the above step inductively to $n-1, \ldots, 1$, and we have

$$\mathbb{E}[e^{\lambda(f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n))}] = \mathbb{E}\big[e^{\lambda \sum_{k=1}^n \Delta_k}\big] \leq \prod_{k=1}^n e^{\lambda^2 (D_k f)^2/8} = e^{\frac{\lambda^2}{8} \sum_{k=1}^n (D_k f)^2}$$

So we complete the proof with the Chernoff bound of the moment-generating function above. □

Now, let's give a concrete example where McDiarmid inequality would be used.

3.2 Example (Uniform rate of kernel density estimator). Let $X_1, \ldots, X_n$ be i.i.d. random variables from density $p(x)$ with support on $[0,1]$. Our goal here is to estimate $p(x)$ by the kernel density estimator:

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where h is called the bandwidth and $K(\cdot)$ is the kernel function. Typically, we will choose the kernel function with bounded support satisfying

$$\int K(t)dt = 1, \int tK(t)dt = 0, \sup_z |K(z)| \le B \text{ and } |K(x) - K(y)| \le L|x - y|,$$

for some upper bound $B > 0$ and Lipschitz constant L. An example is the pyramid kernel $K(x) = \max(1 - |x|, 0)$. We aim to give the uniform statistical rate of $\widehat{p}$ via bounding

$$\|\widehat{p} - p\|_\infty := \sup_{x\in[0,1]} |\widehat{p}(x) - p(x)|.$$

We consider the variance-bias decomposition of $\|\widehat{p} - p\|_\infty$:

$$\text{Variance term : } \|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty \text{ and Bias term : } \|\mathbb{E}[\widehat{p}] - p\|_\infty.$$

We first consider the variance term. Using the boundedness of the kernel function, we have

$$D_i(\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty) \le \frac{\|K\|_\infty}{nh} \le \frac{B}{nh}.$$

By McDiarmid inequality, we have for any $t > 0$,

$$\mathbb{P}\Big(\big|\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty - \mathbb{E}\big[\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty\big]\big| > t\Big) \le 2e^{-2nh^2t^2/B^2}.$$

Therefore, with probability at least $1 - \delta$, we have

$$\big|\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty - \mathbb{E}\big[\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty\big]\big| \le \sqrt{B^2\frac{\log(2/\delta)}{2nh^2}}.$$

Now we bound the bias. The expectation of the kernel density estimator is

$$\begin{aligned}
\mathbb{E}\widehat{p}(x) &= \int \frac{1}{h}K\left(\frac{t-x}{h}\right)p(t)dt \\
&= \int K(u)p(x+uh)du, \quad \text{(by change of variable } u = (t-x)/h) \\
&= \int K(u)(p(x) + p'(x)uh + p''(x)(uh)^2/2 + O(h^3))du,
\end{aligned}$$

where in the last equality we expand $p(x + uh)$ at x. So we can bound the bias as

$$\|\mathbb{E}[\widehat{p}] - p\|_\infty \le \frac{\|p''\|_\infty \int u^2K(u)du}{6}h^2 + o(h^3) = O(h^2).$$

Let us summarize what we have so far:

$$\|\widehat{p} - p\|_\infty \le \underbrace{\Big| \|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty - \mathbb{E}\big[\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty\big] \Big|}_{\text{McDiarmid inequality: } O_P(1/\sqrt{nh^2})} + \underbrace{\mathbb{E}\big[\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty\big]}_{\text{Maximal: ??}} + \underbrace{\|\mathbb{E}[\widehat{p}] - p\|_\infty}_{\text{Bias: } O(h^2)}.$$

We have bounded the first and third terms above. In the next lecture, we will show that the second term has the rate $O(\sqrt{\log n/(nh^2)})$. So we have

$$\|\widehat{p} - p\|_\infty = O_p\bigg(\sqrt{\frac{\log n}{nh^2}} + h^2\bigg).$$

Choosing the bandwidth $h \sim n^{-1/6}$, we have $\|\widehat{p} - p\|_\infty = O_P(\sqrt{\log n} \cdot n^{-1/3})$. □

3.2 Sub-Exponential Random Variables

In statistics, we usually need to estimate the variance or covariance matrix of a random variable X. Therefore, we need to consider X^2. However, even if X is sub-Gaussian, X^2 does not necessarily have a sub-Gaussian tail. The follow example show that even the square of Gaussian is not sub-Gaussian.

3.3 Example (Gaussian). Let $X \sim N(0,1)$. The moment-generating function of X^2 is

$$\mathbb{E}e^{\lambda(X^2-1)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{-\frac{z^2}{2}} dz = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}.$$

We know that X^2 is not sub-Gaussian as its moment-generating function goes to ∞ when $\lambda \ge \frac{1}{2}$. However, if we take a closer look of the moment-generating function, we have

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2}, \text{ as } |\lambda| < 1/2.$$

So if λ is not too large, the moment-generating function of X^2 is still bounded by something like sub-Gaussian. This motivates us for the definition of sub-exponential. □

3.4 Definition (Sub-Exponential). *We say a random variable X is sub-exponential with a parameter α if*

$$\mathbb{E}\big[e^{\lambda(X - \mathbb{E}X)}\big] \le e^{\frac{\lambda^2\alpha^2}{2}} \text{ for all } |\lambda| \le \frac{1}{\alpha}.$$

The key difference between sub-Gaussian and sub-exponential is that for sub-Gaussian, the moment generating function has the Chernoff bound for all $\lambda \in \mathbb{R}$, while for sub-exponential, it holds only for a specific range of λ. By the previous example, if $X \sim N(0,1)$, X^2 is sub-exponential with parameter 2.

Now that we have a shorter range, what will happen to the tail probability of sub-exponential?

3.5 Theorem (Sub-Exponentail tail probability). *If X is sub-exponential with a parameter α, then we have*

$$\mathbb{P}(X - \mathbb{E}X > t) \leq \begin{cases} e^{-\frac{t^2}{2\alpha^2}}, & \text{for } 0 \leq t < \alpha, \\ e^{-\frac{t}{2\alpha}}, & \text{for } t \geq \alpha. \end{cases}$$

Proof. By Chernoff bound, for all $\lambda \in [0, \frac{1}{\alpha}]$, we have

$$\mathbb{P}(X - \mathbb{E}[X] > t) \leq \frac{\mathbb{E}e^{\lambda(X - \mathbb{E}[X])}}{e^{\lambda t}} \leq e^{-\lambda t + \frac{\lambda^2\alpha^2}{2}}.$$

We still want to choose the optimal λ minimizing the right hand side above. If $\frac{t}{\alpha^2} \leq \frac{1}{\alpha}$, the optimal $\lambda = t/\alpha^2$ and we get the same result as sub-Gaussian: $e^{-\lambda t + \lambda^2\alpha^2/2} = e^{-t^2/(2\alpha)}$. If $\frac{t}{\alpha^2} > \frac{1}{\alpha}$, the optimal $\lambda = 1/\alpha$ and we have $e^{-\lambda t + \lambda^2\alpha^2/2} = e^{-t/\alpha + 1/2} \leq e^{-t/(2\alpha)}$ as $1/2 < t/\alpha$. □

The tail probability sub-exponential has two types of range: it behaves like the sub-Gaussian when $0 \leq t < \alpha$ and when $t \geq \alpha$, the tail probability decays like e^{-t}.

Now we can study the sample average of the sub-exponential.

3.6 Theorem. *Let $X_1, \ldots, X_n$ be independent, $\mathbb{E}[X_i] = \mu$ and X_i is is sub-exponential with a parameter α for all $1 \leq i \leq n$, we have*

$$\mathbb{P}(\bar{X}_n - \mu > t) \leq \begin{cases} e^{-\frac{nt^2}{2\alpha^2}} & \text{for } 0 \leq t \leq \alpha, \\ e^{-\frac{nt}{2\alpha}} & \text{for } t \geq \alpha. \end{cases}$$

Proof. By Chernoff bound, for all $\lambda \in [0, \frac{1}{\alpha}]$, we have

$$\mathbb{P}(\bar{X}_n - \mu > t) \leq \frac{\mathbb{E}e^{\lambda(\bar{X}_n - \mu)}}{e^{\lambda t}} = \prod_{i=1}^{n} \mathbb{E}e^{\lambda(X_i - \mu)} e^{-\lambda n t} \leq e^{n(-\lambda t + \lambda^2\alpha^2/2)}.$$

We can find that the right hand side above is almost the same as the corresponding part of a single sub-exponential. We only have an additional n. So following the exactly same analysis in the proof of Theorem 3.5, we can prove the result for sample mean. □

Lecture 4

Bernstein and Maximal Inequalities

4.1 Bernstein Inequality

In the previous lecture, we showed that the sample mean of independent sub-exponential random variables $X_1, \ldots, X_n$ with the parameter α has the tail probability

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}X| > t) \leq= 2e^{-\frac{n}{2}\left(\frac{t^2}{\alpha^2} \wedge \frac{t}{\alpha}\right)},$$

where $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. Therefore, with probability at least $1 - \delta$,

$$|\bar{X}_n - \mathbb{E}X| \leq \sqrt{\frac{\alpha^2}{n} \log\left(\frac{2}{\delta}\right)} \vee \left(\frac{\alpha}{n} \log\left(\frac{2}{\delta}\right)\right).$$

We can see that the two types of sup-exponential tail probability give us two types of rate: $O(\alpha/\sqrt{n})$ and $O(\alpha/n)$. Although the second term is dominated by the first term, it implies the possibility of giving two types of rates in the concentration inequality. We are going to show a stronger concentration inequality of such type.

4.1 Definition (Bernstein Condition). *We say a random variable X with $\mathbb{E}X = 0$ and $Var(X) = \sigma^2$ satisfies the Bernstein condition with parameter b if*

$$\mathbb{E}|X|^k \leq \frac{\sigma^2}{2} k!\, b^{k-2}, \quad \textit{for all } k \geq 3.$$

The Bernstein condition specified $k \geq 3$ because the $k = 1$ and $k = 2$ cases are vacuously true based on the existing assumptions about the mean and variance of X. It is straightforward to check that the bounded random variable satisfies the Bernstein condition.

4.2 Example (Bounded random variable). If $|X| \le B$ and $\mathbb{E}X = 0$, we can check that X satisfies the Bernstein condition with parameter $B/3$ as follows:

$$\mathbb{E}|X|^k \le \mathbb{E}|X|^2 \cdot B^{k-2} \le \sigma^2 B^{k-2} = \sigma^2 \cdot 3^{k-2} \left(\frac{B}{3}\right)^{k-2} \le \sigma^2 \cdot \frac{k!}{2} \left(\frac{B}{3}\right)^{k-2}.$$

The last inequality is because $3^{k-2} \le \frac{k!}{2}$ for all $k \ge 3$. □

In the following theorem, we will show that if X satisfies the Bernstein condition, then X is sub-exponential. We also have a better concentration inequality as follows.

4.3 Theorem (Bernstein Inequality). *Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}X_i = 0$, $\mathrm{Var}(X_i) = \sigma^2$ and X_i satisfies the Bernstein condition with parameter b for all $i = 1, \dots n$. We have*

$$\mathbb{P}(\bar{X}_n > t) \le \exp\left(-\frac{nt^2}{2(\sigma^2 + bt)}\right), \quad \textit{for all } t > 0.$$

Given a bounded random variable $|X| \le B$, Example 4.2 implies that X satisfies the Berstein condition with parameter $B/3$. Applying the Bernstein inequality, we have with probability at least $1 - \delta$,

$$\bar{X}_n - \mathbb{E}X \le \frac{\sigma}{\sqrt{n}} \sqrt{\log(1/\delta)} + \frac{B}{3n} \log(1/\delta).$$

So Bernstein inequality gives us two types of rates like the rate of the sample mean of sub-exponential random variables. However, notice that the constant assigned to the $O(1/\sqrt{n})$ term is the variance σ^2. In contrast, if we apply the Hoeffding inequality, we have at least $1 - \delta$,

$$\bar{X}_n - \mathbb{E}X \le \frac{B}{\sqrt{n}} \sqrt{\log(1/\delta)}.$$

Therefore, Bernstein inequality will give us a better rate than the Hoeffding inequality when $\frac{\sigma}{\sqrt{n}} + \frac{B}{n} = o\left(\frac{B}{\sqrt{n}}\right)$, or equivalently, when $\sigma = o(B)$: the standard deviation is much smaller than the upper bound.

Proof. We prove the Bernstein inequality by two steps. First, we show that if X satisfies the Bernstein condition, then X is sub-exponential. Second, we use the Chernoff bound to control the tail probability.

Step 1: We aim to prove that if X satisfies the Bernstein condition with parameter b, then X is sub-exponential with parameter $2(\sigma^2 \vee b)$.

We expand the moment generating function of X into Taylor series:

$$\begin{aligned}\mathbb{E}e^{\lambda X} &= \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}X^k = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}X^k \\ &\le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{|\lambda|^k}{k!}\frac{\sigma^2}{2}k!b^{k-2} \qquad \text{(by the Bernstein condition)} \\ &= 1 + \frac{\lambda^2\sigma^2}{2}\sum_{k=0}^{\infty}|\lambda|^k b^k.\end{aligned}$$

If $|\lambda| < 1/b$, the infinite summation of the power series above converges and we have

$$\mathbb{E}e^{\lambda X} \le 1 + \frac{\lambda^2\sigma^2}{2}\sum_{k=0}^{\infty}|\lambda|^k b^k = 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-|\lambda|b} \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-|\lambda|b)}\right),$$

where we use $1 + x \le e^x$ in the last inequality. Therefore, if $|\lambda| \le 1/(2(\sigma^2 \vee b))$, we further have

$$\mathbb{E}e^{\lambda X} \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-|\lambda|b)}\right) \le \exp(\lambda^2\sigma^2) \le \exp\left(\frac{\lambda^2}{2}\left(2(\sigma^2 \vee b)\right)\right).$$

Therefore, by definition, X is sub-exponential with parameter $2(\sigma^2 \vee b)$.

Step 2: We apply the Chernoff bound on $\bar{X}_n$. By the analysis in Step 1, for all $|\lambda| < 1/b$, we have

$$\mathbb{P}(\bar{X}_n > t) \le e^{-\lambda n t}\prod_{i=1}^{n} e^{\lambda X_i} \le \exp\left(\frac{n\lambda^2\sigma^2}{2(1-|\lambda|b)} - n\lambda t\right).$$

We then complete the proof by choosing $\lambda = \frac{t}{\sigma^2 + bt}$ as such $\lambda \in (0, 1/b)$. □

4.2 Maximal Inequality

In statistics, we sometimes need to study uniform performance of multiple estimators, especially in the high dimensional scenario. In nonparametric statistics, we also need to study the uniform rate of some function estimator. Therefore, we need to study how large is the maximum of a set of random variables.

We start with the maximum of finite number of random variables. Given $X_1, \ldots, X_d$, we can control the maximum by $\mathbb{E}[\max_{1\le j\le d} X_j] \le \sum_{j=1}^{d}\mathbb{E}|X_j| = O(d)$. However, if X_i's are sub-Gaussian, we can achieve a much better upper bound.

4.4 Theorem (Finite Maximal Inequality). *Let $X_1, \ldots, X_d$ be (not necessarily independent) sub-Gaussian random varibles with variance-proxy σ^2 and $\mathbb{E}X_i = 0$ for all $i = 1, \ldots, d$. We have*

$$\mathbb{E}\left[\max_{1\le j\le d} X_j\right] \le \sigma\sqrt{2\log d},$$

and with probability at least $1-\delta$,

$$\max_{1\le j\le d} X_j \le \sigma\sqrt{2\log(d/\delta)}.$$

Proof. We will apply the Jensen's inequality that given a concave function φ, then $\varphi(\mathbb{E}X) \ge \mathbb{E}[\varphi(X)]$. See Figure 4.1 for an illustration. So we rewrite the expectation as follows.

$$\begin{aligned}
\mathbb{E}\Big[\max_{1\le j\le d} X_j\Big] &= \mathbb{E}\Big[\frac{1}{\lambda}\log\exp\big(\max_{1\le j\le d}\lambda X_j\big)\Big] \\
&\le \frac{1}{\lambda}\log\mathbb{E}\Big[\exp\big(\max_{1\le j\le d}\lambda X_j\big)\Big] \quad \text{(by applying Jensen's inequality to } \log x\text{)} \\
&\le \frac{1}{\lambda}\log\sum_{j=1}^{d}\mathbb{E}e^{\lambda X_j} \le \frac{1}{\lambda}\log\sum_{j=1}^{d} e^{\frac{\lambda^2\sigma^2}{2}} = \frac{\log d}{\lambda} + \frac{\lambda\sigma^2}{2}.
\end{aligned}$$

We obtain the upper bound of expectation by choosing the optimal $\lambda = \sqrt{\frac{2}{\sigma^2}\log d}$.

Now we turn to control the tail probability:

$$\mathbb{P}\Big(\max_{1\le j\le d} X_j > t\Big) = \mathbb{P}\Big(\bigcup_{j=1}^{d}\{X_j > t\}\Big) \le \sum_{j=1}^{d}\mathbb{P}(X_j > t) \le de^{-\frac{t^2}{2\sigma^2}},$$

where the first inequality above is due to the union bound. □

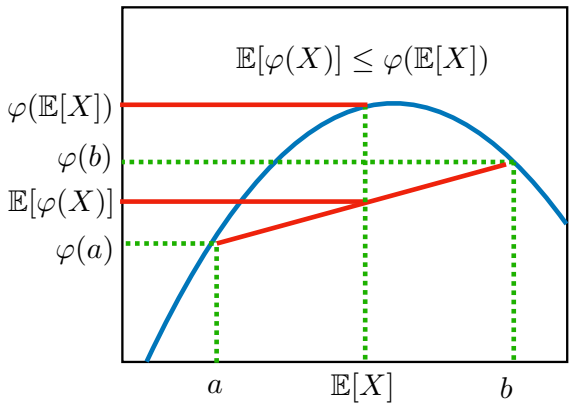


Figure 4.1. *Simple illustration of Jensen's inequality. For concave* φ *(blue line),* $\varphi(\mathbb{E}X)$ *will be greater than or equal to* $\mathbb{E}[\varphi(X)]$.

Now we switch to study the maximal inequality for the suprema of infinite number of random variables. We first consider the ℓ_2-norm of a d-dimensional random vector $X = (X_1,\dots,X_d)^T$ defined as $\|X\| = \sqrt{|X_1|^2+\cdots+|X_d|^2}$. The ℓ_2-norm $\|X\|$ is a suprema due to its variational form

$$\|x\| = \sup_{\|y\|\le 1}\langle y, x\rangle,$$

which can be proved by using the Cauchy-Schwartz inequality $\langle y, x\rangle \le \|x\|\|y\|$ and the equality is obtained by choosing $y = x/\|x\|$.

Denote the ℓ_2-ball as $\mathcal{B}_2^d = \{\theta \in \mathbb{R}^d | \|\theta\| \le 1\}$ and we can write $\|X\| = \sup_{u\in\mathcal{B}_2^d}\langle u, X\rangle$. Now we have a suprema over an ℓ_2-ball which is an infinite set. However, Theorem 4.4 only gives us the maximal inequality for finite random variables. Our idea to bridge the infinite set to a finite maximal inequality is the so-called **discretization trick**: we find finite representive points in the infinte set. This motivates us the following definition of ϵ-net.

4.5 Definition (ϵ-Net). *Given a metrix space $\mathcal{X}$ with the distance function $d(\cdot,\cdot)$, we say $\mathcal{N}_\epsilon \subseteq \mathcal{X}$ is an ϵ-net of $\mathcal{X}$, if for any $x \in \mathcal{X}$, there exists a $y \in \mathcal{N}_\epsilon$ such that $d(x, y) \le \epsilon$.*

As illustrated in Figure 4.2, an ϵ-net collects the centers of the balls with radius ϵ and cover the set $\mathcal{X}$. Therefore, we also define the minimum cardinality of all possible ϵ-net as the **covering number** of $\mathcal{X}$, denoted as $N(\mathcal{X}, d, \epsilon)$. Applying the discritization trick by considering the ϵ-net of the ℓ_2-ball as the representative points, we have the following result on the ℓ_2-norm of a random vector.

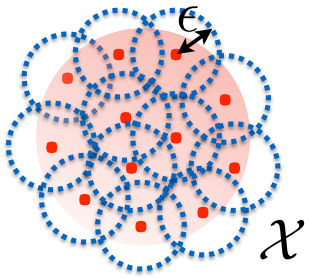


Figure 4.2. *Illustration of the ϵ-net $\mathcal{N}_\epsilon$ (red dots) of $\mathcal{X}$ (pink disc).*

4.6 Theorem (Maximal inequality for ℓ_2-norm). *Given a random vector $X \in \mathbb{R}^d$, if for any $u \in \mathbb{R}^d$, we have $\langle u, X\rangle$ is sub-Gaussian with variance-proxy $\sigma^2\|u\|^2$, then we have*

$$\mathbb{E}\|X\| \le 4\sigma\sqrt{d}$$

and with probability at least $1 - \delta$,

$$\|X\| \le 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

The condition in the theorem that $\langle u, X\rangle$ is sub-Gaussian with variance-proxy $\sigma^2\|u\|^2$ for any $u \in \mathbb{R}^d$ can be interpreted as X is a sub-Gaussian vector. For example, if $X_1, \ldots, X_d$ are independent sub-Gaussian scalars with the variance-proxy σ^2, by Hoeffding inequality, we have $\langle u, X\rangle = \sum_{j=1}^d u_j X_j$ is sub-Gaussian with variance-proxy $\sigma^2\|u\|^2$. Let us see another example for a dependent random vector. If $X \sim N(0, \Sigma)$, then $u^T X \sim N(0, u^T\Sigma u)$ and thus $u^T X$ is sub-Gaussian with the variance-proxy $u^T\Sigma u \le \lambda_{\max}(\Sigma)\|u\|^2$.

Proof. We apply the discretization trick by considering the $1/2$-net of $\mathcal{B}_2^d$, denoted as $\mathcal{N}_{1/2}$. Then for any $u \in \mathcal{B}_2^d$, there exists a $v \in \mathcal{N}_{1/2}$ such that $\|u - v\| \le 1/2$. So we can reformulate the ℓ_2-norm as:

$$\begin{aligned}\|X\| = \max_{u\in\mathcal{B}_2}\langle u, X\rangle &= \max_{u\in\mathcal{B}_2}\left(\langle v, X\rangle + \langle u - v, X\rangle\right)\\ &\le \max_{v\in\mathcal{N}_{1/2}}\langle v, X\rangle + \max_{\|u-v\|\le 1/2}\langle u - v, X\rangle\\ &\le \max_{v\in\mathcal{N}_{1/2}}\langle v, X\rangle + \frac{1}{2}\max_{u\in\mathcal{B}_2}\langle u, X\rangle.\end{aligned}$$

Rearranging the last inequality, we obtain the key inequality in the discretization trick:

$$\max_{u\in\mathcal{B}_2}\langle u, X\rangle \le 2\max_{v\in\mathcal{N}_{1/2}}\langle v, X\rangle,$$

which reduces a suprema over an infinite set to the maximum over a finite set. It follows from Theorem 4.4 that

$$\mathbb{E}\|X\| = \mathbb{E}\max_{u\in\mathcal{B}_2}\langle u, X\rangle \le 2\mathbb{E}\max_{v\in\mathcal{N}_{1/2}}\langle v, X\rangle \le 2\sigma\sqrt{2\log\left(|\mathcal{N}_{1/2}|\right)},$$

and

$$\mathbb{P}(\|X\| > t) = \mathbb{P}\Big(\max_{u\in\mathcal{B}_2}\langle u, X\rangle > t\Big) \le \mathbb{P}\Big(\max_{v\in\mathcal{N}_{1/2}}\langle v, X\rangle > t/2\Big) \le |\mathcal{N}_{1/2}|e^{-\frac{t^2}{8\sigma^2}}.$$

It remains to bound the covering number $|\mathcal{N}_{1/2}|$ by the following lemma.

4.7 Lemma (Covering number of ℓ_2-ball). *Given any $\epsilon \in (0, 1)$, let $\mathcal{N}_\epsilon$ be the ϵ-net of $\mathcal{B}_2^d$. We have $|\mathcal{N}_\epsilon| \le \left(1 + \frac{2}{\epsilon}\right)^d$.*

Implementing the above lemma, we finish the proof of the theorem as

$$\mathbb{E}\|X\| \le 2\sigma\sqrt{2\log\left(|\mathcal{N}_{1/2}|\right)} \le 2\sigma\sqrt{2d\log 5} \le 4\sigma\sqrt{d},$$

and the tail probability has $\mathbb{P}(\|X\| > t) \le |\mathcal{N}_{1/2}|e^{-\frac{t^2}{8\sigma^2}} \le 5^d e^{-\frac{t^2}{8\sigma^2}}$. Letting $\delta = 5^d e^{-\frac{t^2}{8\sigma^2}}$, we solve $t \le 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$. Notice here we relax the constants to make them simplified. □

We now go back and prove Lemma 4.7 below.

Proof. As the ϵ-net implies a covering (see Figure 4.2), we have

$$\bigcup_{x\in\mathcal{N}_\epsilon}\{x + \epsilon\mathcal{B}_2^d\} \supseteq \mathcal{B}_2^d.$$

So the volume of a ℓ_2-ball is bounded by the volumes of all small balls, i.e.,

$$\mathrm{Vol}(\mathcal{B}_2^d) \le \mathrm{Vol}\Big(\bigcup_{x\in\mathcal{N}_\epsilon} \{x + \epsilon\mathcal{B}_2^d\} \Big) \le |\mathcal{N}_\epsilon| \mathrm{Vol}(\epsilon\mathcal{B}_2^d) = |\mathcal{N}_\epsilon| \epsilon^d \mathrm{Vol}(\mathcal{B}_2^d).$$

So this simple observation using the fact of covering gives us the lower bound of the covering number $|\mathcal{N}_\epsilon| \ge (1/\epsilon)^d$, which increases exponentially with the dimension. However, we need a finer analysis to show the upper bound, using a similar **volume argument**.

Let us construct the ϵ-net of $\mathcal{B}_2^d$ by a concrete procedure below:

1. Set $x_1 = 0$;
2. For $i \ge 2$, set $x_i = \mathcal{B}_2^d \setminus \bigcup_{j=1}^{i-1} \{x_j + \epsilon\mathcal{B}_2^d\}$. This is to say, pick a point in $\mathcal{B}_2$ not included in the union of the ϵ-balls around any of the existing x_i's.
3. Output: $\mathcal{N}_\epsilon = \{x_1, \ldots, x_{i-1}\}$ if there is no such x_i exists in Step 2.

This algorithm yields three important facts (see Figure 4.3 for an illustration):

1. $\mathcal{N}_\epsilon$ in an ϵ-net, as we will keep adding points until we have no more space in step 2.
2. The balls $\{x_j + \frac{\epsilon}{2}\mathcal{B}_2^d\}$ for all $x_j \in \mathcal{N}_\epsilon$ are disjoint, as based on this algorithm, $x_j, x_k \in \mathcal{N}_\epsilon, \|x_j - x_k\| > \epsilon$.
3. $\left(1 + \frac{\epsilon}{2}\right)\mathcal{B}_2^d \supseteq \bigcup_{j=1}^{|\mathcal{N}_\epsilon|} \left\{x_j + \frac{\epsilon}{2}\mathcal{B}_2^d\right\}$, as $x_j \in \mathcal{B}_2^d$ and thus the ball $\{x_j + \frac{\epsilon}{2}\mathcal{B}_2^d\}$ are contained in $\left(1 + \frac{\epsilon}{2}\right)\mathcal{B}_2^d$.

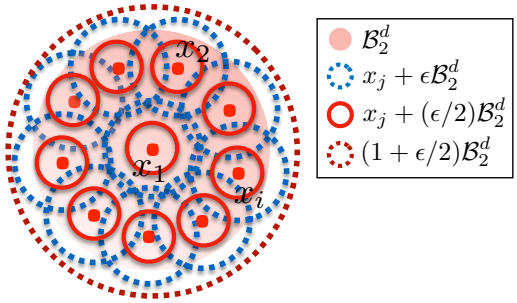


Figure 4.3. *Illustration of volume argument in the proof of Lemma 4.7.*

Using these facts, we use the volume argument that

$$\begin{aligned}\left(1+\frac{\epsilon}{2}\right)^d \mathrm{Vol}(\mathcal{B}_2^d) &= \mathrm{Vol}\left(\left(1+\frac{\epsilon}{2}\right)\mathcal{B}_2^d\right) \\ &\geq \mathrm{Vol}\left(\bigcup_{j=1}^{|\mathcal{N}_\epsilon|}\left\{x_j+\frac{\epsilon}{2}\mathcal{B}_2^d\right\}\right) \qquad \text{(by Fact 3.)} \\ &= \sum_{j=1}^{|\mathcal{N}_\epsilon|}\mathrm{Vol}\left(\frac{\epsilon}{2}\mathcal{B}_2^d\right) = |\mathcal{N}_\epsilon|\left(\frac{\epsilon}{2}\right)^d \mathrm{Vol}(\mathcal{B}_2^d). \qquad \text{(by Fact 2.)}\end{aligned}$$

Solving the inequality above and we have $|\mathcal{N}_\epsilon| \leq \left(1+\frac{2}{\epsilon}\right)^d$. □

We finish this lecture by considering another example of implementing the discretization trick to maximal inequality.

4.8 Example (Uniform rate of the kernel density estimator (continued)). We discuss the kernel density estimator in Example 3.2 in Lecture 3. Given i.i.d. observations $X_1, \ldots, X_n$ generated from the density $p(x)$ supported on $[0,1]$. Recall the kernel density estimator:

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right),$$

where the kernel $K(\cdot)$ is a bounded support function satisfying

$$\int K(t)dt = 1, \int tK(t)dt = 0, \sup_z |K(z)| \leq B \text{ and } |K(x)-K(y)| \leq L|x-y|,$$

for any x, y. We aim to bound the uniform rate $\|\widehat{p}-p\|_\infty := \sup_{x\in[0,1]}|\widehat{p}(x)-p(x)|$.

We can decompose $\|\widehat{p}-p\|_\infty$ into three terms:

$$\|\widehat{p}-p\|_\infty \leq \underbrace{\left|\|\widehat{p}-\mathbb{E}[\widehat{p}]\|_\infty - \mathbb{E}\big[\|\widehat{p}-\mathbb{E}[\widehat{p}]\|_\infty\big]\right|}_{\text{Term I}} + \underbrace{\mathbb{E}\big[\|\widehat{p}-\mathbb{E}[\widehat{p}]\|_\infty\big]}_{\text{Term II}} + \underbrace{\|\mathbb{E}[\widehat{p}]-p\|_\infty}_{\text{Term III}},$$

where in Example 3.2 of Lecture 3, we have shown that Term I has the rate $O_P(1/\sqrt{nh^2})$ by McDiarmid inequality and the bias Term III has the rate $O(h^2)$. It remains to study Term 2, which is an expectation of maximal:

$$\begin{aligned}\text{Term II} = \mathbb{E}\big[\|\widehat{p}-\mathbb{E}[\widehat{p}]\|_\infty\big] &= \mathbb{E}\Big[\sup_{x\in[0,1]}\big|\widehat{p}(x)-\mathbb{E}[\widehat{p}(x)]\big|\Big] \\ &= \mathbb{E}\left[\sup_{x\in[0,1]}\left|\frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right) - \mathbb{E}K\left(\frac{X_i-x}{h}\right)\right|\right].\end{aligned}$$

To simplify our notation, we denote

$$K_i(x) = K\left(\frac{X_i-x}{h}\right) - \mathbb{E}K\left(\frac{X_i-x}{h}\right), \text{ for all } i = 1, \ldots, n.$$

To control the maximal, we use the discretization trick and consider a $1/n$-net of $[0,1]$. We can simply use the grids of the interval $\mathcal{N}_{1/n} = \{x_j = j/n | j = 1, \dots, n\}$. Similar to the proof of Theorem 4.6, we have for any $x \in [0,1]$, there exists a $1 \le j \le n$ such that $|x - x_j| \le 1/n$ and thus

$$\begin{aligned}\mathbb{E}\|\widehat{p} - \mathbb{E}\widehat{p}\|_\infty &= \frac{1}{nh}\mathbb{E}\bigg[\sup_{x\in[0,1]}\bigg|\sum_{i=1}^n K_i(x)\bigg|\bigg] \\ &\le \frac{1}{nh}\mathbb{E}\bigg[\max_{1\le j\le n}\bigg|\sum_{i=1}^n K_i(x_j)\bigg| + \sup_{|x-x_j|\le 1/n}\sum_{i=1}^n |K_i(x) - K_i(x_j)|\bigg]. \quad (4.9)\end{aligned}$$

For the first term in (4.9), it is a finite maximal. As $|K_i(x_j)| \le 2B$ for all $1 \le i, j \le n$, so by Hoeffding inequality, $\sum_{i=1}^n K_i(x_j)$ is sub-Gaussian with variance-proxy nB^2. Applying the finite maximal inequality in Theorem 4.4, we have

$$\frac{1}{nh}\mathbb{E}\bigg[\max_{1\le j\le n}\bigg|\sum_{i=1}^n K_i(x_j)\bigg|\bigg] \le \sqrt{\frac{2B^2\log(2n)}{nh^2}},$$

where we reduce the maximal of absolute values to the maximal inequality without absolute values via the identity $|X| = X \vee (-X)$.

For the second term in (4.9), we apply the Lipschitz property that $|K(x) - K(y)| \le L|x - y|$ for any x, y and have

$$\frac{1}{nh}\mathbb{E}\bigg[\sup_{|x-x_j|\le 1/n}\sum_{i=1}^n |K_i(x) - K_i(x_j)|\bigg] \le \frac{1}{nh}\mathbb{E}\bigg[\sup_{|x-x_j|\le 1/n}\sum_{i=1}^n 2L\Big|\frac{x - x_j}{h}\Big|\bigg] \le \frac{2L}{nh^2}.$$

Summarizing the rates of the two terms in (4.9), we have

$$\mathbb{E}\|\widehat{p} - \mathbb{E}\widehat{p}\|_\infty \le \sqrt{\frac{2B^2\log(2n)}{nh^2}} + \frac{2L}{nh^2} = O\bigg(\sqrt{\frac{\log n}{nh^2}}\bigg).$$

Combining the analysis in Example 3.2 of Lecture 3, we have

$$\begin{aligned}\|\widehat{p} - p\|_\infty &\le \underbrace{\big|\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty - \mathbb{E}\big[\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty\big]\big|}_{\text{McDiarmid inequality}} + \underbrace{\mathbb{E}\big[\|\widehat{p} - \mathbb{E}[\widehat{p}]\|_\infty\big]}_{\text{Maximal inequality}} + \underbrace{\|\mathbb{E}[\widehat{p}] - p\|_\infty}_{\text{Bias}} \\ &= O_P\bigg(\frac{1}{\sqrt{nh^2}}\bigg) + O\bigg(\sqrt{\frac{\log n}{nh^2}}\bigg) + O(h^2) = O_P\bigg(\sqrt{\frac{\log n}{nh^2}} + h^2\bigg).\end{aligned}$$

Choosing the bandwidth $h \sim n^{-1/6}$, we have $\|\widehat{p} - p\|_\infty = O_P(\sqrt{\log n} \cdot n^{-1/3})$. $\square$

Lecture 5

Ordinary Least Squares

5.1 Linear Regression

Given the outcome Y_i and the covariates X_i for $i = 1, \ldots, n$, a regression model assumes

$$Y_i = f(X_i) + \varepsilon_i, \text{ for all } i = 1, \ldots, n,$$

where ε_i is the error/noise. We typically assume that the error terms satisfy $\mathbb{E}\varepsilon_i = 0$ and $\epsilon_1, \ldots, \epsilon_n$ are independent.

A special regression model is the **linear regression model**, where we assume $f(x) = x^\top \beta$ for $\beta \in \mathbb{R}^d$ and thus the regression model becomes

$$Y_i = X_i^\top \beta + \varepsilon_i, \text{ for all } i = 1, \ldots, n.$$

We also want to introduce some matrix notations. We define the **design matrix** $\mathbb{X} = (X_1^\top, \ldots, X_n^\top)^\top \in \mathbb{R}^{n \times d}$, the **response vector:** $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ and the **noise vector** $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. We can then write the linear model as

$$Y = \mathbb{X}\beta + \varepsilon.$$

We will also write the design matrix as the columns $\mathbb{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)$, where $\widetilde{X}_j$ is the jth column of $\mathbb{X}$.

We two typical settings for the linear regression:

- **Fixed design**: the features $X_1, \ldots, X_n$ are deterministic.
- **Random design**: the features $X_1, \ldots, X_n$ random, and we typically assume ε is independent to $\mathbb{X}$.

In this course, we focus on the fixed design setting. Without further specification, we will assume $\mathbb{X}$ is deterministic by default. If $\mathbb{X}$ were actually random, we could condition the design $\mathbb{X}$ and reduce the problem to a fixed design settings in most situations.

There are two major goals in the study of linear regression:

1. **Prediction**. The regression model should predict outcomes well. We can measure the prediction accuracy by the mean squared error of some estimator $\widehat{f}$ of the true function f^*:
$$\mathrm{MSE}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{f}(X_i) - f^*(X_i)\big)^2.$$
We say an estimator $\widehat{f}$ is **persistent** if $\mathrm{MSE}(\widehat{f}) = o_P(1)$.
In the linear regression scenario, we can write the mean squared error as
$$\mathrm{MSE}(\mathbb{X}\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big(X_i^\top\widehat{\beta} - X_i^\top\beta^*\big)^2 = \frac{1}{n}\|\mathbb{X}(\widehat{\beta}-\beta^*)\|^2 = (\widehat{\beta}-\beta^*)^\top\widehat{\Sigma}(\widehat{\beta}-\beta^*),$$
where $\widehat{\Sigma} = \mathbb{X}^\top\mathbb{X}/n = \frac{1}{n}\sum_{i=1}^{n}X_i^\top X_i$ is the sample covariance matrix.
2. **Parameter estimation**. Parameter estimates should be consistent. In the linear case, this means looking at the rate of $\|\widehat{\beta}-\beta^*\|$.

5.2 Ordinary Least Squares

We now start with the most popular estimator for the linear model, typically known as the **ordinary least squares** estimator:
$$\widehat{\beta}^{\mathrm{LS}} = \arg\min_{\beta}\sum_{i=1}^{n}(Y_i - X_i^\top\beta)^2 = \arg\min_{\beta}\|Y-\mathbb{X}\beta\|^2.$$

The following proposition gives the closed form formula for the ordinary lease square.

5.1 Proposition (Closed form of least squares). *The ordinary least squares solution can be explicitly specified by*
$$\widehat{\beta}^{\mathrm{LS}} = (\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top Y,$$
where $A^\dagger$ is the Moore–Penrose pseudo-inverse of A.

Proof. By definition, the critical points of the least square loss at $\widehat{\beta}^{\mathrm{LS}}$ has
$$0 = \frac{\partial}{\partial\beta}\|Y-\mathbb{X}\beta\|^2\Big|_{\beta=\widehat{\beta}^{\mathrm{LS}}} = 2\mathbb{X}^\top(Y-\mathbb{X}\widehat{\beta}^{\mathrm{LS}}).$$
Solve the equation above. We have $\mathbb{X}^\top\mathbb{X}\widehat{\beta}^{\mathrm{LS}} = \mathbb{X}^\top Y$ and thus $\widehat{\beta}^{\mathrm{LS}} = (\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top Y$. □

Although the closed form of ordinary least squares involves a mysterious pseduo-inverse inside, it actually has a clear geometric interpretation. From Figure 12.2, the ordinary least squares essentially aim to find a point in the space spanned by $\widetilde{X}_1,\ldots,\widetilde{X}_d$, denoted by $\mathcal{C}(\mathbb{X}) := \{\mathbb{X}\beta|\beta\in\mathbb{R}^d\}$, such that it is closest to Y. The formulation of β^{LS} tells us such

point is $\mathbb{X}\widehat{\beta}^{\mathrm{LS}}$. Using the geometric language, we say $\mathbb{X}\widehat{\beta}^{\mathrm{LS}}$ is the projection of Y onto the plane $\mathcal{C}(\mathbb{X})$, i.e.,

$$\text{Projection } Y \text{ onto } \mathcal{C}(\mathbb{X}) = \mathbb{X}\widehat{\beta}^{\mathrm{LS}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top Y.$$

Therefore, we define the **projection matrix** $P_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top$. It maps any vector to its projection onto the linear space spanned by the columns of $\mathbb{X}$ (see Figure 12.2). For example, let us consider the projection on a vector $X \in \mathbb{R}^d$. The projection matrix $P_X = XX^\top/\|X\|^2$. So projecting a vector Y onto the vector X can be written as

$$P_X Y = \frac{XX^\top}{\|X\|^2}Y = \frac{\langle X, Y\rangle}{\langle X, X\rangle}X = \left\langle Y, \frac{X}{\|X\|}\right\rangle \frac{X}{\|X\|},$$

which is consistent with the geometric concept of a projection.

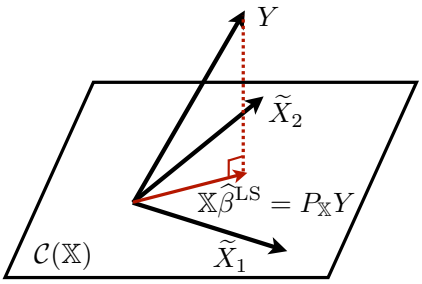


Figure 5.1. *Ordinary least squares is the projection of Y onto $\mathcal{C}(\mathbb{X})$.*

Not only does the entire vector $\widehat{\beta}^{\mathrm{LS}}$ has a geometric interpretation. In fact, the following proposition gives the geometric meaning of each entry $\widehat{\beta}^{\mathrm{LS}}$.

5.2 Proposition (Geometry of $\widehat{\beta}^{\mathrm{LS}}_j$). *Let $\mathbb{X}_{-j} = (\widetilde{X}_i, \ldots, \widetilde{X}_{j-1}, \widetilde{X}_{j+1}, \ldots, \widetilde{X}_d) \in \mathbb{R}^{n\times(d-1)}$ denote the design matrix with the jth column deleted. Define*

$$X_j^\perp = X_j - P_{\mathbb{X}_{-j}}X_j,$$

where $X_j^\perp$ is the projection of X_j on the linear space orthogonal to $\mathcal{C}(\mathbb{X}_{-j})$. See the vector in Figure 5.2. We have

$$\widehat{\beta}^{\mathrm{LS}}_j = \frac{\left\langle Y, X_j^\perp\right\rangle}{\left\langle X_j^\perp, X_j^\perp\right\rangle}.$$

Namely, $\widehat{\beta}^{\mathrm{LS}}_j$ is the coefficient assigned to $X_j^\perp$ when projecting Y onto the vector $X_j^\perp$.

Proof. The proof is straightforward in Figure 5.2. Applying the concept of projection and Euclidean geometry, the key observations in Figure 5.2 is that the projection of Y onto $X_j^\perp$

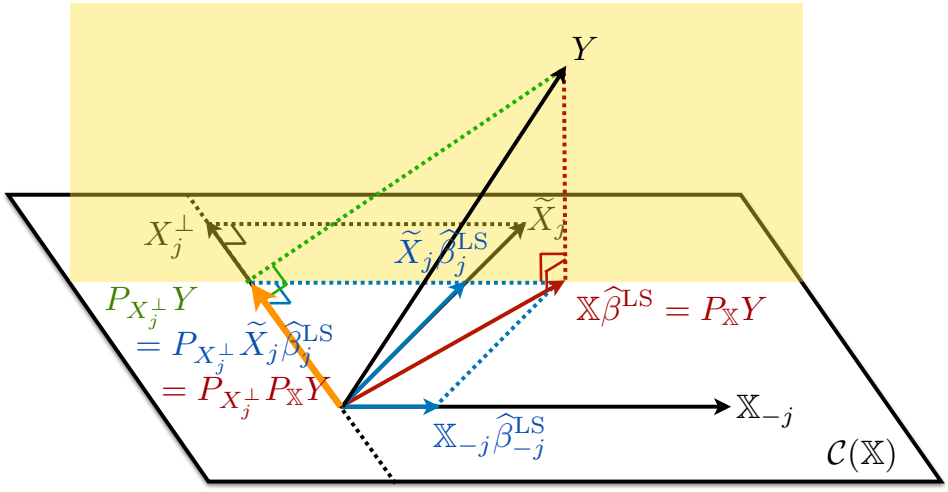


Figure 5.2. *Illustration of Proposition 5.2 and its proof.*

(green dashed line) equals to the sequential projections of Y onto $\mathcal{C}(\mathbb{X})$ (red dashed line) and then onto $X_j^\perp$ (blue dashed line).

Therefore, see in Figure 5.2 that the projection of Y (black vector), $P_{\mathbb{X}}Y$(red vector), $X_j\widehat{\beta}_j^{\mathrm{LS}}$ (blue vector) onto $X_j^\perp$ are the same orange vector, i.e., the following three vectors are identical:

$$P_{X_j^\perp}Y = P_{X_j^\perp}P_{\mathbb{X}}Y = P_{X_j^\perp}X_j\widehat{\beta}_j^{\mathrm{LS}}.$$

As $P_{X_j^\perp}X_j = X_j^\perp$ by the definition of projection, we have $P_{X_j^\perp}Y = \widehat{\beta}_j^{\mathrm{LS}}X_j^\perp$, i.e.,

$$\frac{\left\langle Y, X_j^\perp \right\rangle}{\left\langle X_j^\perp, X_j^\perp \right\rangle}X_j^\perp = \widehat{\beta}_j^{\mathrm{LS}}X_j^\perp,$$

which completes our proof. □

The last theorem of this lecture aims to show the statistical rate of mean squared error for the ordinary least squares.

5.3 Theorem (Mean Squared Error of Least Squares). *If the independent noises $\varepsilon_1, \ldots, \varepsilon_n$ has $\mathbb{E}\varepsilon_i = 0$ and ε_i is sub-Gaussian with variance-proxy σ^2 for all $i = 1, \ldots, n$ and $\mathrm{rank}(\mathbb{X}) = r$, then*

$$\mathbb{E}\left[\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}})\right] \lesssim \frac{\sigma^2 r}{n},$$

and with probability at least $1 - \delta$,

$$\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}) \lesssim \frac{\sigma^2 r}{n} + \frac{\sigma^2}{n}\log\left(\frac{1}{\delta}\right).$$

Here we denote $a_n \lesssim b_n$ if there exists a universal constant C irrelevant to n or any other parameters such that $a_n \leq Cb_n$ for all n.

Proof. The proof strategy is not using the closed form of the least square. Instead, we directly use the fact that $\widehat{\beta}^{\mathrm{LS}}$ is minimizer of the sum of squared errors, which is generally called zero-order condition in optimization. This gives us

$$\|Y - \mathbb{X}\widehat{\beta}^{\mathrm{LS}}\|^2 \leq \|Y - \mathbb{X}\beta^*\|^2 = \|\mathbb{X}\beta^* + \varepsilon - \mathbb{X}\beta^*\|^2 = \|\varepsilon\|^2,$$

while on the other side, we have

$$\|Y - \mathbb{X}\widehat{\beta}^{\mathrm{LS}}\|^2 = \|\mathbb{X}\beta^* + \varepsilon - \mathbb{X}\widehat{\beta}^{\mathrm{LS}}\|^2 = \|\mathbb{X}(\widehat{\beta} - \beta^*)\|^2 - 2\left\langle \varepsilon, \mathbb{X}(\widehat{\beta} - \beta^*)\right\rangle + \|\varepsilon\|^2.$$

Consequently, combing the two inequalities above, we have

$$\|\mathbb{X}(\widehat{\beta} - \beta^*)\|^2 \leq 2\langle \varepsilon, \mathbb{X}(\widehat{\beta} - \beta^*)\rangle = 2\|\mathbb{X}(\widehat{\beta} - \beta^*)\|\left\langle \varepsilon, \frac{\mathbb{X}(\widehat{\beta} - \beta^*)}{\|\mathbb{X}(\widehat{\beta} - \beta^*)\|}\right\rangle. \tag{5.4}$$

In the following part of our proof, we will bound the right hand side of the above inequality via the "sup-out" trick. Let $\Phi = (\phi_1, \ldots, \phi_r) \in \mathbb{R}^{n\times r}$ be orthogonal matrix whose columns are the orthonormal bases of $\mathcal{C}(\mathbb{X})$ such that $\Phi^\top\Phi = \mathrm{I}_r$. Since $\mathbb{X}(\widehat{\beta} - \beta^*) \in \mathcal{C}(\mathbb{X})$, there exists $\nu = (\nu_1, \ldots, \nu_r)^\top \in \mathbb{R}^r$ such that $\mathbb{X}(\widehat{\beta} - \beta^*) = \sum_{j=1}^r \nu_j\phi_j = \Phi\nu$.

Therefore, denote $\widetilde{\varepsilon} = \Phi^\top\varepsilon \in \mathbb{R}^r$ and we have

$$\left\langle \varepsilon, \frac{\mathbb{X}(\widehat{\beta} - \beta^*)}{\|\mathbb{X}(\widehat{\beta} - \beta^*)\|}\right\rangle = \left\langle \varepsilon, \frac{\Phi\nu}{\|\Phi\nu\|_2}\right\rangle = \frac{\varepsilon^\top\Phi\nu}{\|\nu\|} = \left\langle \Phi^\top\varepsilon, \frac{\nu}{\|\nu\|}\right\rangle \leq \sup_{\|u\|\leq 1}\langle\widetilde{\varepsilon}, u\rangle = \|\widetilde{\varepsilon}\|,$$

where we use the "sup-out" trick in the first inequality above.

Combining the above inequality with (5.4), we have

$$\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}) = \frac{1}{n}\|\mathbb{X}(\widehat{\beta} - \beta^*)\|^2 = \frac{4}{n}\left\langle \varepsilon, \frac{\mathbb{X}(\widehat{\beta} - \beta^*)}{\|\mathbb{X}(\widehat{\beta} - \beta^*)\|}\right\rangle^2 \leq \frac{4\|\widetilde{\varepsilon}\|^2}{n}. \tag{5.5}$$

Therefore, we can control the expectation of MSE by

$$\mathbb{E}[\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}})] \leq \frac{4\mathbb{E}\|\widetilde{\varepsilon}\|^2}{n} = \frac{4}{n}\sum_{i=1}^r \mathbb{E}[\widetilde{\varepsilon}_i^2] \leq \frac{4\sigma^2 r}{n},$$

where we use the fact that $\mathbb{E}[\widetilde{\varepsilon}_i^2] = \mathbb{E}(\phi_i^\top\varepsilon)^2 \leq \sigma^2$ as variance is smaller than variance-proxy for sub-Gaussian.

To prove the tail-probability inequality of MSE, we need to apply Theorem 4.6 in Lecture 4, i.e., the maximal inequality for ℓ_2-norm. To implement that theorem, we need to verify that for any $u \in \mathbb{R}^r$, $\langle u, \widetilde{\varepsilon}\rangle$ is sub-Gaussian with variance-proxy $\sigma^2\|u\|^2$, which is true as

$$\mathbb{E}e^{\lambda\langle u, \widetilde{\varepsilon}\rangle} = \mathbb{E}e^{\lambda\langle u, \Phi^\top\varepsilon\rangle} = \mathbb{E}e^{\lambda\langle \Phi u, \varepsilon\rangle} \leq e^{\frac{\lambda^2}{2}\|\Phi u\|^2\sigma^2} = e^{\frac{\lambda^2}{2}\sigma^2\|u\|^2},$$

where in the first inequality, we use the fact that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and we use $\Phi^\top \Phi = \mathrm{I}_r$ in the last inequality

By (5.5), we have with probability at least $1 - \delta$

$$\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}) \leq \frac{4\|\widetilde{\varepsilon}\|^2}{n} \leq \frac{4}{n}\left[4\sigma\sqrt{r} + 2\sigma\sqrt{2\log(1/\delta)}\right]^2 \lesssim \frac{\sigma^2 r}{n} + \frac{\sigma^2}{n}\log\left(\frac{1}{\delta}\right),$$

where we use Theorem 4.6 in Lecture 4 in the second inequality above. □

Lecture 6

Compressive Sensing

6.1 High-dimensional Linear Models

In the high-dimensional setting, we're essentially looking at the same linear model $Y = \mathbb{X}\beta + \varepsilon$ with $\mathbb{X} \in \mathbb{R}^{n\times d}$. However, we now expect the number of features d is much larger than its sample size n. Under the high dimensional setting, the ordinary least squares estimator will have troubles. If the features are linearly independent, we have $\mathrm{rank}(\mathbb{X}) = n$. Then, $\mathbb{X}\widehat{\beta}^{\mathrm{LS}} = P_{\mathbb{X}}Y = Y$, i.e., the ordinary least squares will over-fit. Therefore, we need to invoke the parsimonious principle and introduce the following sparse linear model.

6.1 Definition (Sparse Linear Model). *A sparse linear model satisfies $Y = \mathbb{X}\beta + \varepsilon$ with $\mathbb{X} \in \mathbb{R}^{n\times d}$, but only s entries of β are non-zero. We also define the ℓ_0-norm*

$$\|\beta\|_0 = \sum_{j=1}^{d} \mathbb{1}\{\beta_j \neq 0\} = s.$$

Typically, the number of effective features s is much smaller than d. So the estimation of sparse linear model is like finding a needle in a haystack. A straightforward method to estimate the sparse linear model is to modify the least squares by adding the sparsity constraint:

$$\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2, \text{ s.t. } \|\beta\|_0 \leq s.$$

However, the sparsity constraint $\|\beta\|_0 \leq s$ is very complicated and will cause big troubles in computation. In fact, we can decompose the constraint as

$$\big\{\beta \in \mathbb{R}^d | \|\beta\|_0 \leq s\big\} = \bigcup_{\mathcal{S}:|\mathcal{S}|=s} \big\{\beta \in \mathbb{R}^d | \beta_j \neq 0 \text{ for all } j \in \mathcal{S} \subseteq \{1,\ldots,d\}\big\}.$$

From the decomposition above, there are $\binom{d}{s}$ possible subspace $\mathcal{S}$, which makes us have to solve $\binom{d}{s}$ least squares. This is not acceptable when d is very large.

An alternative idea is to relax the computationally infeasible ℓ_0 norm to some other norms which are computationally friendly. This motivates us to introduce the following ℓ_p-norms.

6.2 Definition (ℓ_p-norm). *The ℓ_p-norm of a d-dimensional vector $x = (x_1, ..., x_d)^\top$ is defined as*

$$\|\boldsymbol{x}\|_p = \Big(\sum_{j=1}^{d} |x_j|^p \Big)^{1/p}, \text{ for } p \in (0, \infty).$$

We also define the ℓ_∞-norm as $\|x\|_\infty = \max_{1 \le j \le d} |x_j|$.

We can see that the ℓ_2-norm used in the previous lectures are consistent with the definition above. Therefore, in the following parts of our lectures, we will add the subscript p to $\|\cdot\|$ to emphasize which ℓ_p-norm we are using. Notice that the triangular inequality $\|x+y\|_p \le \|x\|_p + \|y\|_p$ is only true when $p \ge 1$. Therefore, the ℓ_p-norm for $p \ge 1$ is a real norm while ℓ_p-norm for $p < 1$ is only a pseudo-norm. Similar to the variational form of the ℓ_2-norm: $\|x\|_2 = \sup_{\|y\|_2 \le 1} \langle x, y \rangle$, we also have the following varational form of general ℓ_p-norms for $p \ge 1$.

6.3 Proposition (Variational form of ℓ_p-norm). *For any $1 \le p \le \infty$, we have*

$$\|x\|_p = \sup_{\|y\|_q \le 1} \langle x, y \rangle, \text{ where } \frac{1}{p} + \frac{1}{q} = 1.$$

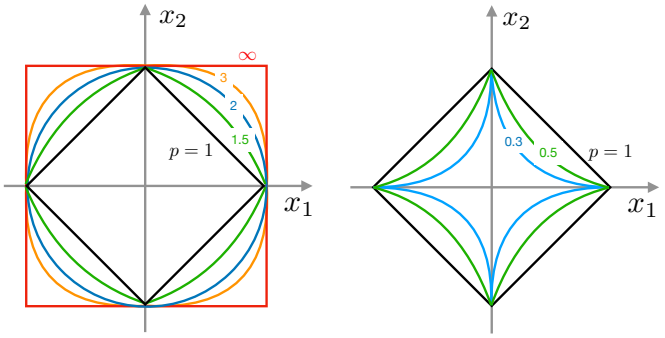
We call ℓ_q-norm is the dual norm of ℓ_p-norm, and vice versa. For example, ℓ_2-norm is the dual norm of itself. ℓ_∞-norm is the dual norm of ℓ_1-norm, and ℓ_1-norm is the dual norm of ℓ_∞-norm. Proposition 6.3 also implies the following inequality:

$$\text{Hölder inequality: } |\langle x, y \rangle| \le \|x\|_p \|y\|_q \text{ for any } \frac{1}{p} + \frac{1}{q} = 1.$$

We can see that the Cauchy-Schwarz inequality $|\langle x, y \rangle| \le \|x\|_2 \|y\|_2$ is a special case of the Hölder inequality for $p = 2$.

Now let us study the geometry of ℓ_p-norms. We visualize the ℓ_p-ball $\{x | \|x\|_p \le 1\}$ in Figure 6.1. In Figure 6.1(a), we can see that the ℓ_1-ball has a diamond shape and the ℓ_∞-ball is a square. When p increases from 1 to ∞, the diamond gradually expands to a square. In Figure 6.1(b), we can see that when $p < 1$, the ℓ_p-ball has a asteroid shape and it keeps shrinking as p goes to zero.

From Figure 6.1, we can see that the ℓ_1-ball is the smallest convex ball. Therefore, the ℓ_1-norm is the *convex relaxation* of ℓ_1-norm. This leads us to estimating the high dimen-



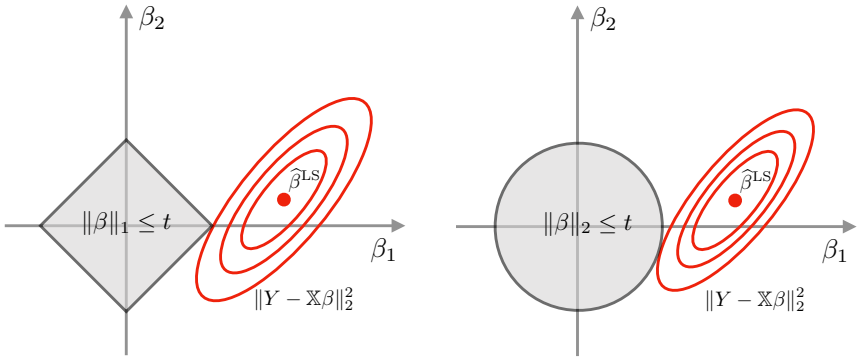
(a) $p \geq 1$ (b) $p \leq 1$

Figure 6.1. *The ℓ_p-ball $\{x|\ \|x\|_p \leq 1\}$ in two dimension space.*

sional linear model via the following convex problem:

$$\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2, \text{ s.t. } \|\beta\|_1 \leq t, \tag{6.4}$$

where t is some tuning parameter. We visualize the above problem in Figure 6.2(a). Comparing to the ℓ_2-norm in Figure 6.2(b), the diamond shape of ℓ_1-ball tends to make the solution of the optimization problem locate on the vertices of the polytope, which corresponding to sparse vectors. This geometric insight also explains why the ℓ_1-norms seems not directly encouraging sparsity like ℓ_0-norm by its definition but it can actually help us to the sparse solutions. On the other hand, the optimization problem on the convex ℓ_1-norm is much more feasible than the ℓ_0-norm. That is why ℓ_1-norm is so welcomed in the high dimensional statistics: it can help us to select variables and it is easy in computation.



(a) ℓ_1 constraint (b) ℓ_2 constraint

Figure 6.2. *The geometry of the optimization problem in* (6.4).

Applying the method of Lagrange multipliers, we know that for any $t > 0$, there exists a λ such that the constraint optimization problem in (6.4) is equivalent to the following unconstrained optimization problem.

The **Lasso** (Least Absolute Shrinkage and Selection Operator) estimator with tuning parameter λ is the minimizer of the optimization problem

$$\min_{\beta} \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

6.2 Compressive Sensing

Before studying Lasso, let us start with a simpler setting when the linear model is noiseless. Namely, we consider the linear equation $Y = \mathbb{X}\beta^*$ with the sparse truth $\|\beta^*\|_0 \leq s$. The noiseless problem can help us to build some intuition on the ℓ_1-norm and the linear regression.

The noiseless setting has its own importance in the applications of signal processing. Suppose Adam wants to sent a d-dimensional signal β^* to Bella. The signal is sparse but the total dimension d is too large such that it will be inefficient to send the original β^*. So can we compress the signal to lower dimensions and save Adam and Bob's time? To be more specific, can we find a matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$ such that Adam can compress β^* into an n-dimensional signal $Y = \mathbb{X}\beta^*$, where n could be much smaller than d. Instead of sending β^*, Adam can send Y to Bella, which is more efficient. If Bella has the matrix $\mathbb{X}$ in prior, she can recover the β^* via some compressive sensing algorithm. We then have two major problems in compressive sensing:

1. What is the algorithm to recover β^*?

2. What kind of matrix $\mathbb{X}$ can guarantee the recovery?

3. How efficiently we can compress β^*, i.e., how small n can be with respect to d?

Let us start with the first question. Ideally, we can find the true β^* by finding the sparsest β satisfying the linear equation $Y = \mathbb{X}\beta$, i.e., we consider the optimization problem

$$\min_{\beta} \|\beta\|_0, \text{ s.t. } Y = \mathbb{X}\beta.$$

However, we know that this problem is challenging in computation. We know that all possible solutions of the linear equation $Y = \mathbb{X}\beta$ is

$$\beta = (\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}Y + \Delta, \text{ for any } \Delta \in \mathrm{Null}(\mathbb{X}),$$

where $\mathrm{Null}(\mathbb{X}) = \{\Delta|\ \mathbb{X}\Delta = 0\}$. However, it is challenging to find the sparsest vector among these solutions as $(\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}Y$ may not be sparse.

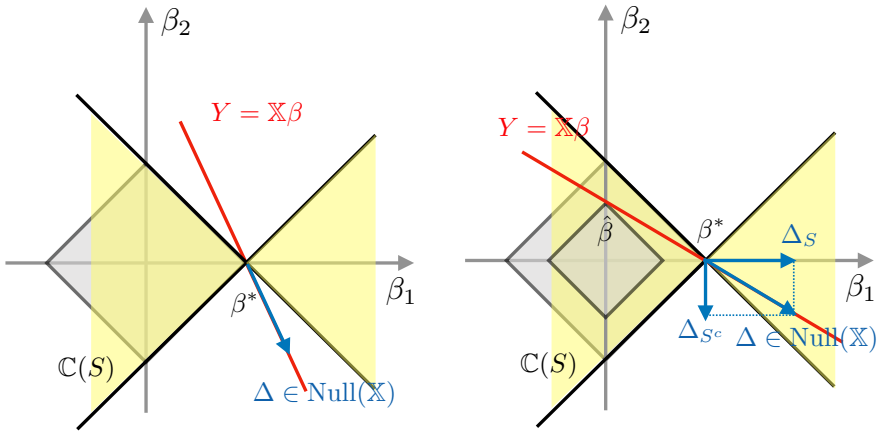
Similar to the discussion in the previous section, we can relax the ℓ_0-norm to ℓ_1-norm and consider the following estimator.

The **basis pursuit** estimator is the minimizer of the linear optimization problem

$$\widehat{\beta} = \arg\min_{\beta} \|\beta\|_1, \text{ s.t. } Y = \mathbb{X}\beta.$$

We can solve this problem efficiently via linear programming algorithms.

Now we study when the basis pursuit can recover the truth, i.e., $\widehat{\beta} = \beta^*$. From Figure 6.3(a), we can see that if the linear space $\{\beta | Y = \mathbb{X}\beta\}$ is tangent to the ℓ_1-ball $\{\beta | \ \|\beta\|_1 \leq \|\beta^*\|_1\}$, the basis pursuit estimator $\widehat{\beta}$ locates at the tangent point β^*, i.e., we have a perfect recovery. On the other hand, in Figure 6.3(b), if the linear space $\{\beta | Y = \mathbb{X}\beta\}$ gets into the ℓ_1-ball, $\widehat{\beta}$ does not equal to β^*. Therefore, Figure 6.3 shows us that the basis pursuit can recover the truth if and only if the null space $\text{Null}(\mathbb{X})$ does not located in the yellow area. In fact, the yellow area is a cone defined as follows.



(a) Perfect recovery (b) No perfect recovery

Figure 6.3. *The geometry of the basis pursuit and the cone condition.*

6.5 Definition (Cone condition). *We denote the support of β^* as $S = \{j | \beta_j^* \neq 0\}$. We also denote $\beta_S \in \mathbb{R}^{|S|}$ as the a subvector of β, whose entries are β_j's for all $j \in S$. We define the cone*

$$\mathbb{C}(S) = \{\Delta | \ \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

The cone $\mathbb{C}(S)$ corresponds to the yellow area in Figure 6.3. It contains the vectors whose ℓ_1-norm not on the support is dominated by the ℓ_1-norm on the support.

The following theorem formalizes our intuition in Figure 6.3 and gives a sufficient and necessary condition of when the basis pursuit is prefect recovery.

6.6 Theorem. *The basis pursuit has a unique solution $\widehat{\beta} = \beta^*$ for any $\beta^* \in \mathbb{R}^d$ with*

support S if and only if

$$\mathbb{C}(S) \bigcap \mathrm{Null}(\mathbb{X}) = \{0\}. \tag{6.7}$$

Proof. We first prove that (6.7) is a sufficient condition for perfect recovery.

Let $\widehat{\Delta} = \widehat{\beta} - \beta^*$. Therefore, $\widehat{\Delta} \in \mathrm{Null}(\mathbb{X})$ as $\mathbb{X}(\widehat{\beta} - \beta^*) = 0$. We aim to show that $\widehat{\Delta} \in \mathbb{C}(S)$ and by (6.7), $\widehat{\Delta} = 0$. By the definition of basis pursuit, we have $\|\beta^*\|_1 \geq \|\widehat{\beta}\|_1$. Moreover, we have

$$\begin{aligned} \|\beta^*_S\|_1 = \|\beta^*\|_1 \geq \|\widehat{\beta}\|_1 &= \|\beta^* + \widehat{\Delta}\|_1 \\ &= \|\beta^*_S + \widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1 \\ &\geq \|\beta^*_S\|_1 - \|\widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1. \end{aligned} \tag{6.8}$$

This implies that $\|\widehat{\Delta}_S\|_1 \geq \|\widehat{\Delta}_{S^c}\|_1$. Hence $\widehat{\Delta} \in \mathbb{C}(S)$ and by (6.7), $\widehat{\beta} = \beta^*$.

Now we prove that (6.7) is necessary for the perfect recovery. We aim to show that if $\widehat{\beta} = \beta^*$ for all β^* such that $\|\beta^*\|_0 \leq s$, then $\mathbb{C}(S) \bigcap \mathrm{Null}(\mathbb{X}) = \{0\}$.

For any $\beta^* \in \mathrm{Null}(\mathbb{X}) \backslash \{0\}$, our goal is to show that $\beta^* \notin \mathbb{C}(S)$. As the basis pursuit can always perfectly recover the truth, we have

$$\begin{pmatrix} \beta^*_S \\ 0 \end{pmatrix} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \|\beta\|_1, \text{ s.t. } \mathbb{X}\beta = \mathbb{X} \begin{pmatrix} \beta^*_S \\ 0 \end{pmatrix}.$$

As $\mathbb{X}\beta^* = 0$, the vector $(0, -\beta^{*\top}_{S^c})^\top$ is also a feasible solution to the problem. Therefore, by uniqueness of basis pursuit, we have $\|\beta^*_S\|_1 < \|\beta^*_{S^c}\|_1$, i.e., $\beta^* \notin \mathbb{C}(S)$. ∎

Lecture 7

Restricted Isometry Property

7.1 Restricted Isometry Property

In the previous lecture, we introduce the problem of compressive sensing: how to find the sparse truth β^* from the linear equation $Y = \mathbb{X}\beta^*$. Recall that we list three major questions for the compressive sensing:

1. What is the algorithm to recover β^*?
2. What kind of matrix $\mathbb{X}$ can guarantee the recovery?
3. How efficiently we can compress β^*, i.e., how small n can be with respect to d?

The first question is solved by the basis pursuit estimator $\widehat{\beta} = \arg\min_\beta \|\beta\|_1$ such that $Y = \mathbb{X}\beta$. The second question is partially answered in Theorem 6.6 of Lecture 6, as we show that the cone condition $\mathbb{C}(S) \bigcap \text{Null}(\mathbb{X}) = 0$ is a sufficient and necessary condition for the perfect recovery of basis pursuit in Theorem 6.6. However, the cone condition is not easy to use in practice. It is not straightforward to construct $\mathbb{X}$ starting from the cone condition. In this lecture, we will discuss another sufficient condition for perfect recovery, called restricted isometry property, which is stronger but easier to implement. We will talk about how to construct $\mathbb{X}$ based on this property and answer the third question.

7.1 Definition (Restricted isometry property). *We say $\mathbb{X}$ satisfies s-RIP with the coefficient $\delta_s \in (0, 1)$ if for any β with $\|\beta\|_0 \leq s$, we have*

$$(1 - \delta_s)\|\beta\|_2^2 \leq \|\mathbb{X}\beta\|_2^2 \leq (1 + \delta_s)\|\beta\|_2^2.$$

The RIP condition implies that the linear map $\mathbb{X}$ only distorts the norm of s-sparse

vectors by δ_s, i.e.,

$$\Big|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\Big| = \Big|\frac{\beta^\top\mathbb{X}^\top\mathbb{X}\beta}{\|\beta\|_2^2}-1\Big| \le \delta_s, \text{ for all } \|\beta\|_0 \le s.$$

Since $\mathbb{X}^\top Y = \mathbb{X}^\top\mathbb{X}\beta^*$, ideally, if $\mathbb{X}^\top\mathbb{X} = \mathrm{I}_d$, then we can directly recover the β^* by $\mathbb{X}^\top Y$. However, in high dimensional setting, we have $\mathrm{rank}(\mathbb{X}^\top\mathbb{X}) \le n \ll d$, thus it is impossible for $\mathbb{X}^\top\mathbb{X}$ to an identity matrix. However, we know β^* is s-sparse, so we only need $\mathbb{X}^\top\mathbb{X}$ to be somehow identity for s-sparse vectors. In that sense, the s-RIP condition assumes that $\mathbb{X}^\top\mathbb{X}$ is an "almost" identity matrix for all s-sparse vectors. And the level of almostness is measured by δ_s.

Therefore, the discussion above provides us the intuition that we can recover the truth if $\mathbb{X}^\top\mathbb{X}$ is almostly an identity. We now formalize this intuition by the theorem below.

7.2 Theorem (Perfect recovery under RIP). *If $\mathbb{X}$ is 3s-RIP with the coefficient $\delta_s < \frac{1}{3}$, then the basis pursuit can perfectly recover the truth, i.e., $\widehat{\beta} = \beta^*$.*

Proof. Let $\Delta = \widehat{\beta} - \beta^*$. Therefore, $\mathbb{X}(\beta - \beta^*) = 0$ and $\Delta \in \mathrm{Null}(\mathbb{X})$. Same as Equation (6.8) in the proof of Theorem 6.6 of Lecture 6, we have

$$\begin{aligned}\|\beta^*_S\|_1 = \|\beta^*\|_1 \ge \|\widehat{\beta}\|_1 &= \|\beta^* + \Delta\|_1 \\ &= \|\beta^*_S + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\ &\ge \|\beta^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1.\end{aligned}$$

Thus $\|\widehat{\Delta}_{S^c}\|_1 \le \|\widehat{\Delta}_S\|_1$. If we can also show that $\|\Delta_S\|_1 \le \rho\|\Delta_{S^c}\|_1$ for some $\rho < 1$, then Δ must be zero and $\widehat{\beta} = \beta^*$. Therefore, to complete the proof, it suffices to find a $\rho < 1$ and prove the inequality $\|\Delta_S\|_1 \le \rho\|\Delta_{S^c}\|_1$ under $3s$-RIP.

Our idea is to apply the $3s$-RIP to Δ. However, the RIP condition only applies to $3s$-sparse vectors. Therefore, our first step is to divide Δ into shorter subvectors as follows:

$$\begin{aligned}&S_0 = S,\\ &S_1 = \text{ entries of the } 2s \text{ largest absolute values in } \Delta_{S^c},\\ &S_2 = \text{ entries of the next } 2s \text{ largest absolute values in } \Delta_{S^c},\\ &\quad\vdots\end{aligned}$$

Recall that $\Delta \in \mathrm{Null}(\mathbb{X})$ and we have

$$0 = \|\mathbb{X}\Delta\|_2 = \|\mathbb{X}\textstyle\sum_{j\ge 0}\Delta_{S_j}\|_2 \ge \|\mathbb{X}(\Delta_{S_0} + \Delta_{S_1})\|_2 - \|\sum_{j\ge 2}\mathbb{X}\Delta_{S_j}\|_2.$$

Therefore, $\|\mathbb{X}(\Delta_{S_0} + \Delta_{S_1})\|_2 \le \|\sum_{j\ge 2}\mathbb{X}\Delta_{S_j}\|_2$. Applying $3s$-RIP on both sides of this inequality, we have

$$\begin{aligned}\sqrt{(1-\delta_{3s})}\|\Delta_{S_0} + \Delta_{S_1}\|_2 &\le \|\mathbb{X}(\Delta_{S_0} + \Delta_{S_1})\|_2 \\ &\le \|\textstyle\sum_{j\ge 2}\mathbb{X}\Delta_{S_j}\|_2 \le \sqrt{(1+\delta_{3s})}\sum_{j\ge 2}\|\Delta_{S_j}\|_2.\end{aligned}$$

The above result has established an inequality between Δ_S and Δ_{S^c}, but in ℓ_2-norm. Recall that our goal is to prove $\|\Delta_S\|_1 \le \rho\|\Delta_{S^c}\|_1$, which is in ℓ_1-norm. Therefore, the remaining part of the proof is to bridge between ℓ_2-norm and ℓ_1-norm. We have

$$\begin{aligned}\sqrt{(1-\delta_{3s})}\|\Delta_S + \Delta_{S_1}\|_2 &\le \sqrt{(1+\delta_{3s})}\sum_{j\ge 2}\|\Delta_{S_j}\|_2 \\ &\le \sqrt{(1+\delta_{3s})}\sum_{j\ge 2}\sqrt{2s}\|\Delta_{S_j}\|_\infty \\ &\le \sqrt{(1+\delta_{3s})}\sum_{j\ge 2}\frac{1}{\sqrt{2s}}\|\Delta_{S_{j-1}}\|_1 = \sqrt{\frac{(1+\delta_{3s})}{2s}}\|\Delta_{S^c}\|_1,\end{aligned}$$

where in the second inequality, we use the fact that $\|x\|_2 \le \sqrt{d}\|x\|_\infty$, for any $x \in \mathbb{R}^d$, and in the last inequality, we use the inequality $\|\Delta_{S_j}\|_\infty \le \frac{1}{2s}\|\Delta_{S_{j-1}}\|_1$ for $j \ge 2$ as we construct Δ_{S_j} in ordered, thus the maximum norm of Δ_{S_j} is smaller than the average of the preceded subvector $\Delta_{S_{j-1}}$.

On the left hand side of the above inequality, we have

$$\sqrt{(1-\delta_{3s})}\|\Delta_S + \Delta_{S_1}\|_2 \ge \sqrt{(1-\delta_{3s})}\|\Delta_S\|_2 \ge \sqrt{\frac{(1-\delta_{3s})}{s}}\|\Delta_S\|_1,$$

where in the last inequality, we use the fact that $\|x\|_1 \le \sqrt{d}\|x\|_2$, for any $x \in \mathbb{R}^d$. Summarizing the inequalities above, we have

$$\|\Delta_S\|_1 \le \sqrt{\frac{(1+\delta_{3s})}{2(1-\delta_{3s})}}\|\Delta_{S^c}\|_1,$$

and therefore, if $\delta_{3s} < 1/3$, we choose $\rho := \sqrt{\frac{(1+\delta_{3s})}{2(1-\delta_{3s})}} < 1$ and $\|\Delta_S\|_1 \le \rho\|\Delta_{S^c}\|_1$. □

Now we are ready to answer the second and third questions listed in the beginning of this lecture. By Theorem 7.2, it suffices to find $\mathbb{X}$ satisfying $3s$-RIP. Although the $3s$-RIP is a deterministic property of a matrix, the easiest way to find such matrix is to select it randomly. As we discussed above, RIP condition implies that $\mathbb{X}^\top\mathbb{X}$ is "almost" identity. If we can randomly select $\mathbb{X}$ such that $\mathbb{E}[\mathbb{X}^\top\mathbb{X}] = \mathrm{I}_d$, then by the concentration principle, we may expect $\mathbb{X}^\top\mathbb{X}$ has similar property as the identity. The following theorem gives us a concrete method to construct $\mathbb{X}$ satisfying $3s$-RIP.

7.3 Theorem. *Suppose the entries of $\mathbb{X} \in \mathbb{R}^{n\times d}$ are i.i.d. $N\Big(0, \frac{1}{n}\Big)$. For any $\varepsilon > 0$, $\delta \in (0,1)$, if $n \ge \frac{96}{\delta^2}s\log(\frac{18d}{\varepsilon})$, then*

$$\mathbb{P}\big(\mathbb{X} \textit{ is 3s-RIP with the coefficient } \delta\big) \ge 1-\varepsilon.$$

This theorem answers the questions of how to construct $\mathbb{X}$ and how efficiently we can compress the signal. We can compress a d-dimensional s-sparse signal into $n = O(s\log d)$ dimensions and perfectly recover it via basis pursuit by Theorem 7.2.

Proof. By the definition of $3s$-RIP, we need to control the probability of the event

$$\left\{\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\right|>\delta \text{ for all } \|\beta\|_0\le 3s\right\}=\left\{\sup_{\|\beta\|_0\le 3s}\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\right|>\delta\right\}.$$

This is an event involving maximums and we will apply the maximal inequality and the discretization trick. Let us start with a fixed β. Given a $\beta\in\mathbb{R}^d$, we have $X_i^\top\beta$ for $i=1,\dots,n$ are i.i.d. $N(0,\frac{1}{n}\|\beta\|_2^2)$. Therefore, we have

$$\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}=\sum_{i=1}^n\Big(\frac{X_i^\top\beta}{\|\beta\|_2^2}\Big)^2\sim\frac{1}{n}\chi_n^2.$$

By the concentration inequality of χ_n^2 (see Homework 1, Q2(b) for example), we have

$$\mathbb{P}\bigg(\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\right|>\delta\bigg)=\mathbb{P}\bigg(\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-\mathbb{E}\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}\right|>\delta\bigg)\le 2e^{-\frac{n\delta^2}{8}}. \tag{7.4}$$

The second step is to apply the ϵ-net. We can decompose the ℓ_0-ball $\|\beta\|_0\le 3s$ into $\binom{d}{3s}$ number of $3s$-dimensional subspaces. Therefore, by the union bound, we have

$$\begin{aligned}\mathbb{P}\bigg(\sup_{\|\beta\|_0\le 3s}\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\right|>\delta\bigg)&\le\binom{d}{3s}\mathbb{P}\bigg(\sup_{\beta\in\mathcal{B}_2^{3s}}\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\right|>\delta\bigg)\\&=\binom{d}{3s}\mathbb{P}\bigg(\sup_{u\in\mathcal{B}_2^{3s}}\big|\|\mathbb{X}u\|_2^2-1\big|>\delta\bigg).\end{aligned}$$

where we abuse the notation in the first inequality: when $\beta\in\mathcal{B}_2^{3s}$, the number of columns of $\mathbb{X}$ becomes $3s$ automatically.

We denote $A=\mathbb{X}^\top\mathbb{X}-\mathrm{I}$ and $\|\mathbb{X}u\|_2^2-1=u^\top(\mathbb{X}^\top\mathbb{X}-\mathrm{I})u=u^\top Au$. Let $\mathcal{N}_{1/4}\subseteq\mathcal{B}_2^{3s}$ be the $\frac{1}{4}$-net of $\mathcal{B}_2^{3s}$. Then for any $u\in\mathcal{B}_2^{3s}$, there exists a $v\in\mathcal{N}_\epsilon$ such that $\|u-v\|_2\le\frac{1}{4}$. Therefore, we have

$$\begin{aligned}\sup_{u\in\mathcal{B}_2^{3s}}|u^\top Au|&\le\sup_{u\in\mathcal{B}_2^{3s}}|(u-v)^\top Au+v^\top A(u-v)+v^\top Av|\\&\le\sup_{u\in\mathcal{B}_2^{3s}}\frac{1}{4}|u^\top Au|+\sup_{u\in\mathcal{B}_2^{3s}}\frac{1}{4}|u^\top Au|+\sup_{v\in\mathcal{N}_{1/4}}|v^\top Av|.\end{aligned}$$

Therefore, $\sup_{u\in\mathcal{B}_2^{3s}}|u^\top Au|\le 2\sup_{v\in\mathcal{N}_{1/4}}|v^\top Av|$ and we have

$$\begin{aligned}\mathbb{P}\bigg(\sup_{\|\beta\|_0\le 3s}\left|\frac{\|\mathbb{X}\beta\|_2^2}{\|\beta\|_2^2}-1\right|>\delta\bigg)&\le\binom{d}{3s}\mathbb{P}\bigg(\sup_{u\in\mathcal{B}_2^{3s}}\big|\|\mathbb{X}u\|_2^2-1\big|>\delta\bigg)\\&\le\binom{d}{3s}\mathbb{P}\bigg(\sup_{v\in\mathcal{N}_{1/4}}\big|\|\mathbb{X}v\|_2^2-1\big|>\frac{\delta}{2}\bigg)\\&\le d^{3s}|\mathcal{N}_{1/4}|\cdot\mathbb{P}\bigg(\left|\frac{\|\mathbb{X}v\|_2^2}{\|v\|_2^2}-1\right|>\frac{\delta}{2}\bigg)\qquad(\text{As }\|v\|_2=1)\\&\le(9d)^{3s}\cdot 2e^{-\frac{n\delta^2}{32}},\end{aligned}$$

where in the last inequality, we use the facts (7.4) and $|\mathcal{N}_{1/4}| \leq 9^{3s}$ by Lemma 4.7 in Lecture 4. Solve $\varepsilon \geq 2(9d)^{3s}e^{-\frac{n\delta^2}{32}}$ and we have $n \geq \frac{96}{\delta^2}s\log(\frac{18d}{\varepsilon})$. □

Lecture 8

Statistical Properties of Lasso

8.1 Restricted Eigenvalue Condition

In this lecture, we return to the noisy linear regression. Recall the sparse linear model $Y = \mathbb{X}\beta^* + \varepsilon$, where $\mathbb{X} \in \mathbb{R}^{n\times d}$ and $\|\beta^*\|_0 \le s$. We estimate the high dimensional linear model via the Lasso estimator

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

In this lecture, we will study the statistical properties of the Lasso estimator. Like the RIP condition for the basis pursuit, we also need conditions for Lasso.

8.1 Definition (Restricted eigenvalue). *Let S be the support of β^*. We sat $\mathbb{X}$ satisfies the restricted eigenvalue condition $\mathrm{RE}(\kappa, \alpha)$ if*

$$\frac{1}{n}\|\mathbb{X}\Delta\|_2^2 \ge \kappa\|\Delta\|_2^2, \text{ for all } \Delta \in \mathbb{C}_\alpha(S) := \{\Delta |\ \|\Delta_{S^c}\|_1 \le \alpha\|\Delta_S\|_1\}. \tag{8.2}$$

We first compare restricted eigenvalue (RE) condition with the restricted isometry property (RIP). The restricted isometry property states that

$$(1-\delta_s)\|\Delta\|_2^2 \le \|\mathbb{X}\Delta\|_2^2 \le (1+\delta_s)\|\Delta\|_2^2,$$

for all $\|\Delta\|_0 \le s$. Comparing to the two-sided inequality on $\|\mathbb{X}\Delta\|_2^2$ in RIP, the restricted eigenvalue condition only impose a one-sided inequality. Moreover, the word "restricted" in RIP means the inequality is truth for $\|\Delta\|_0 \le s$ but the RE restricts Δ in the cone $\|\Delta_{S^c}\|_1 \le \alpha\|\Delta_S\|_1$. Therefore, the RE condition is less restricted comparing to RIP.

In fact, notice that $\frac{1}{n}\|\mathbb{X}\Delta\|_2^2 = \Delta^\top\widehat{\Sigma}\Delta$, where $\widehat{\Sigma} = \mathbb{X}^\top\mathbb{X}/n$ is the sample covariance matrix of the features. Thus the RE condition is also related to $\widehat{\Sigma}$. By Theorem 1.18 in

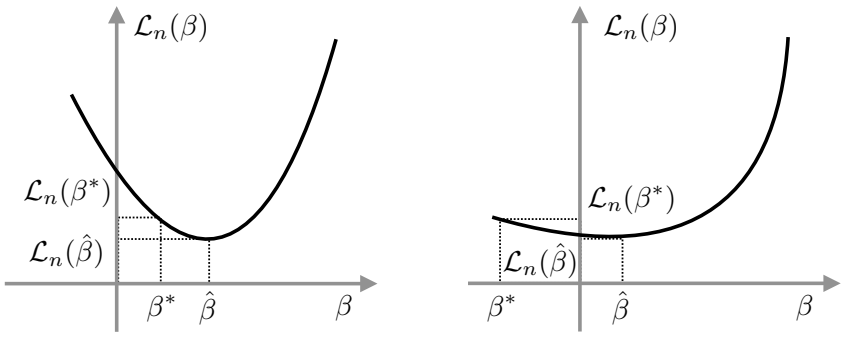
Lecture 1, the minimum eigenvalue of sample covariance matrix $\lambda_{\min}(\widehat{\Sigma})$ has the following variational form

$$\lambda_{\min}(\widehat{\Sigma}) = \min_{\Delta} \frac{\Delta^\top \widehat{\Sigma} \Delta}{\|\Delta\|_2^2} = \min_{\Delta} \frac{\Delta^\top \mathbb{X}^\top \mathbb{X} \Delta}{n\|\Delta\|_2^2} = \min_{\Delta} \frac{1}{n} \frac{\|\mathbb{X}\Delta\|_2^2}{\|\Delta\|_2^2}. \tag{8.3}$$

In comparison, the RE condition in (8.2) is equivalent to the formulation

$$\min_{\Delta \in \mathbb{C}_\alpha(S)} \frac{1}{n} \frac{\|\mathbb{X}\Delta\|_2^2}{\|\Delta\|_2^2} \geq \kappa.$$

This is why we call it the restricted eigenvalue condition, as the κ is somehow the minimum eigenvalue of $\widehat{\Sigma}$ restricted on the cone $\mathbb{C}_\alpha(S)$. We know that a matrix A is positive definition, denoted as $A \succ 0$, if $\lambda_{\min}(A) > 0$. However, under the high dimensional setting when $d \gg n$, we have $\operatorname{rank}(\mathbb{X}) \leq n \ll d$ so $\lambda_{\min}(\widehat{\Sigma}) = 0$. Namely, $\widehat{\Sigma}$ is positive semi-definite, denoted as $\widehat{\Sigma} \succeq 0$. In that sense, the RE condition assumes that $\mathbb{X}^\top \mathbb{X}/n$ is "almost" positive definite and the word "almost" means we restricted on the cone. Recall that in the previous lecture, we say RIP assumes that $\mathbb{X}^\top \mathbb{X}$ is almost identity, which is more stringent than positive definiteness. This is because we need perfect recovery in the compressive sensing, while we only expect good statistical rate for Lasso.

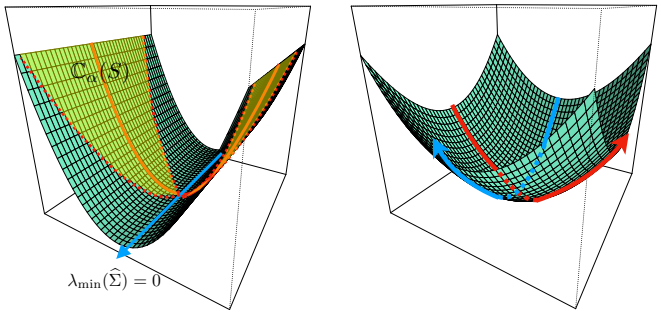


(a) Curved (b) Flat

Figure 8.1. *The curvature of $\mathcal{L}_n(\beta)$ and its impact to the rate of $\widehat{\beta}$.*

Now we provide some intuition why the analysis of Lasso requires the RE condition. This is due to the landscape of the least squares loss $\mathcal{L}_n(\beta) = \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2$. By concentration principle, we may expect that $\mathcal{L}_n(\widehat{\beta})$ is close to $\mathcal{L}_n(\beta^*)$. From Figure 8.1(a), we can see that if $\mathcal{L}_n(\beta)$ is convex enough, then $\widehat{\beta}$ is also close to β^*. However, if $\mathcal{L}_n(\beta)$ is very flat, then $\widehat{\beta}$ could be far away from β^* even if $\mathcal{L}_n(\widehat{\beta})$ is close to $\mathcal{L}_n(\beta^*)$. This means that the curvature of $\mathcal{L}_n(\beta)$. We know that the curvature of a curve can be characterized by the Hessian matrix $\nabla^2 \mathcal{L}_n(\widehat{\beta})$. The larger the minimum eigenvalue of the Hessian $\lambda_{\min}(\nabla^2 \mathcal{L}_n(\widehat{\beta}))$ is, the more convex $\mathcal{L}_n(\widehat{\beta})$ is. However, for the least squares loss in the high dimensional setting, $\nabla^2 \mathcal{L}_n(\widehat{\beta}) = \mathbb{X}^\top \mathbb{X}/n = \widehat{\Sigma}$, whose minimum eigenvalue $\lambda_{\min}(\widehat{\Sigma}) = 0$. So the square loss is like Figure 8.2(a). In comparison the case when the Hessian is positive definition, see Figure 8.2(b), we can see that the curved space in Figure 8.2(a) is flat along one

direction and curved along another. But it is not the end of the world. Remember that β^* is sparse, so we do not need the loss to be curved along all directions, but only along some special directions related to the sparsity of β^*. In specific, the RE condition assumes that the loss is curved along the cone $\mathbb{C}_\alpha(S)$.



(a) $\widehat{\Sigma} \succeq 0$ (b) $\widehat{\Sigma} \succ 0$

Figure 8.2. *The surface of two types of Hessian matrices.*

8.2 Statistical Rate of Lasso

With the restricted eigenvalue condition, we can present the rate of Lasso estimator.

8.4 Theorem (Rate of Lasso). *Let the cardinality of the support of β^* be $|S| = s$. Suppose $\mathbb{X}$ has $\mathrm{RE}(\kappa, 3)$. If we choose the tuning parameter $\lambda \geq \frac{2}{n}\|\mathbb{X}^\top \varepsilon\|_\infty$, then*

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda. \tag{8.5}$$

This theorem is totally deterministic: we do not impose any random assumption on the the noise ε. On the right hand side of (8.5), we can see that the rate of $\widehat{\beta}$ depends on κ, s and λ. The larger κ is, i.e., the more curved the loss is, the better is the rate. The rate also depends on the choice of the tuning parameter. This implies that as lambda increases, the rate will suffer as the estimator is more biased. The theorem suggests us to choose $\lambda = \frac{2}{n}\|\mathbb{X}^\top \varepsilon\|_\infty$; however, ε is unknown. We will handle this issue in the following corollary. Before that, we first present the proof of the theorem.

Proof. In the first part of the proof, We aim to show that if $\lambda \geq \frac{2}{n}\|\mathbb{X}^\top \varepsilon\|_\infty$, then the error vector $\Delta = \widehat{\beta} - \beta^*$ has $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$.

Our proof strategy is to begin with the zero-order optimization condition. Knowing the loss function is minimized for $\widehat{\beta}$, we have

$$\frac{1}{2n}\|Y - \mathbb{X}\widehat{\beta}\|_2^2 + \lambda\|\widehat{\beta}\|_1 \leq \frac{1}{2n}\|Y - \mathbb{X}\beta^*\|_2^2 + \lambda\|\beta^*\|_1 = \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda\|\beta^*\|_1$$

We expand the left hand side of the above inequality as

$$\begin{aligned}\frac{1}{2n}\|Y-\mathbb{X}\widehat{\beta}\|_2^2+\lambda\|\widehat{\beta}\|_1 &= \frac{1}{2n}\|\mathbb{X}\beta^*-\mathbb{X}\widehat{\beta}+\|\varepsilon\|_2^2+\lambda\|\widehat{\beta}\|_1\\ &= \frac{1}{2n}\|\mathbb{X}\Delta\|_2^2-\left\langle\varepsilon,\frac{1}{n}\mathbb{X}^\top\Delta\right\rangle+\frac{1}{2n}\|\varepsilon\|_2^2+\lambda\|\widehat{\beta}\|_1.\end{aligned}$$

Rearranging the above two inequalities, we have

$$\begin{aligned}0\le\frac{1}{2n}\|\mathbb{X}\Delta\|_2^2 &\le \frac{\varepsilon^\top\mathbb{X}\Delta}{n}+\lambda\|\beta^*\|_1-\lambda\|\widehat{\beta}\|_1\\ &\le \frac{1}{n}\|\mathbb{X}^\top\varepsilon\|_\infty\|\Delta\|_1+\lambda\|\beta_S^*\|_1-\lambda\|\widehat{\beta}_S\|_1-\lambda\|\widehat{\beta}_{S^c}\|_1 \quad \text{(By Hölder inequality)}\\ &\le \frac{\lambda}{2}\|\Delta\|_1+\lambda\|\Delta_S\|_1-\lambda\|\Delta_{S^c}\|_1 \qquad (\text{As } \lambda\ge 2\|\mathbb{X}^\top\varepsilon\|_\infty/n)\\ &= \frac{\lambda}{2}(3\|\Delta_S\|_1-\|\Delta_{S^c}\|_1),\end{aligned}$$

where in the third inequality, we apply the triangle inequality $\|\beta_S^*\|_1-\|\widehat{\beta}_S\|_1\le\|\widehat{\beta}_S-\beta_S^*\|_1$ and $\Delta_{S^c}=\widehat{\beta}_{S^c}$ as $\beta_{S^c}^*=0$. Therefore, we have $\|\Delta_{S^c}\|_1\le 3\|\Delta_S\|_1$ and thus $\Delta\in\mathbb{C}_3(S)$.

In the second part of the proof, we show the rate of Δ. As $\Delta\in\mathbb{C}_3(S)$, by $\mathrm{RE}(\kappa,3)$, $\frac{1}{n}\|\mathbb{X}\Delta\|_2^2\ge\kappa\|\Delta\|_2^2$. Going back to the inequality above, we have

$$\kappa\|\Delta\|_2^2\le\frac{1}{n}\|\mathbb{X}\Delta\|_2^2\le\lambda(3\|\Delta_S\|_1-\|\Delta_{S^c}\|_1)\le 3\lambda\|\Delta_S\|_1\le 3\lambda\sqrt{s}\|\Delta_S\|_2\le 3\lambda\sqrt{s}\|\Delta\|_2,$$

where in the fourth inequality, we use $\|x\|_1\le\sqrt{d}\|x\|_2$ for $x\in\mathbb{R}^d$. Therefore, we have $\|\Delta\|_2\le\frac{3}{\kappa}\sqrt{s}\lambda$. □

Now, we discuss how to choose λ when the noise ε is random.

8.6 Corollary. *If the noises $\varepsilon_1,\dots,\varepsilon_n$ are independent, and ε_i is subgaussian with variance proxy σ^2 for all $i=1,\dots,n$. The design matrix $\mathbb{X}$ is normalized such that the variance of the jth column of the design matrix $\mathbb{X}_j$ has $\frac{1}{n}\|\mathbb{X}_j\|_2^2\le 1$ for all $1\le j\le d$. If $\mathbb{X}$ satisfies $\mathrm{RE}(\kappa,3)$ and we choose $\lambda=\sigma\sqrt{\log(2d/\delta)/(2n)}$, then with probability at least $1-\delta$, we have*

$$\|\widehat{\beta}-\beta^*\|_2\le\frac{3\sigma}{2\kappa}\sqrt{\frac{2s\log(2d/\delta)}{n}}.$$

This corollary implies that if we choose the tuning parameter $\lambda=C\sqrt{\log d/n}$ for some sufficiently large constant C, the Lasso estimator has the ℓ_2-norm rate

$$\|\widehat{\beta}-\beta^*\|_2=O_P\left(\sqrt{\frac{s\log d}{n}}\right).$$

The Lasso estimator is consistent as long as $s\log d/n=o(1)$: if s is fixed, the dimension d can increase at the rate of the exponential of sample size. If we know the true support S

and run the least squares $\widehat{\beta}^{\mathrm{LS}} = \arg\min_{\beta_S} \|Y - \mathbb{X}_S\beta_S\|_2^2$, Theorem 5.3 in Lecture 5 shows the rate of OLS is

$$\|\widehat{\beta}^{\mathrm{LS}} - \beta^*\|_2 = O_P(\sqrt{r/n}).$$

In comparison, the Lasso estimator has an additional $\sqrt{\log d}$ term in the rate, which is the price for variable selection since Lasso does not know the true support.

Proof. By Theorem 8.4, we need to control $\|\mathbb{X}^\top\varepsilon\|_\infty/n$. We have

$$\mathbb{P}\bigg(\frac{1}{n}\|\mathbb{X}^\top\varepsilon\|_\infty > t\bigg) = \mathbb{P}\bigg(\max_{1\le j\le d}\Big|\frac{1}{n}\mathbb{X}_j^\top\varepsilon\Big| > t\bigg) \le d\mathbb{P}(|X_j^\top\varepsilon| > nt).$$

By Hoeffing inequality (Theorem 2.10 in Lecture 2), for any $j = 1,\dots,d$, $\mathbb{X}_j^\top\varepsilon$ is sub-Gaussian with the variance proxy $\sigma^2\|\mathbb{X}_j\|_2^2 \le \sigma^2 n$. The sub-Gaussian tail gives us

$$\mathbb{P}\bigg(\frac{1}{n}\|\mathbb{X}^\top\varepsilon\|_\infty > t\bigg) \le d\mathbb{P}(|\mathbb{X}_j^\top\varepsilon| > nt) \le 2de^{-nt^2/(2\sigma^2)}.$$

Therefore, if we choose $\lambda = \sqrt{\frac{\sigma^2\log(2d/\delta)}{2n}}$, with probability at least $1-\delta$, we have $\lambda \ge \frac{2}{n}\|\mathbb{X}^\top\varepsilon\|_\infty$, and by Theorem 8.4,

$$\|\widehat{\beta} - \beta^*\|_2 \le \frac{3}{\kappa}\sqrt{s}\lambda = \frac{3}{\kappa}\sqrt{\frac{\sigma^2 s\log(2d/\delta)}{2n}}.$$

□

In the last part of the lecture, we give a concrete example of the design matrix $\mathbb{X}$ satisfying the restricted eigenvalue condition. Similar to how we construct $\mathbb{X}$ satisfying RIP in Theorem 7.3 in Lecture 7, we also generate $\mathbb{X}$ randomly.

8.7 Proposition. *Let the rows of $\mathbb{X}$: $X_1,\dots,X_n$ be i.i.d. $N(0,\Sigma)$. We define the matrix maximum norm $\|A\|_{\max} = \max_{i,j}|A_{ij}|$. For any $\delta \in (0,1)$, if $64s\sqrt{\frac{\log(d/\delta)}{n}} \le \frac{\lambda_{\min}(\Sigma)}{\|\Sigma\|_{\max}}$, then*

$$\mathbb{P}\big(\mathbb{X} \textit{ satisfies } \mathrm{RE}(\lambda_{\min}(\Sigma)/2, 3)\big) \ge 1-\delta.$$

Therefore, if $s\sqrt{\log d/n} = o(1)$ and $X_i \overset{\text{iid}}{\sim} N(0,\Sigma)$, then $\mathbb{X}$ satisfies the RE condition with high probability. In fact, the scaling condition here is suboptimal. It is proved in [1] that the normal design is RE with high probability if $\sqrt{s\log d/n} = o(1)$.

Proof. We first control the rate of the sample covariance matrix in $\|\cdot\|_{\max}$-norm. Since $X_i \sim N(0,\Sigma)$, for any $1 \le i,j \le d$, $X_{ij}X_{ik}$ is sub-exponential with the parameter

$\Sigma_{jj}\Sigma_{kk} \le \|\Sigma\|_{\max}^2$. Therefore, we have

$$\begin{aligned}
\mathbb{P}\Big(\|\widehat{\Sigma}-\Sigma\|_{\max} > t\Big) &= \mathbb{P}\Big(\max_{1\le j\le k\le d} |\widehat{\Sigma}_{jk}-\Sigma_{jk}| > t\Big) \\
&\le d^2\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^n \Big(X_{ij}X_{ik}-\mathbb{E}[X_{ij}X_{ik}]\Big)\Big| > t\right) \qquad \text{(By union bound)} \\
&\le d^2 e^{-\frac{nt^2}{2\|\Sigma\|_{\max}^2}}. \qquad \text{(By the concentration of sub-exponential)}
\end{aligned}$$

Therefore, with probability at least $1-\delta$,

$$\|\widehat{\Sigma}-\Sigma\|_{\max} \le \|\Sigma\|_{\max}\sqrt{\frac{4\log(d/\delta)}{n}}.$$

Second, we will show $\mathbb{X}$ satisfies RE. We have

$$\frac{1}{n}\|\mathbb{X}\Delta\|_2^2 = \Delta^\top\widehat{\Sigma}\Delta = \Delta^\top(\widehat{\Sigma}-\Sigma)\Delta+\Delta^\top\Sigma\Delta \ge \lambda_{\min}(\Sigma)\|\Delta\|_2^2-\|\widehat{\Sigma}-\Sigma\|_{\max}\|\Delta\|_1^2, \quad (8.8)$$

where in the last inequality, we use (8.3) and the matrix Hölder inequality

$$|x^\top Ax| \le \|x\|_1\|Ax\|_\infty \le \|A\|_{\max}\|x\|_1^2.$$

When $\Delta \in \mathbb{C}_3(S)$, we have $\|\Delta\|_1 = \|\Delta_S\|_1+\|\Delta_{S^c}\|_1 \le 4\|\Delta_S\|_1 \le 4\sqrt{s}\|\Delta\|_2$. Therefore, with probability at least $1-\delta$,

$$\|\widehat{\Sigma}-\Sigma\|_{\max}\|\Delta\|_1^2 \le 32\|\Sigma\|_{\max}\sqrt{\frac{s^2\log(d/\delta)}{n}}\|\Delta\|_2^2 \le \frac{\lambda_{\min}(\Sigma)}{2}\|\Delta\|_2^2,$$

as we assume $64s\sqrt{\frac{\log(d/\delta)}{n}} \le \frac{\lambda_{\min}(\Sigma)}{\|\Sigma\|_{\max}}$. Combining with (8.8), we have

$$\frac{1}{n}\|\mathbb{X}\Delta\|_2^2 \ge \frac{1}{2}\lambda_{\min}(\Sigma)\|\Delta\|_2^2.$$

□

Lecture 9

Variations of Lasso

9.1 Limitations of Lasso

In the previous lecture, we study the high dimensional linear model $Y = \mathbb{X}\beta^* + \epsilon$, with $\mathbb{X} \in \mathbb{R}^{n\times d}$ and $\|\beta^*\|_0 \le s$. We propose to estimate β^* via Lasso estimator

$$\widehat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

We consider two assumptions: (1) the design matrix satisfies the restricted eigenvalue condition and (2) the noise ε are independent sub-Gaussians with variance proxy σ^2. If we choose $\lambda = C\sigma\sqrt{\log d/n}$ for some sufficiently large constant C, we show that the Lasso estimator has the statistical rate $\|\widehat{\beta}^{\text{Lasso}} - \beta^*\|_2 = O_P(\sqrt{s\log d/n})$.

However, the Lasso estimator is not perfect. In this lecture, we will discuss the extensions of Lasso. Due to the scope of our lecture, we could not cover all details of these extensions. We will list the related references for the further reading[2]. In order to extend Lasso, we first list the limitations of Lasso below. We will elaborate these limitations in the remaining of this lecture.

Limitations of Lasso

1. **Model**: Lasso is restricted to the linear model;
2. **Penalty**: The ℓ_1-penalty of Lasso can only regularize the sparsity;
3. **Bias**: Lasso estimator is biased;
4. **Tuning**: The choice of λ depends on the unknown σ;
5. **Robustness**: Lasso is consistent only when ε has light tail.

[2]The list is highly incomplete and we refer to the book Fan et al. (2020) for a more thorough review.

9.2 Beyond Linear Model

We first generalize the Lasso to general high dimensional models. In general, if we observe the samples $(Y_1, X_1), \ldots, (Y_n, X_n)$ i.i.d. generated from the density $p_{\beta^*}(y, x)$. If the dimension of the parameter $\beta^* \in \mathbb{R}^d$ is sparse, we can estimate β^* via adding a ℓ_1-penalty to the log-likelihood. Therefore, we have the following general Lasso estimation

$$\min_{\beta} \sum_{i=1}^{n} -\log p_{\beta}(Y_i, X_i) + \lambda \|\beta\|_1.$$

In fixed dimensional statistics, besides the maximal likelihood estimator, we may also consider some other loss between Y_i and X_i's, denoted as $\mathcal{L}_n(\beta)$. In principle, we can transform them to estimate sparse parameters by considering the problem $\min_\beta \mathcal{L}_n(\beta) + \lambda\|\beta\|_1$. We will give a few concrete examples below.

9.2.1 High Dimensional Classification

Let us start with the classification problem. Comparing to the regression problem, the classification has the binary response $Y \in \{-1, +1\}$. The arguably most popular model in classification is the logistic model. The **linear logistic model** assumes

$$\mathbb{P}(Y = +1|X = x) = \frac{1}{1 + e^{-x^\top \beta^*}} \text{ and } \mathbb{P}(Y = -1|X = x) = \frac{1}{1 + e^{x^\top \beta^*}}.$$

If we observe i.i.d. samples $(Y_1, X_1), \ldots, (Y_n, X_n)$ from the linear logistic model and β^* is sparse, we can estimate β^* via adding the ℓ_1-penalty to the logistic loss as follows.

The ℓ_1**-penalized logistic regression**:

$$\min_{\beta} \sum_{i=1}^{n} \log\left(1 + \exp(-Y_i X_i^\top \beta)\right) + \lambda||\beta||_1.$$

The statistical properties of the ℓ_1-penalized logistic regression can be found in van de Geer (2008). We now switch to another widely used classification model. In logistic model, we impose the model on $\mathbb{P}(Y|X)$. We can also turn to model $\mathbb{P}(X|Y)$. In specific, the **discriminative classification model** assumes the data (Y, X) is generated from the above model follows the mechanism:

1. Generate a Bernoulli random variable $Y \sim \text{Bernoulli}(\eta)$;
2. If $Y = +1$, generate $X \sim p_+(x)$, otherwise generate $X \sim p_-(x)$.

This model is also called the **mixture model** as it is a mixture of two distributions.

The **Gaussian discriminant model** further assumes that p_+ and p_- are Gaussian distributions. In specific, we denote

$$p_+(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_+)^\top \Sigma^{-1}(x-\mu_+)\right);$$
$$p_-(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_-)^\top \Sigma^{-1}(x-\mu_-)\right).$$

Given an input $X = x$, we can use the following Bayes rule to decide whether the corresponding Y is $+1$ or -1:

$$\text{Bayes rule: } h(x) = \begin{cases} +1 & \text{if } \mathbb{P}(Y=+1|X=x) \geq \mathbb{P}(Y=-1|X=x); \\ -1 & \text{otherwise.} \end{cases}$$

We can see that the Bayes rule classifies Y as $+1$ as long as the condition probability of $Y = +1$ with the prior information of $X = x$ is larger than the probability that $Y = -1$. For the discriminative classification model, we can get the closed forms of these two probabilities in the Bayes rule:

$$\mathbb{P}(Y=+1|X=x) = \frac{p_+(x)\eta}{p_+(x)\eta + p_-(x)(1-\eta)},$$
$$\mathbb{P}(Y=-1|X=x) = \frac{p_-(x)(1-\eta)}{p_+(x)\eta + p_-(x)(1-\eta)}.$$

Therefore, we can reformulate the Bayes rule as

$$\log\frac{\mathbb{P}(Y=+1|X=x)}{\mathbb{P}(Y=-1|X=x)} = \log\frac{p_+(x)}{p_-(x)} + \log\frac{\eta}{1-\eta} \geq 0.$$

Plugging in the Gaussian densities, under the balanced setting, i.e., $\eta = 1/2$, the Bayes rule becomes

$$f(x) = \underbrace{(\mu_+ - \mu_-)^\top \Sigma^{-1}}_{\beta^{*\top}}\Big(x - \underbrace{\frac{\mu_+ + \mu_-}{2}}_{\bar{\mu}}\Big) \geq 0,$$

where we can see that the decision rule is linear. Therefore, we call this method as **linear discriminant analysis** (LDA). When $\eta \neq 1/2$, it is easy to find that the decision boundary is also linear.

Suppose $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d. from the Gaussian discriminant model. In order to get the decision boundary $f(x) = \beta^{*\top}(x - \bar{\mu})$, we need to estimate

$$\beta^* = \Sigma^{-1}(\mu_+ - \mu_-) \text{ and } \bar{\mu} = \frac{\mu_+ + \mu_-}{2}.$$

In high dimensional LDA, we assume that β^* is sparse as the parsimonious principle implies that only a few variables among d covariates will be involved in the decision rule. We start with estimating the mean and covariance matrices as

$$n_+ = \textstyle\sum_{i=1}^n \mathbb{I}(Y_i = +1), \qquad n_- = \sum_{i=1}^n \mathbb{I}(Y_i = -1);$$
$$\widehat{\mu}_+ = \frac{1}{n_+}\sum_{i:Y_i=+1} X_i, \qquad \widehat{\mu}_- = \frac{1}{n_-}\sum_{i:Y_i=-1} X_i;$$
$$\widehat{\Sigma}_+ = \frac{1}{n_+}\sum_{i:Y_i=+1}(X_i - \widehat{\mu}_+)(X_i - \widehat{\mu}_+)^\top, \quad \widehat{\Sigma}_- = \frac{1}{n_-}\sum_{i:Y_i=-1}(X_i - \widehat{\mu}_-)(X_i - \widehat{\mu}_-)^\top.$$

We can then estimate $\bar{\mu}$ and the covariance matrix estimator as

$$\widehat{\bar{\mu}} = \frac{\widehat{\mu}_+ + \widehat{\mu}_-}{2} \text{ and } \widehat{\Sigma} = \frac{n_+\widehat{\Sigma}_+ + n_-\widehat{\Sigma}_-}{n_+ + n_-}.$$

Notice that Σ^{-1} is involved in the definition of β^*. We know that in the high dimensional setting, $\widehat{\Sigma}$ is not invertible. However, we observe that $\Sigma\beta^* = \mu_+ - \mu_-$. Therefore, we would expect $\widehat{\Sigma}\beta^*$ is close to $\widehat{\mu}_+ - \widehat{\mu}_-$. This motivates us to consider the following high dimensional LDA.

The **high dimensional linear discriminant analysis**:

$$\widehat{\beta}^{\text{LDA}} = \arg\min_{\beta} \|\beta\|_1, \text{ s.t. } \|\widehat{\Sigma}\beta - (\widehat{\mu}_+ - \widehat{\mu}_-)\|_\infty \leq \lambda,$$

and the decision rule is $f(x) = (x - \widehat{\bar{\mu}})^\top \widehat{\beta}^{\text{LDA}} \geq 0$. The problem above can be solved via linear programming.

The statistical properties of the high dimensional LDA is studied by Cai and Liu (2011).

9.2.2 High Dimensional Graphical Model

Networks are widely used as an interpretable way to visualize the data. An undirected graph $G = (V, E)$ has the vertex set V and edge set E. We denote the number of nodes $d = |V|$. We propose the following graphical model connecting the

9.1 Definition (Graphical Model). *A random vector $X \in \mathbb{R}^d$ is Markov with respect to a graph $G = (V, E)$ if*

$$(j, k) \notin E \Leftrightarrow X_j \perp\!\!\!\perp X_k | \text{ other } X_i\text{'s} .$$

By the definition, the node j is connected to node k if X_j are dependent to X_k conditioning on the rest $\{X_\ell | \ell \neq j, k\}$. The graphical model is widely used in real applications. In practice, X could be gene expressions, brain imaging, or social media. The graph G corresponding to X then becomes the genomic network, brain network, or social network. Comparing to the unconditional dependency like correlation, the conditional dependency can characterize direct connections between variables.

For the normal distributions, the following proposition implies that we can characterize the graphical model directly with the inverse covariance matrix.

9.2 Proposition (Gaussian graphical model). *If $X \sim N(0, \Sigma)$, then X is markov to $G = (V, E)$ if*

$$(j, k) \in E \Leftrightarrow (\Sigma^{-1})_{jk} \neq 0.$$

This proposition says the edges of Gaussian graphical model is determined by the inverse covariance matrix $\Theta = \Sigma^{-1}$, which is also called the precision matrix.

A sparse graphical model assumes that the graph G is sparse, i.e., the number of edges are much smaller than the number of nodes d. For the Gaussian graphical model, this is to assume that the precision matrix is sparse. Let $X_1, \ldots, X_n$ be i.i.d. samples from $N(0, \Sigma)$. Under the high dimensional setting, we cannot estimate the precision matrix via directly inverting the sample covariance matrix $\widehat{\Sigma}$ as it is not invertible. However, we can combine the negative log-likelihood of $N(0, \Sigma)$ with the ℓ_1-penalty.

The **graphical Lasso** (GLasso) estimator is

$$\min_{\Theta \succeq 0} \operatorname{Tr}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda ||\Theta||_{1,1},$$

where the matrix ℓ_1-norm is $||\Theta||_{1,1} = \sum_{j \neq k} |\Theta_{jk}|$.

More discussion on the GLasso estimator can be found in Friedman et al. (2008).

9.3 Beyond the ℓ_1-penalty

Sparsity plays a vital role in the high dimensional statistics. We know that the ℓ_1-norm $||\beta||_1$ encourages sparsity of β. However, we may need to impose other structural assumptions on the parameter β. Then we need to consider other penalty terms on β.

9.3.1 Group Lasso

The Lasso estimator can be used to select variables from the high dimensional features. Sometimes the features are in groups, and we want to select variables in groups.

Suppose $\beta \in \mathbb{R}^d$ has J groups. We denote each group as $S_j \subset \{1, \ldots, d\}$ for $j = 1, \ldots, J$. Therefore, we want to select subvectors $\beta_{S_1}, \ldots, \beta_{S_J}$.

For example, we want to predict the number of COVID cases tomorrow. The covariates to predict Y are groups: (1) the group of features related to the number of cases in the past: the number of case today, the number of case yesterday, the number of case in the past month, etc; (2) the group of features related to the weather: temperature, precipitation, and so on; (3) the group of features related to the social distancing: the number of people working from home, the number of restaurants opened, and so on; (4) the group of features related to Trump's activities: the number of Trump's tweets, the number of days Trump recovered from COVID, and so on. We may expect that the the number of COVID cases are related to some group of these variables.

If β is sparse in groups, we have the vector $(||\beta_{S_1}||_2, ||\beta_{S_2}||_2, \ldots, ||\beta_{S_J}||_2)^\top \in \mathbb{R}^J$ is

sparse. Therefore, we consider the group Lasso penalty as

$$\left\|(||\beta_{S_1}||_2, ||\beta_{S_2}||_2, \ldots, ||\beta_{S_J}||_2)\right\|_1 = \sum_{j=1}^{J} \|\beta_{S_j}\|_2.$$

Therefore, the group Lasso estimator is formulated as follows.

The **group Lasso** estimator

$$\min_{\beta} \left\|Y - \sum_{j=1}^{J} \mathbb{X}_{S_j} \beta_{S_j}\right\|_2^2 + \lambda \sum_{j=1}^{J} ||\beta_{S_j}||_2.$$

You can find the visualization of the group Lasso ball in Figure 9.1. More details on the group Lasso can be found in Yuan and Lin (2006).

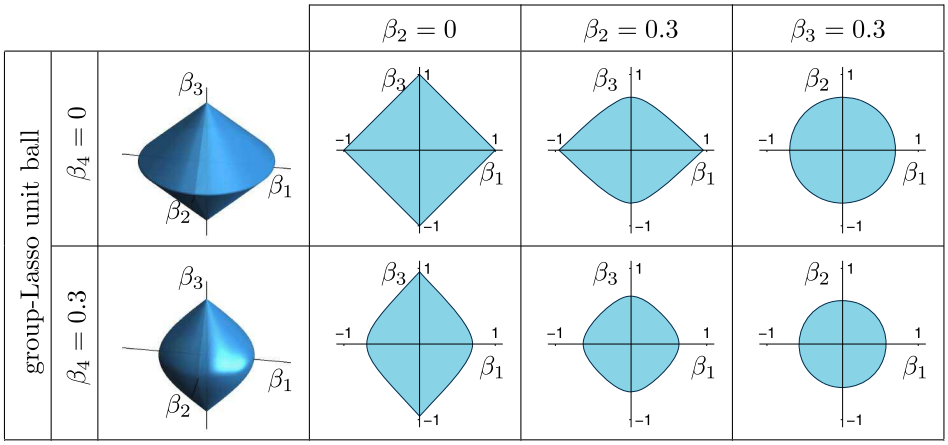


Figure 9.1. *The unit ball of the group Lasso penalty* $\sqrt{\beta_1^2 + \beta_2^2} + \sqrt{\beta_3^2 + \beta_4^2} \leq 1$. *Credit to Chiquet et al. (2012).*

An important application of the group Lasso is to the **sparse additive model** (SPAM):

$$Y_i = \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i, \text{ for } i = 1, \ldots, n,$$

where only s of these univariate functions f_j's are non-zero. To estimate f_j's, we expand the function in basis:

$$f_j(x_j) = \sum_{k=1}^{\infty} \beta_{jk}^* \phi_k(x_j), \text{ for } j = 1, \ldots, d,$$

where $\{\phi_k\}_{k=1}^{\infty}$ are basis functions, e.g., the polynomial bases $\{x^k\}_{k=1}^{\infty}$, the trigonometric bases $\{\sin(kx), \cos(kx)\}_{k=1}^{\infty}$, the B-splines, etc.

Therefore, if we want to select the nonparametric functions f_j's, it is equivalent to select the basis coefficients $\{\beta_{jk}^*\}_{k=1}^{\infty}$ for $j = 1, \ldots, d$ in group. This motivates us to consider the sparse additive model estimator with group Lasso penalty as follows.

The **spare additive model estimator**:

$$\min_{\beta_{jk}} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{d} \sum_{k=1}^{m} \beta_{jk} \phi_k(X_{ij}) \Big)^2 + \lambda \sum_{j=1}^{d} \Big(\sum_{k=1}^{m} \beta_{jk}^2 \Big)^{\frac{1}{2}},$$

where m is the number of basis functions we choose to approximate the true function.

More details of the sparse additive model can be found in Ravikumar et al. (2009).

9.3.2 Fused Lasso

In many applications, the parameters β are not sparse but piece-wise constant. For example, for the genomic data in Figure 9.2, the expression levels of neighbor genes are the same.

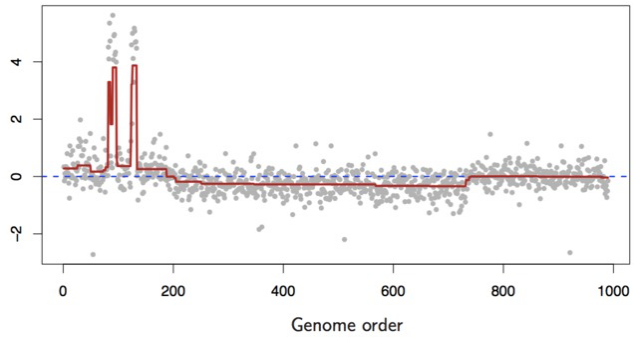


Figure 9.2. *The genomic sequencing data (grey dots) and the truth (red lines). We can see that the signal is a mixture of spiked and piece-wise constant signals.*

Although β is not spares, but if β is piece-wise constant, then the differences $(\beta_1 - \beta_2, \beta_2 - \beta_3, \ldots, \beta_{d-1} - \beta_d)^\top = D\beta$, where D is the differential map

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(d-1)\times d},$$

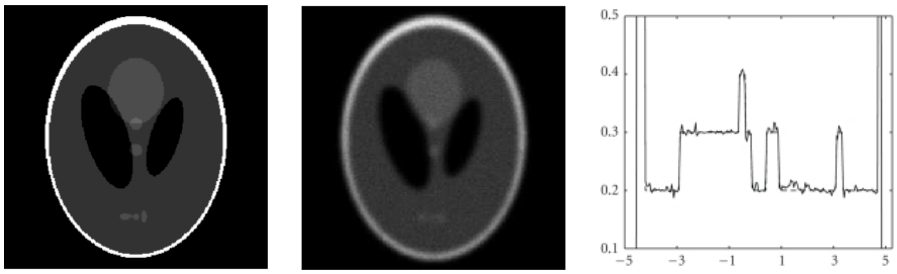
is sparse. Therefore, instead of penalizing $\|\beta\|_1$, we should penalize the ℓ_1-norm of $D\beta$. This gives us the Fused Lasso estimator.

The **fused Lasso** estimator:

$$\min_{\beta} ||Y - \mathbb{X}\beta||_2^2 + \lambda ||D\beta||_1.$$

We refer to Tibshirani et al. (2005) for more details. We can also apply the fused Lasso to 2D image data. For example, in Figure 9.3, β^* is brain imaging data and Y is a noisy observation. We can see that the β is a "patched" image: it is a 2-dimensional piece-wise constant. Therefore, we can apply the differential map D to the rows and columns of β and expect that both of them are sparse. Then we can recover the truth from the noisy observation via

$$\min_{\beta} \sum_{jk} |Y_{jk} - \beta_{jk}|^2 + \lambda ||D\beta||_1 + \lambda ||D\beta^\top||_1.$$



(a) Truth (b) Noisy (c) Sliced Signal

Figure 9.3. *Shepp-Logan phantom image. Panel (a) is the original signal, panel (b) is the noisy image, and panel (c) is one row of the image.*

9.4 Beyond the Biasedness

Now we go back to the linear model $Y = \mathbb{X}\beta^* + \varepsilon$ with the sparsity assumption. Even if β is sparse, the ℓ_1-norm may not be perfect. Ideally, we want to regularize the ℓ_0-norm of β, but we know that the computation of ℓ_0-norm is challenging. That is why we introduce the convex relaxation of the ℓ_0-norm, which is ℓ_1-norm. Despite its advantage in computation, from Figure 9.4(a), we can see that the larger β_j is, the more the ℓ_1-norm penalizes it, while the ℓ_0-norm does not. Therefore, Lasso is biased for large β_j's.

A straightforward approach to fix this problem is through refitting. Let $\widehat{S}$ be the support of $\widehat{\beta}^{\text{Lasso}}$ and we can refit the linear model on $\widehat{S}$ via least squares. Denote the refitted least squares estimator as

$$\widehat{\beta}^{\text{LS}}_{\widehat{S}} = (\mathbb{X}_{\widehat{S}}^\top \mathbb{X}_{\widehat{S}})^\dagger \mathbb{X}_{\widehat{S}}^\top Y \text{ and } \widehat{\beta} = (\widehat{\beta}_{\widehat{S}}^\top, \widehat{\beta}_{\widehat{S}^c}^\top)^\top, \text{ where } \widehat{\beta}_{\widehat{S}} = \widehat{\beta}^{\text{LS}}_{\widehat{S}}, \widehat{\beta}_{\widehat{S}^c} = 0.$$

Since the least square is unbiased, so the refitted estimator should remove the bias of Lasso estimator. We refer to Meinshausen (2007) for more details.

Besides the refitted estimator, Zou (2006) proposed the adaptive Lasso to fix the bias. The idea of the adaptive Lasso is that if β_j^* is large, we should penalize less on β_j. Let $\widehat{\beta}^{\text{ini}}$ be some initial estimator of β^*, which could be Lasso estimator or others.

The **adaptive Lasso** estimator is:

$$\min_{\beta} ||Y - \mathbb{X}\beta||_2^2 + \lambda \sum_{j=1}^{d} \frac{|\beta_j|}{|\widehat{\beta}_j^{\text{ini}}|^\gamma}, \text{ where } \gamma > 0.$$

We can see that the adaptive Lasso has a weighted ℓ_1-norm as the penalty. The larger $|\widehat{\beta}_j^{\text{ini}}|$ is, the weight on $|\beta_j|$ is smaller and thus the bias is smaller.

9.4.1 Nonconvex Penalties

The previous two approaches fixing the bias of Lasso depend on some initial estimator. In fact, we can formulate a one-step estimator. From Figure 9.4(a), the ℓ_1-norm is convex but biased while ℓ_0 is discontinuous but unbiased. As a trade-off, we can consider a penalty which is between ℓ_1 and ℓ_0. See the red curve in Figure 9.4(a). The SCAD penalty proposed by Fan and Li (2001) is this kind of penalty.

The SCAD penalty is

$$p_\lambda^{\text{SCAD}}(t) = \int_0^t \Big(\mathbb{I}\{t \le \lambda\} + \frac{a\lambda - x}{(a-1)\lambda} \mathbb{I}\{t > \lambda\} \Big) dx,$$

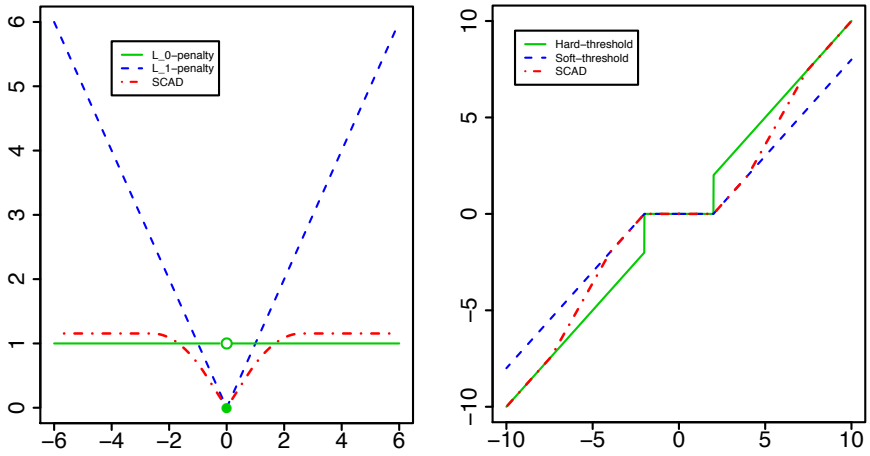
and the SCAD estimator is

$$\widehat{\beta}^{\text{SCAD}} = \arg\min_{\beta} ||Y - \mathbb{X}\beta||_2^2 + 2 \sum_{j=1}^{d} p_\lambda^{\text{SCAD}}(\beta_j).$$

When $\mathbb{X}$ is identity, the penalized least square $\min_\beta \|Y - \beta\|_2^2 + p_\lambda(\beta)$ has closed form solution for the penalty p_λ to be ℓ_0, ℓ_1, or SCAD. We visualize the solution in Figure 9.4(b) and we can see that the Lasso (blue line) gives us biased estimation while the ℓ_0 (green line) and SCAD (red line) are unbiased for large signals.

We can see that SCAD penalty is nonconvex. However, there is efficient algorithms to solve SCAD estimator. We refer to Loh and Wainwright (2015) and Wang et al. (2014) for details. Denote two index sets

- Strong signals: $S_1 = \left\{ j \in S | |\beta_j^*| \ge C\sqrt{\log d/n} \right\}$ for some sufficiently large constant C;



(a) ℓ_0, ℓ_1, and SCAD penalties (b) Solutions for identity design

Figure 9.4. *The penalties for p_λ being ℓ_0, ℓ_1, and SCAD and the solution to* $\min_\beta \|Y-\beta\|_2^2+p_\lambda(\beta)$.

- Weak signals: $S_2 = S \backslash S_1$.

It can be shown (Loh and Wainwright, 2015; Wang et al., 2014) that, if we choose the tuning parameter properly, the SCAD estimator has the statistical rate

$$||\widehat{\beta}^{\text{SCAD}} - \beta^*||_2 = O_P\left(\sqrt{\frac{|S_1|}{n}} + \sqrt{\frac{|S_2|\log d}{n}}\right).$$

In comparison, the Lasso estimator has the rate (Corollary 8.6 in Lecture 8)

$$||\widehat{\beta}^{\text{Lasso}} - \beta^*||_2 = O_P\left(\sqrt{\frac{(|S_1|+|S_2|)\log d}{n}}\right).$$

As we discussed in the previous lecture, the $\sqrt{\log d}$ term in the rate can be treated as a price to select s variables from the d features. The statistical rate of SCAD estimator shows that SCAD does not need to pay the $\sqrt{\log d}$ price to detect strong signals, which illustrates how SCAD does not have bias for strong enough signals in theory.

9.5 Beyond Tuning Sensitive

Recall that in Corollary 8.6 in Lecture 8, if the noise ε are independent sub-Gaussians with variance proxy σ^2, we suggest to choose the tuning parameter of Lasso as $\lambda = C\sigma\sqrt{\log d/n}$ for some sufficiently large constant C. However, in practice, we do not know the variance proxy σ^2. Although we can always use cross-validation in practice, the tuning procedure will be sensitive. We will introduce a tuning insensitive variations of Lasso.

9.5.1 Square-Root Lasso

Intuitively, the reason why Lasso is tuning sensitive of the parameter of ε is that the gradient of the least square loss $\mathcal{L}_n(\beta) = \|Y - \mathbb{X}\beta\|_2^2$ at the truth is $\nabla\mathcal{L}_n(\beta^*) = \mathbb{X}^\top\varepsilon$, which is related to ε. This is why in Theorem 8.4 of Lecture 8, we suggest to choose $\lambda \geq \frac{2}{n}\|\mathbb{X}^\top\varepsilon\|_\infty$. If we can design some loss whose gradient is irrelavant to ε, then we can expect the estimator to be tuning insensitive. A candidate loss is the square-root loss $\mathcal{L}_n^{1/2}(\beta) = \|Y - \mathbb{X}\beta\|_2$ whose gradient at the truth is

$$\nabla\mathcal{L}_n^{1/2}(\beta^*) = \frac{\mathbb{X}^\top\varepsilon}{2\|Y - \mathbb{X}\beta^*\|_2} = \frac{\mathbb{X}^\top\varepsilon}{2\|\varepsilon\|_2}.$$

Although it is still related to ε, we can see that the gradient of the square-root loss is somehow normalized and thus insensitive to the magnitude of ε. This motivates us to consider the square-root Lasso.

The **square-root Lasso** estimator is

$$\widehat{\beta}^{\sqrt{\text{Lasso}}} = \arg\min_\beta \frac{1}{\sqrt{n}}||Y - \mathbb{X}\beta||_2 + \lambda\|\beta\|_1.$$

Under mild regularity conditions on the noise ε, it can be show that if we choose $\lambda = C\sqrt{\log d/n}$ for some universal constant C irrelevant to ε, the square-root Lasso has the rate $\|\widehat{\beta}^{\sqrt{\text{Lasso}}} - \beta^*||_2 = O_P(\sqrt{s\log d/n})$. We refer to Belloni et al. (2011a) for detailed discussion.

9.6 Beyond Sub-Gaussian

Recall that in Corollary 8.6 in Lecture 8, the Lasso has the rate $O_P(\sqrt{s\log d/n})$ if the noise ε are independent sub-Gaussians. We aim to generalize the estimator to heay-tailed noises. If ε is heavy-tailed, its the moment generating function may be infinite but the quantile of ϵ is always finite. This motive us to consider to utilize the property of quantiles.

9.6.1 Quantile Regression

We start with formulating the quantile as an optimization problem.

9.3 Proposition (Variational form of quantiles). *Let $\rho_\tau(y) = y(\tau - \mathbb{I}\{y < 0\})$. Then τ-quantile of a random variable $X = \arg\min_x \mathbb{E}[\rho_\tau(X - x)]$. Specifically, we have*

$$\textit{Median of } X = \arg\min_x \mathbb{E}|X - x|.$$

Proof. We denote $x_\tau = \arg\min_x \mathbb{E}[\rho_\tau(X - x)]$. Therefore, we have

$$0 = \frac{d}{dx}\mathbb{E}[\rho_\tau(X - x)]\Big|_{x=x_\tau} = \tau - \mathbb{E}[\mathbb{I}\{X - x_\tau < 0\}].$$

Thus $\mathbb{P}(X < x_\tau) = \tau$, i.e., x_τ is the τ-quantile of X. □

Therefore, to have a robust estimator for the linear model, we can replace the least squares loss in Lasso by the $\rho_\tau(y) = y(\tau - \mathbb{I}\{y < 0\})$.

The **high dimensional quantile regression** estimator is

$$\min_\beta \sum_{i=1}^n \rho_\tau(Y_i - X_i^\top \beta) + \lambda||\beta||_1.$$

When $\tau = 0.5$, it reduces to the **least absolute deviation estimator** (LAD).

$$\min_\beta \|Y - \mathbb{X}\beta\|_1 + \lambda||\beta||_1.$$

Since $\rho_\tau(\cdot)$ is piece-wise linear, so the quantile regression can be solved via linear programming.

We refer the further discussion on quantile regression to Belloni et al. (2011b).

Lecture 10

Convexity and Subgradient

10.1 Convex Optimization

From the previous lectures, we can see that many estimators can be formulated as an optimization problem. One example is the Lasso estimator

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{2n} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_1.$$

The Lasso estimator is an unconstrained problem as we find the minimizer over the entire $\mathbb{R}^d$. In comparison, the constrained Lasso

$$\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2, \text{ s.t. } \|\beta\|_1 \leq t,$$

has the constraint $\|\beta\|_1 \leq t$. More examples can be found in the examples of Lecture 9. Therefore, the next part of our class will focus on how to solve these optimization problems and how well these algorithms perform.

Under the setting of high dimensional statistics, both the sample size n and the feature dimension d are very large. We talk about how to handle this challenge in probability and statistics. The big data challenge is also crucial in optimization from two perspectives: storage and computation. Given a d-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$, in optimization, we will usually use its gradient $\nabla f \in \mathbb{R}^d$ and Hessian matrix $\nabla^2 f \in \mathbb{R}^{d \times d}$. In high dimensional optimization, we will try to only use the gradients and avoid using the Hessian matrix as it is much slower to conduct matrix operations than vector operations. Moreover, when d is ultra-large, we may not even afford to store the d by d matrix, not mentioning the matrix operations. For example, many genomics datasets have expression levels of millions of genes, then the size of Hessian matrix related to the question will be trillions.

The optimization algorithms only involving the gradients are called the **first order method**. Therefore, in our class, we will focus on the first order method. From the unconstrained and constrained Lasso above, we found that both the least squares loss $\|Y - \mathbb{X}\beta\|_2^2$ and the ℓ_1-ball $\|\beta\|_1 \leq t$ are convex. Convexity is a very important concept in optimization. It covers a lot of useful examples and will induce many nice properties.

Let us start with formally defining a convex set and convex function.

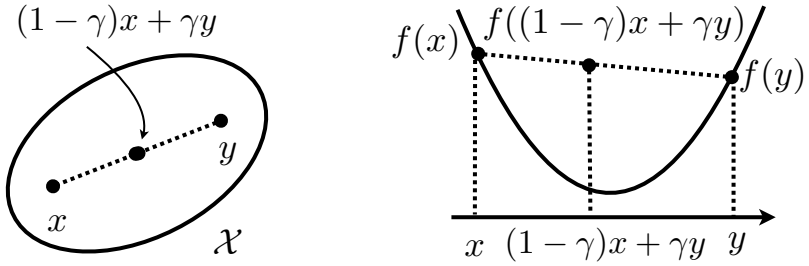
10.1 Definition (Convex set and function). *A set M is convex if*

$$(1-\gamma)x + \gamma y \in M, \text{ for any } x, y \in M, \gamma \in [0,1].$$

A function $f : M \to \mathbb{R}$ is convex if

$$f((1-\gamma)x + \gamma y) \leq (1-\gamma)f(x) + \gamma f(y), \text{ for any } x, y \in M, \gamma \in [0,1].$$

From Figure 10.1, we can see that a set M is convex if it contains all its segments and a function f is convex if it always lies below its chords.



(a) Convex set (b) Convex function

Figure 10.1. *A set M is convex if it contains all its segments and a function f is convex if it always lies below its chords.*

We next give a general form of optimization problems we are interested in our class.

10.2 Definition (Convex optimization). *An optimization problem is called convex optimization if it can be formulated as*

$$\begin{aligned} &\min_x f(x) \\ &\textit{s.t. } x \in M, \end{aligned}$$

where both f and M are convex.

We can see that both the unconstrained and constrain Lasso are convex optimization. We can check that all the examples given by Lecture 9 (expect SCAD) are convex optimizations as well.

10.2 Subgradient

As we like the first order method in high dimension optimization, we need to study the properties of gradients. However, many convex functions are not smooth and does not have

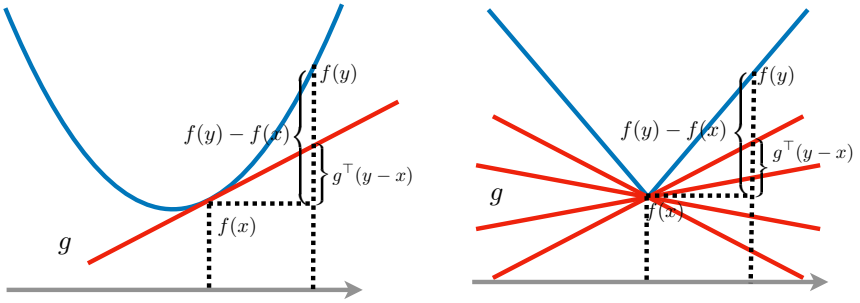
gradients. For example, the ℓ_1-norm $\|x\|_1$ is not differentiable at $x = 0$. Therefore, we will generalize the concept of gradient to non-smooth convex functions.

10.3 Definition (Subgradient). *Let $f : M \to \mathbb{R}$. We say $g \in \mathbb{R}^d$ is a subgradient of f at x if for any $y \in M$, we have*

$$f(y) - f(x) \geq g^\top (y - x).$$

We denote the set of all subgradients of f at x as $\partial f(x)$.

Figure 10.2 illustrates the geometric interpretation of the subgradient. The hyperplane tangent to f at x is $p(y) = g^\top (y - x) + f(x)$. We can see in Figure 10.2 that the inequality $f(y) - f(x) \geq g^\top (y - x)$ implies that f is above the hyperplane. So we also call $p(y)$ as a supporting hyperplane.



(a) Smooth function (b) Non-Smooth function

Figure 10.2. *The geometry of the subgradient. For smooth functions, the subgradient is unique but for non-smooth functions, the subgradient may be not unique.*

Notice that $\partial f(x)$ is a set. It can be shown (or seen from Figure 10.2) that when a function is convex, the subgradient must exist. When f is differentiable and convex, by Definition 10.1, we have

$$\begin{aligned} f(y) &\geq \frac{f((1-\gamma)x + \gamma y) - (1-\gamma)f(x)}{\gamma} \\ &= f(x) + \frac{f(x + \gamma(y - x)) - f(x)}{\gamma} \xrightarrow[\gamma \to 0]{} f(x) + \nabla f(x)^\top (y - x). \end{aligned}$$

So $\nabla f(x) \in \partial f(x)$. However, the following example shows that the subgradient may be non-unique.

10.4 Example (Subgradient of ℓ_1-norm). The most widely used nonsmooth convex function in our class is the ℓ_1-norm. We first start with the univariate case. Let $f(x) = |x|$ and we know f is differentiable when $x \neq 0$. Therefore, $\partial |x| = \operatorname{sign}(x)$ when $x \neq 0$. When $x = 0$, $|x|$ is non-smooth. From Figure 10.2(b), we can see that there are multiple lines which can "support" $|x|$. It is easy to check that for any $g \in [-1, 1]$, we have $|y| \geq gy$ for any y. Therefore, $\partial |x| = [-1, 1]$ when $x = 0$. We can generalize to the ℓ_1-norm $\|x\|_1$.

Following the same analysis of the univariate case, if $g \in \partial \|x\|_1$, $g_j = \text{sign}(x_j)$ for $x_j \neq 0$ and $g_j \in [-1, 1]$ for $x_j = 0$. □

We mentioned above that convexity will give us many nice properties. The following proposition is one of the reason why convex optimization is much easier than nonconvex optimization.

10.5 Proposition (Local minima are global minima). *Let f be convex. If x^* is local minimum of f, then x^* is its global minimum. This happens if and only if $0 \in \partial f(x^*)$.*

Proof. We say x^* is local minimum of f if for any y, there exists a small enough γ such that

$$f(x^*) \leq f((1-\gamma)x^* + \gamma y) \leq (1-\gamma)f(x^*) + \gamma f(y),$$

where the last inequality is due to the convexity of f. Therefore, $f(x^*) \leq f(y)$ and x^* is global minimum.

If x^* is global minimum, $f(y) - f(x^*) \geq 0$ for any y. So $0 \in \partial f(x^*)$ by definition. It is straightforward to check the reverse direction. □

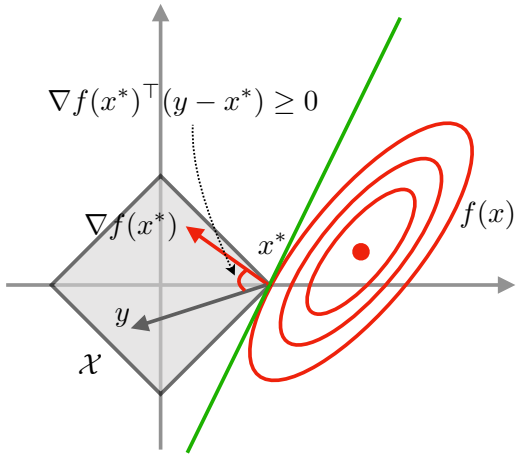


Figure 10.3. *The geometry of the first order optimality condition $\nabla f(x^*)^\top (y - x^*) \geq 0$, for all $y \in M$ for the problem $\min_x f(x)$ s.t. $x \in M$.*

If x^* is global minimum of f if and only if $0 \in \partial f(x^*)$. We also call $0 \in \partial f(x^*)$ as the **first order optimality condition**. Comparing to the **zero order optimality condition** $f(x^*) \leq f(y)$ for any y, the first order optimality condition is called first as it involves the derivatives. Notice that $0 \in \partial f(x^*)$ is the first order optimality condition only for the unconstrained optimization. The following proposition gives us the optimality condition for unconstrained problem.

10.6 Proposition (First order optimality condition). *Given a convex set M and a convex*

differentiable function $f : M \to \mathbb{R}$, $x^* \in \arg\min_x f(x)$ s.t. $x \in X$ if and only if

$$\nabla f(x^*)^\top (y - x^*) \geq 0, \text{ for all } y \in M.$$

Before we prove the proposition, we illustrate the geometry of the first order optimality condition in Figure 10.3. We use the constrained Lasso as an example. We can see that the first order optimality condition essentially implies that the constrain M is on the one side of the hyperplane $p(y) = \nabla f(x^*)^\top (y - x^*)$. Specifically, $\nabla f(x^*)^\top (y - x^*) \geq 0$ implies that the angle between the gradient $\nabla f(x^*)$ and $y - x^*$ is always acute.

The optimality conditions for the convex optimization $x^* \in \arg\min_x f(x)$ s.t. $x \in M$: for any $y \in M$ (when the problem is uncontrained $M = \mathbb{R}^d$), we have

	Unconstrained	Constrained
Zero order	$f(x^*) \leq f(y)$	$f(x^*) \leq f(y)$
First order	$0 \in \partial f(x^*)$	$\nabla f(x^*)^\top (y - x^*) \geq 0$

Proof. We first prove the "if" part. By the definition of subgradient, we have

$$f(y) - f(x^*) \geq \nabla f(x^*)^\top (y - x^*) \geq 0$$

Therefore, we have $f(y) \geq f(x^*)$ and thus x^* is minimizer.

To prove the "only if" part, it suffices to show that if there exists a $y \in M$ such that $\nabla f(x^*)^\top (x^* - y) > 0$, then x^* is not a minimum. Define $h(t) = f(x^* + t(y - x^*))$. We have $h'(0) = \nabla f(x^*)^\top (y - x^*) < 0$

$$h'(0) = \nabla f(x^*)^\top (y - x^*) < 0$$

So $h(t)$ is decreasing at 0, i.e., $f(x)$ will decrease along the direction of $y - x^*$. Therefore, x^* is not a minimizer. □

Lecture 11

Gradient Descent

11.1 Gradient Descent

In this lecture, we will start designing algorithms to solve the convex optimization

$$\min_{x \in \mathcal{X}} f(x), \text{ where } f \text{ and } M \text{ are convex.}$$

Our goal is to find the minimizer $x^* = \arg\min_{x \in M} f(x)$. Let us start with the unconstrained problem first with $M = \mathbb{R}^d$. If we start our search for x^* at some value x_0, we aim to move to the next point such that the value of $f(x)$ becomes smaller. If we go along the direction d with an infinitesimal step size τ, the decrease of the objective function is

$$\lim_{\tau \to 0} \frac{f(x_0 + \tau d) - f(x_0)}{\tau} = \nabla f(x_0)^\top d. \tag{11.1}$$

We hope to choose the direction d such the the decrease is maximized, and from the result above, we can see that the optimal direction is $d = -\nabla f(x_0)$. Therefore, if we can only utilize the local information of $f(x)$ at the point x_0, the negative gradient $-\nabla f(x_0)$ is the steepest descent direction. This simple strategy motivates the gradient descent algorithm.

The **gradient descent algorithm** is an iterative algorithm starting at the point $x_0 \in \mathbb{R}^d$ and it iterates as

$$x_{t+1} = x_t - \eta_t \nabla f(x_t),$$

where η_t is the step-size (also called learning rate).

In order to guarantee good convergence properties of the gradient descent, the landscape of $f(x)$ cannot be too steep or too rough. We formalize this intuition by the following definition of L-smooth.

11.2 Definition (L-smooth). *A continuously differentiable function f is L-smooth if*

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|x - y\|_2^2,$$

for any x and y.

We can see that $f(x) + \nabla f(x)^\top (y - x)$ is the first order Taylor expansion of $f(y)$. So the smoothness of f implies that the difference between the function value and its first order Taylor can be controlled by the quadratic function $\frac{L}{2}\|y - x\|_2^2$. If f has the second derivative, the 2nd order Taylor expansion implies

$$f(y) \approx f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x). \tag{11.3}$$

Compare it with Definition 11.2, the L-smooth implies that the Hessian $\nabla^2 f(x)$ cannot be too large. In fact, the L-smooth is equivalent to following properties.

- If f has the second derivatives, (11.3) is equivalent to $\nabla^2 f(x) \preceq L\mathbf{I}_d$ for any x, which prevents f from being too curved.
- (11.3) is equivalent to $\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2$, i.e., the gradient of f is Lipschitz. So L-smoothness implies that the gradient of f does not change too drastically.

For example, the least squares loss $f(\beta) = \frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2$ has the Hessian matrix $\nabla^2 f(x) = \mathbb{X}\mathbb{X}^\top$. So the least squares loss $\frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2$ is L-smooth if the maximum eigenvalue of $\mathbb{X}\mathbb{X}^\top$ is smaller than L. On the other hand, the ℓ_1-norm $\|x\|_1$ is not smooth as its derivative is not continuous (not mentioning L-Lipschitz) at $x = 0$.

The following theorem shows that how fast the gradient descent converges.

11.4 Theorem (Convergence of gradient descent). *Let f be convex and L-smooth. If we choose $\eta_t = 1/L$ and the gradient descent has*

$$f(x_t) - f(x^*) \le \frac{2L\|x_0 - x^*\|_2^2}{t}.$$

We say the convergence rate of gradient is $O(1/t)$, or to rephrase, the gradient descent can achieve the ϵ-accuracy $f(x_t) - f(x^*) \le \varepsilon$ within $O(1/\epsilon)$ steps. We omit the proof in the class but we will list the proof here.

Proof. By the smoothness of $f(x)$ and the gradient descent iteration $x_{t+1} = x_t - \eta_t \nabla f(x_t)$, we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &= -\eta_t \|\nabla f(x_t)\|_2 + \frac{\eta_t^2 L}{2}\|\nabla f(x_t)\|_2^2. \end{aligned}$$

In order to maximize the descent, we choose $\eta_t = 1/L$ to maximize the right hand side of the inequality and get

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L}\|\nabla f(x_t)\|_2^2. \tag{11.5}$$

This inequality shows that the objective value of the gradient descent is monotonic decreasing.

In fact, not only the objective value is decreasing, we can also show that the distance between x_t and x^* is also decreasing. To prove this, we need the following lemma derived from the L-smoothness.

11.6 Lemma. *f is L-smooth. Then for any $x, y \in \mathbb{R}^n$, one has*

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Proof. Let $z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x))$. Then one has

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \leq \nabla f(x)^\top (x - z) + \nabla f(y)^\top (z - y) + \frac{L}{2}\|z - y\|_2^2 \\ &= \nabla f(x)^\top (x - y) + (\nabla f(x) - \nabla f(y))^\top (y - z) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\ &= \nabla f(x)^\top (x - y) - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2. \end{aligned}$$

□

Now we show that $\|x_t - x^*\|_2$ is decreasing with t. Using Lemma 11.6 one immediately gets

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

We use this inequality by setting $x = x_{t+1}, y = x^*$ as follows (together with $\nabla f(x^*) = 0$)

$$\begin{aligned} \|x_{t+1} - x^*\|_2^2 &= \left\| x_t - \frac{1}{L}\nabla f(x_t) - x^* \right\|_2^2 \\ &= \|x_t - x^*\|_2^2 - \frac{2}{L}\nabla f(x_t)^\top (x_t - x^*) + \frac{1}{L^2}\|\nabla f(x_t)\|_2^2 \\ &\leq \|x_t - x^*\|_2^2 . \end{aligned}$$

Denoting $\delta_t = f(x_t) - f(x^*)$, we can rewrite (11.5) as

$$\delta_{t+1} \leq \delta_t - \frac{1}{2L}\|\nabla f(x_t)\|_2^2 .$$

One also has by convexity (Definition 10.3),

$$\delta_t \leq \nabla f(x_t)^\top (x_t - x^*) \leq \|x_t - x^*\|_2 \|\nabla f(x_t)\|_2 .$$

As $\|x_t - x^*\|_2$ is decreasing with t, which with the two above displays will imply

$$\delta_{t+1} \leq \delta_t - \frac{1}{2L\|x_0 - x^*\|_2^2}\delta_t^2.$$

Let us see how to use this last inequality to conclude the proof. Let $\omega = \frac{1}{2L\|x_1 - x^*\|_2^2}$, then

$$\omega\delta_t^2 + \delta_{t+1} \leq \delta_t \Leftrightarrow \omega\frac{\delta_t}{\delta_{t+1}} + \frac{1}{\delta_t} \leq \frac{1}{\delta_{t+1}} \Rightarrow \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \geq \omega \Rightarrow \frac{1}{\delta_t} \geq \omega(t-1).$$

□

11.2 Frank-Wolfe Algorithm

We now consider the constrained problem $\min_{x\in M} f(x)$. The gradient descent algorithm might make x_t escape the constraint $\mathcal{X}$. As illustrated in Figure 11.1, we still want to find the steepest descent d but we also need to make sure that x_t belongs to M for all $t = 0, 1, 2, \ldots$. Therefore, similar to (11.1), starting with $x_0 \in M$, we want to find the steepest descent $d = x - x_0$ with $x \in M$ via the following steepest descent

$$\arg\min_{x\in\mathcal{X}} \langle \nabla f(x_0), x - x_0\rangle = \arg\min_{x\in\mathcal{X}} \langle \nabla f(x_0), x\rangle.$$

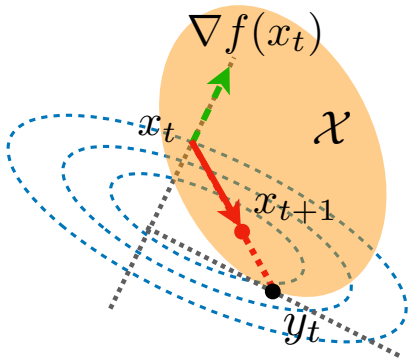


Figure 11.1. *The illustration of the Frank-Wolfe algorithm.*

This motivates the following Frank-Wolfe algorithm.

The **Frank-Wolfe algorithm** performs the update

$$y_t = \arg\min_{x\in\mathcal{X}} \langle \nabla f(x_t), x\rangle;$$
$$x_{t+1} = x_t + \eta_t(y_t - x_t),$$

where η_t is the step-size.

11.7 Example (Power iteration). We know that the leading eigenvector of a positive semi-definite matrix A is the solution of the following optimization problem (see Theorem 1.18 in Lecture 1)

$$\max_{\|x\|_2\le 1} x^\top Ax = \min_{\|x\|_2\le 1} -x^\top Ax.$$

Notice that $f(x) = -x^\top Ax$ is concave as $A \succeq 0$. So the above problem is not convex. However, we can still use the Frank-Wolfe algorithm to solve it and find the leading eigenvector of A. Implementing the Frank-Wolfe algorithm, we need to solve

$$y_t = \arg\min_{\|x\|_2\le 1} \langle \nabla f(x_t), x\rangle = -\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|_2} = \frac{Ax_t}{\|Ax_t\|_2},$$

where we use the variational form of ℓ_2-norm $\arg\max_{\|x\|_2\le 1}\langle y, x\rangle = y/\|y\|_2$. This gives us the updating rule

$$x_{t+1} = x_t + \eta_t\Big(\frac{Ax_t}{\|Ax_t\|_2} - x_t\Big) = (1-\eta_t)x_t + \eta_t\frac{Ax_t}{\|Ax_t\|_2}.$$

We want to choose the step-size $\eta_t \in [0,1]$ to minimize $f(x_{t+1}) = -x_{t+1}^\top Ax_{t+1}$. It can be proved that the optimal $\eta_t = 1$ and thus $x_{t+1} = \frac{Ax_t}{\|Ax_t\|_2}$, which is called **power iteration** to solve the leading eigenvector of A. □

11.8 Example (Constrained Lasso). We can see that the following two constrained Lasso are equivalent:

$$\begin{cases} \min\limits_{\beta} \|Y - \mathbb{X}\beta\|_2^2 \\ \text{s.t. } \|\beta\|_1 \le \lambda, \end{cases} \iff \begin{cases} \min\limits_{\beta} \|Y/\lambda - \mathbb{X}\beta\|_2^2 \\ \text{s.t. } \|\beta\|_1 \le 1. \end{cases}$$

Therefore, without loss of generality, we can assume $\lambda = 1$ in the constrained Lasso. To solve it by Frank-Wolfe algorithm, we have

$$y_t = \arg\min_{\|\beta\|_1\le 1} \langle \nabla f(\beta_t), \beta\rangle,$$

where $\nabla f(\beta_t) = \nabla_\beta \|Y - \mathbb{X}\beta_t\|_2^2 = -2\mathbb{X}^\top(Y - \mathbb{X}\beta_t)$. By the variational form of ℓ_1-norm (see Proposition 6.3 in Lecture 6), we know the dual norm of ℓ_1-norm is $\|x\|_\infty = \max_{\|y\|_1\le 1}\langle y, x\rangle$ and

$$\arg\max_{\|y\|_1\le 1} \langle y, x\rangle = (0,\dots,0,\text{sign}\,(x_{j^*}),0,\dots,0)^\top, \text{ where } j^* = \arg\max_{1\le j\le d} |x_j|.$$

Plugging this back into FW algorithm, we get

$$(y_t)_j = \begin{cases} -\text{sign}\,(\nabla_j f(\beta_t)), & \text{if } j = \arg\max_k \nabla_k f(\beta_t); \\ 0, & \text{otherwise.} \end{cases}$$
$$\beta_{t+1} = (1-\eta_t)\beta_t + \eta_t y_t.$$

Notice that only one entry of y_t is nonzero, therefore, in each iteration, β_{t+1} at most has one more nonzero entry than x_t. So if we start at $\beta_0 = 0$, we have $\|x_t\|_0 \le t$. This implies that the algorithm is very fast, and also provides further intuition as to why the ℓ_1 constraint implies sparsity in the optimal Lasso solution. □

The following theorem shows how to choose the step-size in Frank-Wolfe and how fast it converges.

11.9 Theorem (Convergence rate of Frank-Wolfe algorithm). *Let f be convex and L-smooth. If we choose $\eta_t = \frac{2}{t+2}$, then the Frank-Wolfe algorithmhas*

$$f(x_t) - f(x^*) \le \frac{2Ld_{\mathcal{X}}^2}{t+2},$$

where $d_{\mathcal{X}}^2 = \sup_{x,y\in\mathcal{X}} \|x-y\|_2^2$.

Similar to the converge rate of gradient descent in Theorem 11.4, for convex and smooth function, the Frank-Wolfe algorithm has the convergence rate $O(1/t)$, or to rephrase, we can achieve the ϵ-accuracy within $O(1/\epsilon)$ steps. However, unlike the constant step-size for the gradient descent, the suggested step-size of the Frank-Wolfe algorithm is shrinking and is irrelevant to L.

Proof. By smoothness, we have

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\le \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &\le \eta_t \nabla f(x_t)^\top (y_t - x_t) + \frac{L}{2}\eta_t^2 \|y_t - x_t\|_2^2 && (\text{By } x_{t+1} - x_t = \eta_t(y_t - x_t)) \\ &\le \eta_t \nabla f(x_t)^\top (x^* - x_t) + \frac{L}{2}\eta_t^2 d_{\mathcal{X}}^2 && (\text{By } y_t = \arg\min_{x\in\mathcal{X}} \nabla f(x_t)^\top x) \\ &\le \eta_t (f(x^*) - f(x_t)) + \frac{L}{2}\eta_t^2 d_{\mathcal{X}}^2 && (\text{By convexity, Definition 10.3}) \end{aligned}$$

Now let $\Delta_t = f(x_t) - f(x^*)$, we get

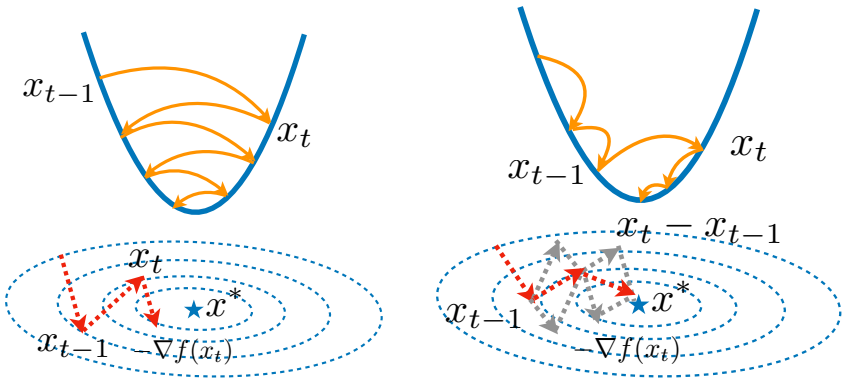
$$\Delta_{t+1} - \Delta_t \le -\eta_t \Delta_t + \frac{L}{2}\eta_t^2 d_{\mathcal{X}}^2 \implies \Delta_{t+1} \le (1-\eta_t)\Delta_t + \frac{L}{2}\eta_t^2 d_{\mathcal{X}}^2.$$

We set $\eta_t = 2/(t+2)$, and it follows from induction that $\Delta_t \le \frac{2L}{t+2} d_{\mathcal{X}}^2$. □

11.3 Accelerated Gradient Descent

From Figure 11.2(a), we notice that the trajectory of gradient descent is zigzag. This is due to the steep descent direction in (11.1) only depends on the local information of the objective function and thus is short-sighted. Can we find an algorithm converges faster than the gradient descent? We may consider the following two strategies to remedy the problems of the gradient descent we mentioned above:

- Exploit the history of the trajectory;
- Add buffers to have a smoother trajectory.



(a) Gradient Descent (b) Accelerated Gradient Descent

Figure 11.2. *The trajectories of the gradient descent and the accelerated gradient descent.*

In Figure 11.2(b), instead of using the local gradient at x_t, we can use the information in the history and combine $-\nabla f(x_t)$ with the previous points $x_t - x_{t-1}$. We can see that the combined direction is more stable and less zigzag. Nesterov proposed the following accelerated gradient descent which by its name converges faster than the gradient descent.

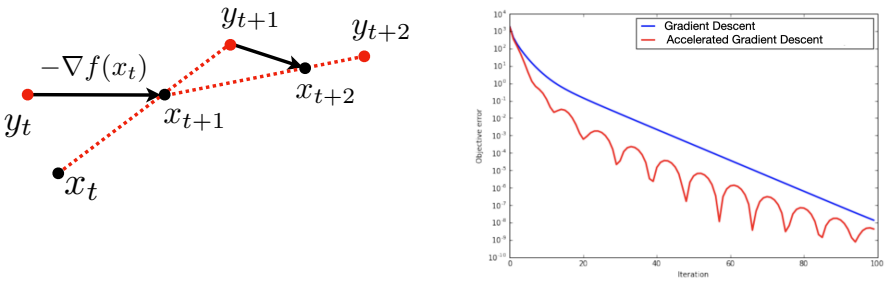
The **Nesterov's accelerated gradient descent** (AGD) algorithm: Initialize $x_0 = y_0$,

$$\begin{aligned} x_{t+1} &= y_t - \eta_t \nabla f(y_t) \\ y_{t+1} &= x_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(x_{t+1} - x_t), \end{aligned}$$

where $\lambda_0 = 1, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$.

From Figure 11.3(a), we can see that comparing to the gradient descent, the updating rule of the accelerated gradient descent utilizes the history of the trajectory through the

$(x_{t+1} - x_t)$ term. However, the weight $\frac{\lambda_t - 1}{\lambda_{t+1}}$ is somehow mysterious. We will try to provide some explaination in the next lecture. Since the AGD utilizes the history, unlike the gradient descent, the trajectory of AGD is not monotonically decreasing towards the minimum. See Figure 11.3(b).



(a) The trajectory of AGD (b) Comparison between GD and AGD

Figure 11.3. *The objective values of the gradient descent is monotonically decreasing while the AGD does not.*

The following theorem shows the convergence rate of AGD.

11.10 Theorem (Convergence of AGD). *Let f be convex and L-smooth. If we choose $\eta_t = 1/L$, then Nestorov's accelerated gradient descent has*

$$f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|_2^2}{(t+1)^2}$$

This means that the converge rate of AGD is $O(1/t^2)$, which is faster than the $O(1/t)$ rate of the gradient descent in Theorem 11.4. Or to rephrase, for convex and smooth objective functions, we can achieve ϵ-accuracy within $O(1/\sqrt{\epsilon})$ steps.

We will show the proof of this theorem and give more insights of the AGD next time.

Lecture 12

Proximal Gradient Descent

12.1 Proximal Perspective

In the previous lecture, we introduce the gradient descent and accelerated gradient algorithm to solve the unconstrained optimization. We show the the convergence rates of these two algorithms when the objective function is smooth. However, in Lasso $\min_\beta \frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1$, the ℓ_1-norm penalty term is not smooth. For many other high dimensional M-estimators, e.g., these examples we introduced in Lecture 9, they can formulated as the optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) = \min_{x\in\mathbb{R}^d} f(x) + h(x),$$

where $f(x)$ is the loss function, which is typically convex and smooth, and $h(x)$ is the penalty term, which is convex but usually non-differentiable. If we directly apply the (sub)-gradient descent algorithm to these problems, the convergence rate will be deteriorated by the non-smooth part of the objective function.

In this lecture, we will focus on the algorithms solving this type of composite loss and expect them to converge as fast as the gradient descent for smooth objective functions. Before we introduce the new algorithm, we first start with giving some new insight of the gradient descent

$$x_{t+1} = x_t - \eta_t \nabla f(x_t).$$

In the previous lecture, we motivated the gradient descent by showing the $-\nabla f(x_t)$ is the steepest descent direction. An alternative perspective to understand the gradient descent is that we aim to approximate the objective function $f(x)$ around $x = x_t$ by a quadratic function

$$f(x) \approx \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t\rangle}_{\text{1st order Taylor expansion}} + \underbrace{\frac{1}{2\eta_t}\|x - x_t\|_2^2}_{\text{proximal term}}.$$

Instead of minimizing $f(x)$, we first minimize its quadratic approximation

$$x_{t+1} = \arg\min_{x\in\mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\}, \tag{12.1}$$

see Figure 12.1 for an illustration.

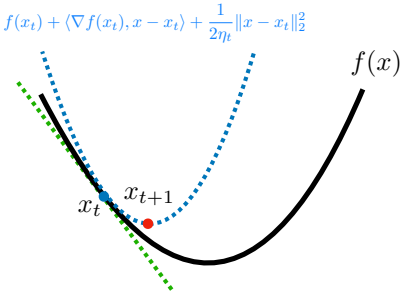


Figure 12.1. *The proximal perspective of gradient descent*

The problem in (12.1) has a closed form solution which is exactly gradient descent:

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in\mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\} \\ &= \arg\min_{x\in\mathbb{R}^d} \left\{ \frac{1}{2}\|x - (x_t - \eta_t \nabla f(x_t))\|_2^2 \right\} \\ &= x_t - \eta_t \nabla f(x_t). \end{aligned}$$

So from the proximal perspective, in each iteration, the gradient descent tries to minimize a local quadratic approximation of the objective function.

Now let us go back to the composite loss $F(x) = f(x) + h(x)$. We can modify the proximal perspective of the gradient descent in (12.1) as

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in\mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 + h(x) \right\} \\ &= \arg\min_{x\in\mathbb{R}^d} \left\{ \frac{1}{2}\|x - (x_t - \eta_t \nabla f(x_t))\|_2^2 + \eta_t h(x) \right\}. \end{aligned}$$

We then will have the following algorithm to solve $\min_{x\in\mathbb{R}^d} f(x) + h(x)$.

Define the **proximal operator** as

$$\operatorname{prox}_h(x) = \arg\min_{z\in\mathbb{R}^d} \left\{ \frac{1}{2}\|z - x\|_2^2 + h(z) \right\}.$$

. The **proximal gradient descent** can be written as

$$x_{t+1} = \operatorname{prox}_{\eta_t h}\big(x_t - \eta_t \nabla f(x_t)\big).$$

We will give a few examples of the proximal gradient descent.

12.2 Example (Constrained optimization). Although the proximal gradient descent is designed for an unconstrained problem, we can reformulate the constrained optimization $\min_{x\in M} f(x)$ as the unconstrained composite form $\min_{x\in\mathbb{R}^d} f(x) + h(x)$, where the indicator function

$$h(x) = \begin{cases} 0, & \text{if } x \in \mathcal{X}; \\ \infty, & \text{if } x \notin \mathcal{X}. \end{cases}$$

We can solve the proximal operator

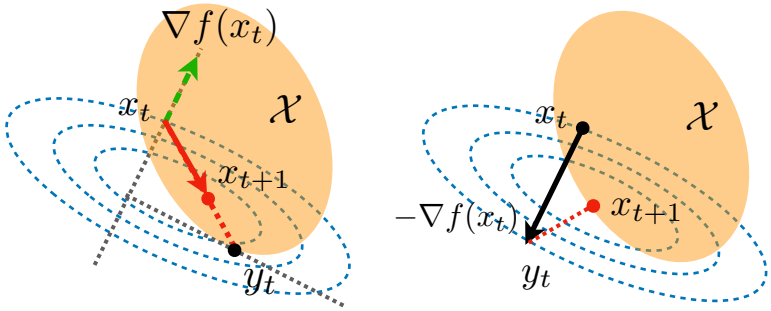
$$\text{prox}_h(x) = \arg\min_{z\in\mathbb{R}^d} \frac{1}{2}\|z - x\|_2^2 + h(z) = \arg\min_{z\in\mathcal{X}} \|z - x\|_2^2,$$

which is projecting x to the constraint M. Thus, the proximal gradient descent gives us the projected gradient descent algorithm to solve the constrained optimization.

The **projected gradient descent** is

$$\begin{aligned} y_t &= x_t - \eta_t \nabla f(x_t); \\ x_{t+1} &= \arg\min_{x\in\mathcal{X}} \|y_t - x\|_2^2. \end{aligned}$$

We illustrated the projected gradient descent in Figure 12.2. Now we have two algorithms to solve the constrained optimization, the Frank-Wolfe algorithm and the projected gradient descent. Both algorithms need to solve a sub-problem in the iterations. The projected gradient descent need to find the projection, while the Frank-Wolfe algorithm needs to solve $\arg\min_{x\in\mathcal{X}} \langle \nabla f(x_{t-1}), x\rangle$. In practice, we will tend to choose the algorithm such that the sub-problem is easier to compute. □



(a) Frank-Wolfe Algorithm (b) Projected Gradient Descent

Figure 12.2. *The comparison between the projected gradient descent and the Frank-Wolfe algorithm.*

12.3 Example (Lasso). The ℓ_1-penalized optimization has the objective function $\min_x f(x)+\lambda\|x\|_1$. The proximal operator

$$\text{prox}_h(x) = \arg\min_{z\in\mathbb{R}^d} \frac{1}{2}\|z-x\|_2^2 + \lambda\|z\|_1$$

becomes the soft-threshold operator

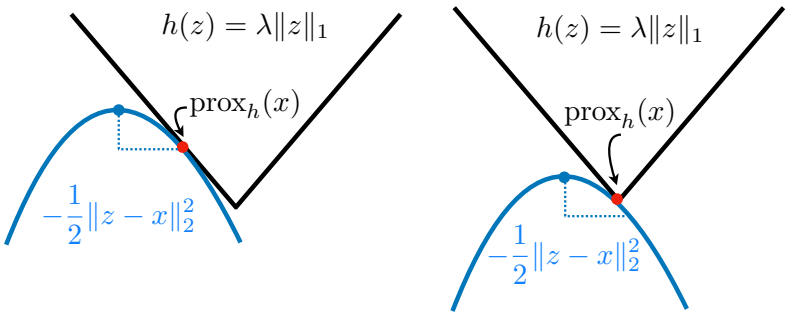
$$[\text{prox}_h(x)]_j = [\text{Soft-Threshold}(x,\lambda)]_j = \begin{cases} x_j - \lambda, & \text{if } x_j \geq \lambda; \\ x_j + \lambda, & \text{if } x_j \leq -\lambda; \\ 0, & \text{otherwise,} \end{cases} \tag{12.4}$$

for all $j = 1,\dots,d$. We visualize the proximal operator of ℓ_1-norm in Figure 12.3. The proximal gradient descent reduces to the following algorithm.

The Iterative Shrinkage-Thresholding Algorithm (**ISTA**) solves the Lasso problem $\min_{x\in\mathbb{R}^d} f(x)+\lambda\|x\|_1$ as

$$\begin{aligned} y_t &= x_t - \eta_t \nabla f(x_t); \\ x_{t+1} &= \text{Soft-Threshold}(y_t, \lambda\eta_t). \end{aligned}$$

□



(a) $|x| \geq \lambda$ (b) $|x| < \lambda$

Figure 12.3. *The proximal operator of ℓ_1-norm and the soft-thresholding.*

Even if the objective function $F(x) = f(x) + h(x)$ has the non-smooth $h(x)$, the following theorem shows that the proximal gradient descent has the same convergence rate $O(1/t)$ as the gradient descent.

12.5 Theorem (Convergence rate of proximal gradient descent). *Suppose f is convex and L-smooth and h is convex. If $\eta_t = 1/L$, the proximal gradient descent has*

$$F(x_t) - F(x^*) \leq \frac{L\|x_0 - x^*\|_2^2}{2t}. \tag{12.6}$$

Before we prove the theorem, we need to establish a fundamental inequality in the following lemma.

12.7 Lemma. *Suppose f is convex and L-smooth and h is convex. For any $x, y \in \mathbb{R}^d$, let $y^+ = \operatorname{prox}_{\frac{1}{L}h}\big(y - \frac{1}{L}\nabla f(y)\big)$, then*

$$F(y^+) - F(x) \le \frac{L}{2}\|x - y\|_2^2 - \frac{L}{2}\|x - y^+\|_2^2.$$

This lemma provides a very useful inequality which bounds the objective function by the distance of points.

Proof. [Proof of Lemma 12.7] We have

$$\begin{aligned} F(y^+) - F(x) &= \underbrace{f(y^+) - f(y)}_{L\text{-smooth}} + \underbrace{f(y) - f(x)}_{\text{convexity}} + \underbrace{h(y^+) - h(x)}_{\text{convexity}} \\ &\le \nabla f(y)^\top (y^+ - y) + \frac{L}{2}\|y^+ - y\|_2^2 + \nabla f(y)^\top (y - x) + \partial h(y^+)^\top (y^+ - x) \\ &= \langle \nabla f(y) + L(y^+ - y) + \partial h(y^+), y^+ - x\rangle \\ &\quad + \frac{L}{2}\|y^+ - y\|_2^2 - L\langle y^+ - y, y^+ - x\rangle, \end{aligned}$$

where in the first inequality, we apply the inequalities on the L-smoothness in Definition 11.2 and the subgradient in Definition 10.3. Recall that we define y^+ as

$$y^+ = \operatorname{prox}_{\frac{1}{L}h}\Big(y - \frac{1}{L}\nabla f(y)\Big) = \arg\max_z \underbrace{f(y) + \langle \nabla f(y), z - y\rangle + \frac{L}{2}\|z - y\|_2^2 + h(z)}_{\phi(z)}.$$

By the first optimality condition, we have

$$\partial\phi(y^+) = \nabla f(y) + L(y^+ - y) + \partial h(y^+) = 0.$$

Therefore, we have

$$F(y^+) - F(x) \le \frac{L}{2}\|y_t - y\|_2^2 - L\langle y^+ - y, y^+ - x\rangle = \frac{L}{2}\|x - y\|_2^2 - \frac{L}{2}\|x - y^+\|_2^2,$$

where in the last equality, we use the identity $\|a\|_2^2 - 2\langle a, b\rangle = \|a - b\|_2^2 - \|b\|_2^2$. □

Now we can go back to the proof of the convergence rate of the proximal gradient descent.

Proof. [Proof of Theorem 12.5] First, we prove that the values of the objective function is decreasing with t. We apply Lemma 12.7 by letting $x = x_t$, $y = x_t$, $y^+ = x_{t+1}$ and have

$$F(x_{t+1}) - F(x_t) \le -\frac{L}{2}\|x_{t+1} - x_t\|_2^2 < 0.$$

So $F(x_t)$ decreases as t increases.

Second, we apply Lemma 12.7 again by letting $x = x^*$, $y = x_t$, $y^+ = x_{t+1}$ and have

$$F(x_{t+1}) - F(x^*) \leq \frac{L}{2}\|x_t - x^*\|_2^2 - \frac{L}{2}\|x_{t+1} - x^*\|_2^2.$$

We sum up the above inequality from 0 to $t-1$ and get

$$\sum_{s=0}^{t-1}(F(x_{s+1}) - F(x^*)) \leq \frac{L}{2}\|x_0 - x^*\|_2^2 - \frac{L}{2}\|x_t - x^*\|_2^2.$$

As in the first part, we showed $F(x_s) - F(x^*)$ decreases as t increases, we have

$$F(x_t) - F(x^*) \leq \frac{L\|x_0 - x^*\|_2^2}{2t}.$$

□

12.2 Accelerated Proximal Gradient Descent

Theorem 12.5 shows that the proximal gradient descent has the convergence rate $O(1/t)$, which is same as the gradient descent. In the previous lecture, we introduced the Nesterov's accelerated gradient descent and shows that it converges faster than the gradient descent with the rate $O(1/t^2)$.

We can also apply Nesterov's idea to the proximal gradient descent and derive the following algorithm.

The **accelerated proximal gradient descent** algorithm: Initialize $x_0 = y_0$,

$$x_{t+1} = \text{prox}_{\eta_t h}\big(y_t - \eta_t \nabla f(y_t)\big);$$
$$y_{t+1} = x_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(x_{t+1} - x_t),$$

where $\lambda_0 = 1, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$.

For example, we can accelerated ISTA for Lasso $\min_x f(x) + \lambda\|x\|_1$ in Example 12.3 via the following algorithm.

The Fast Iterative Shrinkage-Thresholding Algorithm (**FISTA**) solves the Lasso problem $\min_{x\in\mathbb{R}^d} f(x) + \lambda\|x\|_1$ as

$$x_{t+1} = \text{Soft-Threshold}(y_t - \eta_t \nabla f(y_t), \lambda\eta_t);$$
$$y_{t+1} = x_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(x_{t+1} - x_t),$$

where $\lambda_0 = 1, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $x_0 = y_0$.

We can show that the convergence rate of the accelerated proximal gradient descent is $O(1/t^2)$.

12.8 Theorem (Convergence rate of accelerated proximal gradient descent). *Suppose f is convex and L-smooth and h is convex. If $\eta_t = 1/L$, the accelerated proximal gradient descent has*

$$F(x_t) - F(x^*) \le \frac{2L\|x_0 - x^*\|_2^2}{(t+1)^2}.$$

When $h = 0$, the accelerated proximal gradient descent reduces to AGD. So Theorem 11.10 in Lecture 11 is a special case of Theorem 12.8.

Proof. Recall that the unlike the gradient descent, the AGD is not monotonically decreasing. So our idea to prove the theorem is to construct a Lyapunov function which is monotonically decreasing using the accelerated proximal gradient descent.

The following lemma gives the construction of the Lyapunov function.

12.9 Lemma (Lyapunov function). *Let $u_t = \lambda_{t-1}x_t - (x^* + (\lambda_{t-1} - 1)x_{t-1})$ We define the Lyapunov function*

$$L_t = \|u_t\|_2^2 + \frac{2}{L}\lambda_{t-1}^2(F(x_t) - F(x^*)). \tag{12.10}$$

Then L_t is decreasing with t, i.e. $L_t \le L_{t-1} \le \ldots \le L_1$.

We first focus on the proof of Theorem 12.8 and leave the proof of Lemma 12.9 to the end of this lecture.

We apply Lemma 12.7 by letting $x = x^*$, $y = y_{t-1}$, $y^+ = x_t$ and have

$$F(x_t) - F(x^*) \le \frac{L}{2}\left(\|y_{t-1} - x^*\|_2^2 - \|x_t - x^*\|_2^2\right).$$

Let $t = 1$ and we have

$$\frac{2}{L}(F(x_1) - F(x^*)) \le \|y_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2 = \|x_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2,$$

as we initialize $x_0 = y_0$. Recall the definition of u_t and L_t in Lemma 12.9. As $u_1 = \lambda_0 x_1 - (x^* + (\lambda_0 - 1)x_0) = x_1 - x^*$ and $\lambda_0 = 1$, we have

$$L_1 = \|u_1\|_2^2 + \frac{2\lambda_0^2}{L}(F(x_1) - F(x^*)) = \|x_1 - x^*\|_2^2 + \frac{2}{L}(F(x_1) - F(x^*)) \leq \|x_0 - x^*\|_2^2.$$

By Lemma 12.9, the Lyapunov function is decreasing $L_t \leq L_1$, i.e.,

$$L_t = \|u_t\|_2^2 + \frac{2}{L}\lambda_{t-1}^2(F(x_t) - F(x^*)) \leq L_1 \leq \|x_0 - x^*\|_2^2.$$

This implies that

$$F(x_t) - F(x^*) \leq \frac{L\|x_0 - x^*\|_2^2}{2\lambda_{t-1}^2}.$$

By induction, we can show that $\lambda_t \geq \frac{t+2}{2}$, and therefore we have the convergence rate

$$F(x_t) - F(x^*) \leq \frac{2L\|x_0 - x^*\|_2^2}{(t+1)^2}.$$

□

We complete our lecture by showing that the Lyapunov function is monotonically decreasing.

Proof. [Proof of Lemma 12.9] We apply Lemma 12.7 by letting $x = {\lambda_t}^{-1}x^* + (1 - {\lambda_t}^{-1})x_t$, $y^+ = y_t$, $y_t = x_{t+1}$ and have

$$\begin{aligned} &F(x_{t+1}) - F\big({\lambda_t}^{-1}x^* + (1 - {\lambda_t}^{-1})x_t\big) \\ &\leq \tfrac{L}{2}\|{\lambda_t}^{-1}x^* + (1 - {\lambda_t}^{-1})x_t - y_t\|_2^2 - \tfrac{L}{2}\|{\lambda_t}^{-1}x^* + (1 - {\lambda_t}^{-1})x_t - x_{t+1}\|_2^2 \\ &= \tfrac{L}{2\lambda_t^2}(\|x^* + (\lambda_t - 1)x_t - \lambda_t y_t\|_2^2 - \|x^* + (\lambda_t - 1)x_t - \lambda_t x_{t+1}\|_2^2). \end{aligned} \tag{12.11}$$

Recall in Lemma 12.9 we define $u_{t+1} = \lambda_t x_{t+1} - (x^* + (\lambda_t - 1)x_t)$ and the AGD gives $y_t = x_t + \frac{\lambda_{t-1}-1}{\lambda_t}(x_t - x_{t-1})$. So

$$\begin{aligned} \lambda_t y_t - (x^* + (\lambda_t - 1)x_t) &= \lambda_t x_t + (\lambda_{t-1} - 1)(x_t - x_{t-1}) - (x^* + (\lambda_t - 1)x_t) \\ &= \lambda_{t-1}x_t - (x^* + (\lambda_{t-1} - 1)x_{t-1}) = u_t. \end{aligned}$$

By (16.2), we have

$$\begin{aligned} &F(x_{t+1}) - F\big({\lambda_t}^{-1}x^* + (1 - {\lambda_t}^{-1})x_t\big) \\ &\leq \frac{L}{2\lambda_t^2}(\|x^* + (\lambda_t - 1)x_t - \lambda_t y_t\|_2^2 - \|x^* + (\lambda_t - 1)x_t - \lambda_t x_{t+1}\|_2^2) \\ &= \frac{L}{2\lambda_t^2}(\|u_t\|_2^2 - \|u_{t+1}\|_2^2). \end{aligned}$$

Therefore, the above inequality derives

$$\begin{aligned}\frac{L}{2}(\|u_t\|_2^2 - \|u_{t+1}\|_2^2) &\geq \lambda_t^2\Big[F(x_{t+1}) - F\big({\lambda_t}^{-1}x^* + (1-{\lambda_t}^{-1})x_t\big)\Big] \\ &\geq \lambda_t^2\Big[F(x_{t+1}) - {\lambda_t}^{-1}F(x^*) - (1-{\lambda_t}^{-1})F(x_t)\Big] \\ &= \lambda_t^2(F(x_{t+1} - F(x^*)) - (\lambda_t^2 - \lambda_t)(F(x_t) - F(x^*)),\end{aligned}$$

where we use the convexity of $F(x)$ in the second inequality. Recall the sequence

$$\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2} \implies (2\lambda_t - 1)^2 = 1 + 4\lambda_{t-1}^2 \implies \lambda_{t-1}^2 = \lambda_t^2 - \lambda_t.$$

Therefore, we have

$$\begin{aligned}\frac{L}{2}(\|u_t\|_2^2 - \|u_{t+1}\|_2^2) &\geq \lambda_t^2(F(x_{t-1}) - F(x^*)) - (\lambda_t^2 - \lambda_t)(F(x_t) - F(x^*)) \\ &= \lambda_t^2(F(x_{t-1}) - F(x^*)) - \lambda_{t-1}^2(F(x_t) - F(x^*)).\end{aligned}$$

Recall the Lyapunov function $L_t = \|u_t\|_2^2 + \frac{2}{L}\lambda_{t-1}^2(F(x_t) - F(x^*))$ and thus the above inequality implies $L_t \geq L_{t+1}$. □

Lecture 13

Mirror Descent

13.1 Bregman Divergence

In the previous lecture, we introduce the proximal perspective of the gradient descent. To minimize $f(x)$, we approximate the objective function $f(x)$ around $x = x_t$ by a quadratic function

$$f(x) \approx \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}_{\text{1st order Taylor expansion}} + \underbrace{\frac{1}{2\eta_t}\|x - x_t\|_2^2}_{\text{proximal term}}.$$

We consider the constrained optimization $\min_{x\in M} f(x)$. If we start at x_t, we can update the next step via minimizing the quadratic approximation

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in M} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\} \\ &= \arg\min_{x\in M} \left\{ \frac{1}{2}\|x - (x_t - \eta_t \nabla f(x_t))\|_2^2 \right\}. \end{aligned}$$

If there is no constrained, i.e., $M = \mathbb{R}^d$, the above step $x_{t+1} = x_t - \eta_t \nabla f(x_t)$. Otherwise, it is the projected gradient descent. We also notice that if there is no proximal term, it reduces to the Frank-Wolfe algorithm. We add the proximal term $\frac{1}{2\eta_t}\|x - x_t\|_2^2$ in the approximation to prevent x_{t+1} from being to far away from x_t. Otherwise, if there is no constraint, $\min_{x\in\mathbb{R}^d} \nabla f(x_t)^\top x = -\infty$. A natural question is why we have to use the ℓ_2-norm in the proximal term? Is it possible to use another distance?

The following simple example shows that we can probably get a better algorithm by changing the distance in the proximal term.

13.1 Example. Consider the quadratic optimization

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{2}(x - x^*)^\top Q(x - x^*),$$

where Q is a positive definite matrix.

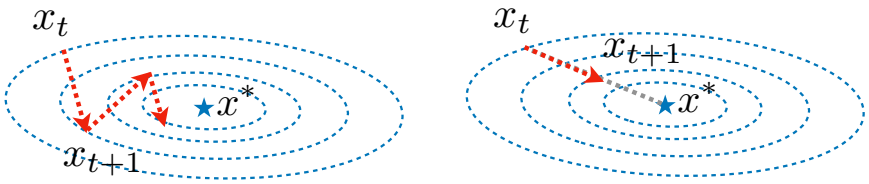
If we use the ℓ_2-norm in the proximal term, we have the gradient descent

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in\mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\} \\ &= x_t - \eta_t Q(x_t - x^*). \end{aligned}$$

In Figure 13.1(a), we can see that the trajectory of the gradient descent is zigzag.

On the other hand, we can imagine that why the gradient descent has a zigzag trajectory is because the ℓ_2-norm is not the perfect distance for the objective $f(x) = \frac{1}{2}(x-x^*)^\top Q(x-x^*)$. The contour is scaled by the matrix Q. What if we consider the norm $\|x\|_Q^2 = x^\top Q x$ in the proximal term? We therefore update x_{t+1} as

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in\mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\eta_t}\|x - x_t\|_Q^2 \right\} \\ &= \arg\min_{x\in\mathbb{R}^d} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\eta_t}(x - x_t)^\top Q(x - x_t) \right\} \\ &= x_t - \eta_t Q^{-1}\nabla f(x_t) = x_t - \eta_t(x_t - x^*). \end{aligned}$$



(a) $\|\cdot\|_2$-norm (b) $\|\cdot\|_Q$-norm

Figure 13.1. *The trajectory using the ℓ_2-norm versus using the $\|\cdot\|_Q$-norm in the proximal term.*

In Figure 13.1(b), we can see the descent direction directly points to the minimizer x^* which gives us a much faster algorithm. □

The previous example shows that we need to find a better distance fitting the geometry of the problem. This motivates us to define the following distance.

13.2 Definition (Bregman divergence). *Let $\varphi : M \to \mathbb{R}$ be a convex[3] and differentiable function, then the Bregman divergence between x and z is*

$$D_\varphi(x, z) = \varphi(x) - \varphi(z) - \langle \nabla\varphi(z), x - z\rangle.$$

By convexity, $D_\varphi(x, z) \geq 0$, so it is a kind of distance. Moreover, by mean-value theorem, $\varphi(x) - \varphi(z) - \langle \nabla\varphi(z), x - z\rangle \approx \frac{1}{2}(x - z)^\top \nabla^2\varphi(\xi)(x - z)$, where ξ depends

[3] Actually, a standard definition of the Bregman divergence needs to assume φ is strictly convex, i.e., $\varphi((1-\gamma)x + \gamma y) < (1-\gamma)\varphi(x) + \gamma\varphi(y)$, for any $x, y \in M, \gamma \in (0, 1)$. If φ is strictly convex, then $D_\varphi(x, z) = 0$ if and only if $z = x$.

on x and z. So the Bregman divergence is like a generalization of the quadratic norm $\|\cdot\|_Q$ in Example 13.1. In fact, when $\varphi(x) = \frac{1}{2}x^\top Qx$, we can easily see that $D_\varphi(x,z) = (x-z)^\top Q(x-z)$.

13.2 Mirror Descent

If we replace the ℓ_2-norm in the proximal term by the Bregman divergence, we have

$$x_{t+1} = \arg\min_{x\in M} \Big\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \Big\}.$$

This motivates the following algorithm.

The **mirror descent** algorithm is

$$x_{t+1} = \arg\min_{x\in M} \Big\{ \langle \nabla f(x_t), x\rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \Big\}.$$

Example 13.1 shows that we need to consider the Bregman divergence other than ℓ_2-norm due to the geometry of the objective function. However, for most of cases, we need to find a proper Bregman divergence due to the geometry of the constraint. The following example illustrates how to choose the proper Bregman divergence under a specific constraint.

13.3 Example (Probability simplex). In many statistical problems, we need to estimate the probability mass function of a discrete distribution. So the parameters are constrained to the the **probability simplex**, defined as

$$\Delta = \Big\{ x \in \mathbb{R}^d \;\Big|\; \sum_{i=1}^d x_i = 1, x_i \geq 0 \text{ for all } i = 1, \dots, d \Big\}.$$

See a visualization of a 2-dimensional probability simplex in 3-dimensional space in Figure 13.2.

A widely used Bregman divergence for the probability simplex use φ as the negative entropy:

$$\varphi(x) = \sum_{i=1}^d x_i \log x_i.$$

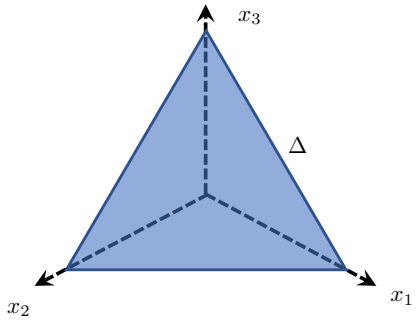


Figure 13.2. *The 2-dimensional probability simplex in 3-dimensional space*

By definition, the corresponding Bregman divergence becomes

$$\begin{aligned} D_\varphi(x,z) &= \varphi(x) - \varphi(z) - \langle \nabla\varphi(z), x - z\rangle \\ &= \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d z_i \log z_i - \sum_{i=1}^d (1+\log z_i)(x_i - z_i) \\ &= \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d (x_i - z_i) - \sum_{i=1}^d x_i \log z_i \\ &= \sum_{i=1}^d x_i \log \frac{x_i}{z_i}, \end{aligned}$$

where the last equality is due to $\sum_{i=1}^d x_i = \sum_{i=1}^d z_i = 1$. We also call $D_\varphi(x,z) = \sum_{i=1}^d x_i \log \frac{x_i}{z_i}$ as the Kullback–Leibler (KL) divergence, denoted as $D_{\mathrm{KL}}(x\|z)$. In fact, we have been using the KL-divergence all the time while we apply the maximum log-likelihood estimator

$$\arg\max_\theta \mathbb{E}_{\theta^*}[\log P_\theta(x)] = \arg\max_\theta \mathbb{E}_{\theta^*} \log \frac{P_\theta(x)}{P_{\theta^*}(x)} = \arg\min_\theta D_{\mathrm{KL}}(P_\theta \| P_{\theta^*}).$$

Therefore, the maximum log-likelihood estimator is essentially finding a distribution P_θ closest under the KL-divergence to the true distribution P_{θ^*}. From this example, we can also see that the Bregman divergence is not necessarily symmetric, i.e., $D_\varphi(x,z) \neq D_\varphi(z,x)$.

Now let us consider the constrained optimization problem $\min_{x\in\Delta} f(x)$. So far, we have learned three algorithms solving the constrained optimization.

1. **Frank-Wolfe algorithm**: We can see that the Frank-Wolfe algorithm is essentially the mirror descent with $\varphi = 0$. It involves solving a sub-problem

   $$\min_{x\in\Delta} \langle \nabla f(x_t), x\rangle,$$

   which is a linear programming. There is no closed form solution so we need to solve a linear programming in each iteration which is time-consuming.

2. **Projected gradient descent**: We can see that the projected gradient descent algorithm is essentially the mirror descent use the ℓ_2-norm as the proximal term, i.e., $\varphi(x) = \frac{1}{2}\|x\|_2^2$. It involves solving a sub-problem

$$\min_{x\in\Delta} \|x - (x_t - \eta_t \nabla f(x_t))\|_2^2,$$

which is a quadratic programming without closed form solution either.

3. **Mirror Descent**: If we use the KL-divergence $D_{\mathrm{KL}}(x\|x_t)$ in the mirror descent, the sub-problem becomes

$$\arg\min_{x\in\Delta} \left\{ \langle \nabla f(x_t), x\rangle + \frac{1}{\eta_t}\sum_{i=1}^{d} x_i \log\frac{x_i}{x_i^t} \right\}. \tag{13.4}$$

The sub-problem seems more complicated than the Frank-Wolfe and the projected gradient descent, however, unlike the linear or quadratic programming, (13.4) actually has a closed form solution.

We will derive the solution to (13.4) now. Apply the Lagrange multiplier

$$L(x,\lambda) = \langle \nabla f(x_t), x\rangle + \frac{1}{\eta_t}\sum_{i=1}^{d} x_i \log\frac{x_i}{x_i^t} + \lambda\Big(\sum_{i=1}^{d} x_i - 1\Big).$$

Then we have

$$\frac{\partial L(x,\lambda)}{\partial x_i}\Big|_{x=x_{t+1}} = \nabla_i f(x_t) + \frac{1}{\eta_t}\big[(1+\log x_i^{t+1}) - \log x_i^t\big] + \lambda = 0,$$

which implies that

$$x_i^{t+1} = x_i^t \exp(-\eta_t \nabla_i f(x_t)) \exp(-\eta_t\lambda - 1).$$

Apply the constraint $\sum_{i=1}^d x_i^{t+1} = 1$ and therefore we have

$$x_i^{t+1} = \frac{x_i^t \exp(-\eta_t \nabla_i f(x_t))}{\sum_{j=1}^d x_j^t \exp(-\eta_t \nabla_j f(x_t))}, \text{ for } i = 1,\dots,d.$$

We can see that the mirror descent has the closed form and it updates like

$$\begin{aligned} \log(y_{t+1}) &= \log(x_t) - \eta_t \nabla f(x_t), \\ x_{t+1} &= y_{t+1}/\|y_{t+1}\|_1. \end{aligned}$$

So we can see why we call the algorithm as the mirror descent. We run the gradient descent in the "mirror space" of $\log(x_t)$ and map it back to the probability simplex via normalization. □

In Figure 13.3, we give widely used Bregman divergence under different constraints.

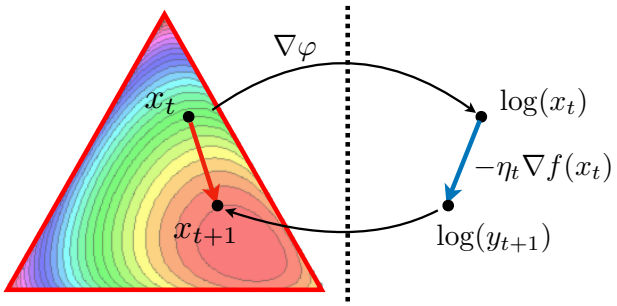


Figure 13.3. *The visualization of the mirror descent*

Function Name	$\varphi(x)$	dom φ	$D_\varphi(x;y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x - x$	$[0,+\infty)$	$x\log\frac{x}{y} - x + y$
Bit entropy	$x\log x + (1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y} - \log\frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \quad (0<p<1)$	$[0,+\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p \quad (1<p<\infty)$	$(-\infty,+\infty)$	$\lvert x\rvert^p - p\,x\operatorname{sgn} y\,\lvert y\rvert^{p-1} + (p-1)\,\lvert y\rvert^p$
Exponential	$\exp x$	$(-\infty,+\infty)$	$\exp x - (x-y+1)\exp y$
Inverse	$1/x$	$(0,+\infty)$	$1/x + x/y^2 - 2/y$

Figure 13.4. *Common φ functions and the corresponding Bregman divergences, taken from I. Dhillon & J. Tropp, 2007.*

13.3 Nesterov's Smoothing

We have introduce the proximal algorithm to solve the problem

$$\min_x f(x) + h(x).$$

When f is smooth and the accelerated proximal gradient descent has the convergence rate $O(1/t^2)$. What if the objective is not smooth? For example, in Lecture 9, we introduce the square-root Lasso

$$\min_\beta \|Y - \mathbb{X}\beta\|_2 + \lambda\|\beta\|_1.$$

We know that the ℓ_2-norm $\|x\|_2$ is not differentiable at 0. Can we somehow approximate the non-smooth function by the smooth function? In fact, Nesterov's smoothing idea is to

1. Approximate the non-smooth objective function by a smooth function.
2. Minimize the smooth approximation via (proximal) gradient descent or (proximal) accelerated gradient descent.

We rigor the idea above via the following definition.

13.5 Definition. *A convex function f is (α, β)-smoothable if for any $\mu > 0$, there exists a convex approximation f_μ such that*

1. *$f_\mu(x) \le f(x) \le f_\mu(x) + \beta\mu$, for all x,*
2. *f_μ is $\frac{\alpha}{\mu}$-smooth.*

We also call f_μ as the $\frac{1}{\mu}$-smooth approximation of f with parameters (α, β).

13.6 Example (ℓ_1-norm). We can approximate the absolute value $f(z) = |z|$ via the Huber loss

$$h_\mu(z) = \begin{cases} z^2/(2\mu), & \text{if } |z| \le \mu; \\ |z| - \mu/2, & \text{otherwise.} \end{cases}$$

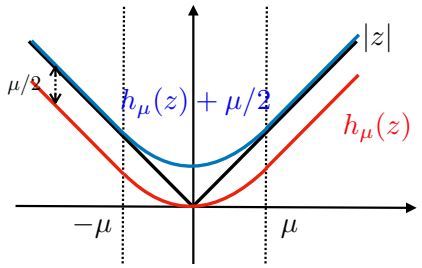


Figure 13.5. *The Huber loss function*

By Figure 13.5, we can see that $h_\mu(z) \le |z| \le h_\mu(z) + \mu/2$. As $h_\mu(z)$ is $\frac{1}{\mu}$-smooth, so $|z|$ is $(1, \frac{1}{2})$-smoothable.

For the ℓ_1-norm $f(z) = \|z\|_1$, we can approximate it by $\sum_{i=1}^d h_\mu(z_i)$. Therefore, we have

$$\sum_{i=1}^d h_\mu(z_i) \le \|z\|_1 \le \sum_{i=1}^d h_\mu(z) + d\mu/2.$$

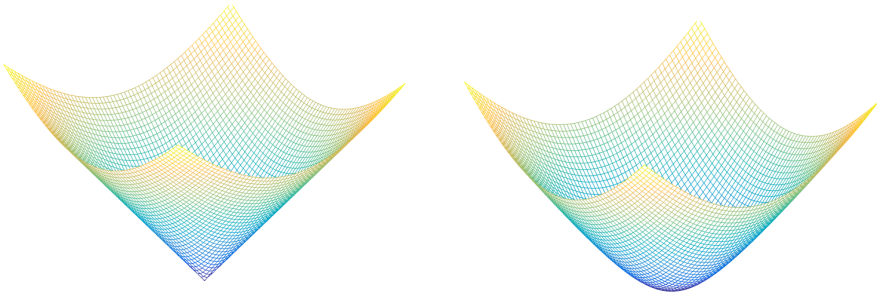
Therefore, $\|z\|_1$ is $(1, \frac{d}{2})$-smoothable. □

13.7 Example (ℓ_2-norm). We can approximate the ℓ_2-norm $f(x) = \|x\|_2$ by

$$f_\mu(x) = \sqrt{\|x\|_2^2 + \mu^2} - \mu.$$

We have

$$f_\mu(x) \le \|x\|_2 + \mu - \mu \le \|x\|_2 \le \sqrt{\|x\|_2^2 + \mu^2} = f_\mu(x) + \mu.$$



(a) $\|x\|_2$ (b) $\sqrt{\|x\|_2^2 + \mu^2} - \mu$

Thus, $\|x\|_2$ is $(1, 1)$-smoothable. Notice that comparing to the ℓ_1-norm, the smoothing parameters of $\|x\|_2$ is dimension free. Therefore, for the square-root loss $f(\beta) = \|Y - \mathbb{X}\beta\|_2$, we can approximate it by $f_\mu(\beta) = \sqrt{\|Y - \mathbb{X}\beta\|_2^2 + \mu^2} - \mu$. We can check that $f(\beta)$ is $(\|\mathbb{X}\|_{\text{op}}^2, 1)$-smoothable. □

The following theorem shows the convergence rate using Nesterov's smoothing idea.

13.8 Theorem. *Given the objective function $F(x) = f(x) + h(x)$, where f is (α, β)-smoothable and h is convex, let f_μ be the $\frac{1}{\mu}$-smooth approximation to f. If we apply the accelerated proximal gradient descent to $F_\mu(x) = f_\mu(x) + h(x)$, where we choose $\mu = \epsilon/(2\beta)$, then*

$$F(x_t) - F(x^*) \le \epsilon \text{ if } t \gtrsim \frac{\sqrt{\alpha\beta}}{\epsilon}.$$

Proof. First, as f_μ is α/μ-smooth, Theorem 12.8 shows that we can achieve ϵ-accuracy

$$F_\mu(x_t) - \min_x F_\mu(x) \le \epsilon$$

within $t = O(\sqrt{\frac{\alpha}{\mu}}\frac{1}{\sqrt{\epsilon}})$ steps. Second, by Definition 13.5, we have $|f(x_t) - f_\mu(x_t)| \le \beta\mu = \frac{\epsilon}{2}$ as we choose $\mu = \epsilon/(2\beta)$. Therefore, we have

$$F(x_t) - F(x^*) \le |f(x_t) - f_\mu(x_t)| + |F_\mu(x_t) - \min_x F_\mu(x)| \le \epsilon$$

within $t = O(\sqrt{\frac{\alpha}{\mu}}\frac{1}{\sqrt{\epsilon}}) = O(\sqrt{\frac{\alpha\beta}{\epsilon}}\frac{1}{\sqrt{\epsilon}}) = O(\frac{\sqrt{\alpha\beta}}{\epsilon})$ steps. □

For the square-root Lasso, if we apply the accelerated proximal gradient descent to

$$\min_\beta \sqrt{\|Y - \mathbb{X}\beta\|_2^2 + \mu^2} + \lambda\|\beta\|_1.$$

As $f(\beta) = \|Y - \mathbb{X}\beta\|_2$ is $(\|\mathbb{X}\|_{\text{op}}^2, 1)$-smoothable, as long as the maximum singular value of the design matrix $\mathbb{X}$ is bounded, we can achieve ϵ-accuracy of square-root Lasso within $O(1/\epsilon)$ steps.

Another idea is why not we apply the Huber loss to approximation the ℓ_1-norm in Lasso. Instead of applying FISTA, we apply the accelerated gradient descent to

$$\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \sum_{j=1}^{d} h_\mu(\beta_j),$$

where $h_\mu(z)$ is defined in Example 13.6. As the ℓ_1-norm $\|\beta\|_1$ is $(1, \frac{d}{2})$-smoothable, we achieve ϵ-accuracy within $O(\frac{\sqrt{d}}{\epsilon})$ steps. When the dimension d is large, the convergence rate is not good.

Lecture 14

Duality and ADMM

14.1 Composite Objective Function

In the previous lectures, we introduce the proximal gradient descent to solve the optimization problem $\min_x f(x) + g(x)$, where f is smooth but g is not differentiable. A key step in the proximal gradient descent is the solve the proximal operator

$$\text{prox}_g(x) = \arg\min_{z \in \mathbb{R}^d} \left\{ \frac{1}{2}\|z - x\|_2^2 + g(z) \right\}.$$

For example, when $g(x) = \lambda\|x\|_1$, the proximal operator is soft-thresholding. For the fused Lasso problem

$$\min_\beta \|Y - \mathbb{X}\beta\|_2^2 + \lambda\|D\beta\|_1,$$

where D is the differential map (See Section 9.3.2 in Lecture 9), if we apply the proximal gradient descent to solve the above problem, we need to solve the sub-problem

$$\arg\min_{z \in \mathbb{R}^d} \left\{ \frac{1}{2}\|z - x\|_2^2 + \lambda\|Dz\|_1 \right\}.$$

Unlike the ℓ_1-norm, the above problem does not have a closed form solution. Therefore, it will be inefficient to implement proximal gradient descent to solve fused Lasso.

In this lecture, we will discuss how to solve this type of composite objective functions. In general, we are interested in the following composite optimization problem

$$\min_{x,y} f(x) + g(y), \text{ s.t. } Ax + By = c, \tag{14.1}$$

for some convex f and g. We can see that the fused Lasso can be reduced the formality above by letting $f(\beta) = \|Y - \mathbb{X}\beta\|_2^2$, $g(y) = \lambda\|y\|_1$ so

$$\min_\beta \|Y - \mathbb{X}\beta\|_2^2 + \lambda\|D\beta\|_1 \Longleftrightarrow \min_{\beta,y} f(\beta) + g(y), \text{ s.t. } D\beta - y = 0.$$

Another example is the basis pursuit

$$\min_\beta \|\beta\|_1 \text{ s.t. } \mathbb{X}\beta - Y = 0.$$

14.2 Duality

In order to solve the composite optimization problem in (14.1), we start with reviewing the concept of duality in optimization. We define the Langrange multiplier function of (14.1) as

$$L(x, y, \lambda) = f(x) + g(y) + \lambda^\top (Ax + By - c).$$

This allows us to convert the **primal problem** in (14.1) to the

$$\textbf{Dual problem} : \max_{\lambda} h(\lambda), \text{ where } h(\lambda) = \min_{x,y} L(x, y, \lambda).$$

Duality is an important concept in optimization. The following theorem implies that we can solve the primal problem via the dual problem.

14.2 Theorem (Strong Duality). *Given convex functions f, g, the primal problem has the solution*

$$(x^*, y^*) = \underset{x,y}{\textit{arg min}}\, f(x) + g(y), \ \textit{s.t. } Ax + By = c.$$

Consider the dual problem $\lambda^* = \arg\max_\lambda h(\lambda)$, where $h(\lambda) = \min_{x,y} L(x, y, \lambda)$. We can solve (x^*, y^*) via

$$(x^*, y^*) = \underset{x,y}{\textit{arg min}}\, L(x, y, \lambda^*).$$

As $L(x, y, \lambda^*)$ is decomposable, we further have

$$x^* = \underset{x}{\textit{arg min}}\, f(x) + \lambda^{*\top} Ax \textit{ and } y^* = \underset{y}{\textit{arg min}}\, g(y) + \lambda^{*\top} By.$$

The above strong duality property is useful when we want to convert a linear constraint problem to an unconstrained dual problem. The following example shows how we can find the dual problem of Lasso.

14.3 Example (Duality of Lasso). As Lasso is an unconstrained problem, we first convert it to the standard formality in (14.1) as

$$\min_{\beta} \frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1 \iff \min_{\beta, z} \frac{1}{2}\|z\|_2^2 + \lambda\|\beta\|_1, \text{ s.t. } Y - \mathbb{X}\beta = z.$$

The Lagrange function is

$$L(\beta, z, \mu) = \frac{1}{2}\|z\|_2^2 + \lambda\|\beta\|_1 + \mu^\top (Y - \mathbb{X}\beta - z)$$

Notice that $L(\beta, z, \mu)$ is decomposable

$$\min_{\beta,z} L(\beta, z, \mu) = \min_{\beta} \left\{\lambda\|\beta\|_1 - \mu^\top \mathbb{X}\beta\right\} + \min_{z} \left\{\frac{1}{2}\|z\|_2^2 - \mu^\top z\right\} + \mu^\top Y$$

The second problem above is quadratic and we have

$$z^* = \arg\min_z \left\{ \frac{1}{2}\|z\|_2^2 - \mu^\top z \right\} \implies z^* - \mu = 0. \tag{14.4}$$

We also apply the first order optimality condition the first problem and get

$$\beta^* = \arg\min_\beta \left\{ \lambda\|\beta\|_1 - \mu^\top \mathbb{X}\beta \right\} \implies 0 = -\mathbb{X}^\top \mu + \lambda g, \text{ for some } g \in \partial\|\beta^*\|_1.$$

As $\|g\|_\infty \leq 1$, the equality $0 = -\mathbb{X}^\top\mu + \lambda g$ has a finite solution only if $\lambda \geq \|\mathbb{X}^\top\mu\|_\infty$. Moreover, if $\lambda \geq \|\mathbb{X}^\top\mu\|_\infty$, we multiple β^* on the both side of the equation $0 = -\mathbb{X}^\top\mu + \lambda g$ and have

$$0 = -\beta^{*\top}\mathbb{X}^\top\mu + \lambda\beta^{*\top} g = -(\beta^*)^\top\mathbb{X}^\top\mu + \lambda\|\beta^*\|_1,$$

where we use the fact that $\beta^{*\top} g = \beta^{*\top}\text{sign}(\beta^*) = \|\beta^*\|_1$ if $g \in \partial\|\beta^*\|_1$. Therefore, we have

$$\min_\beta \left\{ \lambda\|\beta\|_1 - \mu^\top\mathbb{X}\beta \right\} = \begin{cases} 0 & \text{if } \lambda \geq \|\mathbb{X}^\top\mu\|_\infty; \\ -\infty & \text{if } \lambda < \|\mathbb{X}^\top\mu\|_\infty. \end{cases} \tag{14.5}$$

Combining (16.2) and (16.4), the dual problems becomes

$$\begin{aligned} \max_\lambda \min_{\beta,z} L(\beta, z, \mu) &= \max_\lambda \left\{ \min_\beta \left\{ \lambda\|\beta\|_1 - \mu^\top\mathbb{X}\beta \right\} + \min_z \left\{ \frac{1}{2}\|z\|_2^2 - \mu^\top z \right\} + \mu^\top Y \right\} \\ &= \max_\lambda -\frac{1}{2}\|\mu\|_2^2 + \mu^\top Y \text{ s.t. } \lambda \geq \|\mathbb{X}^\top\mu\|_\infty, \end{aligned}$$

as when $\lambda < \|\mathbb{X}^\top\mu\|_\infty$, the objective value of the dual becomes $-\infty$ which could not be the maximum. Therefore, the dual problem of Lasso is

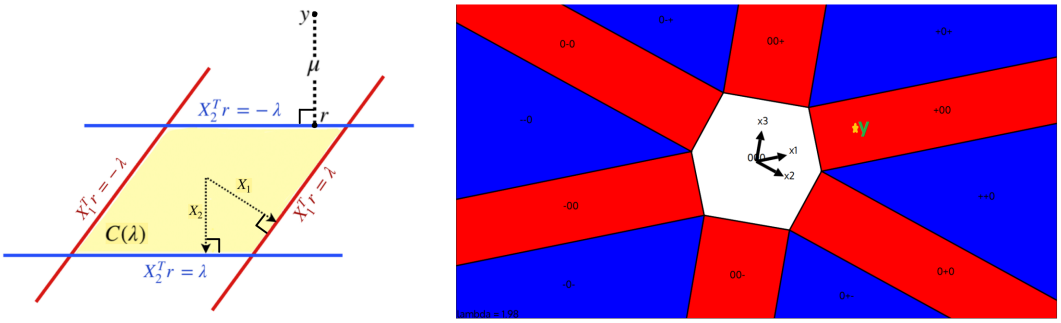
$$\max_\mu L(\beta^*, z^*, \mu) = \max_\mu -\frac{1}{2}\|\mu\|_2^2 + \mu^\top Y \text{ s.t. } \|\mathbb{X}^\top\mu\|_\infty \leq \lambda.$$

The above problem is equivalent to

$$\textbf{Duality of Lasso} : \min_\mu \|Y - \mu\|_2^2, \text{ s.t. } \|\mathbb{X}^\top\mu\|_\infty \leq \lambda. \tag{14.6}$$

By (16.2), we can see that the dual variable $\mu^* = z^* = Y - \mathbb{X}\beta^*$ is the residual. So duality of Lasso gives us a new geometric insight to Lasso. This duality form illustrates that, just like OLS, the Lasso is also a projection. Instead of projecting Y on the hyperplane $\mathcal{C}(\mathbb{X})$ for OLS, the Lasso projects Y on the polytope $C(\lambda) = \{\mu |\ \|\mathbb{X}^\top\mu\|_\infty \leq \lambda\}$. See Figure 14.1(a). The residual $\mu^* = Y - \mathbb{X}\beta^* = P_{C(\lambda)}(Y)$, where P_M is the projection operator such that $P_M(x) = \arg\min_{z\in M}\|z - x\|_2$.

In Figure 14.1(b), if Y locates in the white region, i.e., $\|\mathbb{X}^\top Y\|_\infty \leq \lambda$, then $P_{C(\lambda)}(Y) = Y = Y - \mathbb{X}\beta^*$. This implies that $\beta^* = 0$ as all the other vector in null($\mathbb{X}$) has a larger ℓ_1-norm. If Y lies in the red region, the projection is to one of the faces of the polytope. If Y lies in the blue region, the projection is to one of the vertex of the polytope. If we shrink the value of λ, the projection of Y will remain on the same face until Y locates in the blue region and may further pivots to another red region, which leads to a piece-wise linear regularization path of Lasso. □



(a) $d = n = 2$ (b) $d = 3, n = 2$

Figure 14.1. *A geometric interpretation of the dual Lasso problem. (Yes, I know you would say the Union Jack.)*

14.3 Alternating Direction Method of Multipliers

We consider a equivalent form of (14.1) by adding a quadratic term as follows

$$\begin{cases} \min\limits_{x,y} f(x) + g(y) \\ \text{s.t. } Ax + By = c, \end{cases} \iff \begin{cases} \min\limits_{x,y} f(x) + g(y) + \dfrac{\rho}{2}\|Ax + By - c\|_2^2 \\ \text{s.t. } Ax + By = c. \end{cases}$$

We then introduce **augmented Lagrangian** for the second problem above

$$L_\rho(x, y, \lambda) = f(x) + g(y) + \lambda^\top (Ax + By - c) + \frac{\rho}{2}\|Ax + By - c\|_2^2.$$

In order to solve the dual problem $\max_\lambda \min_{x,y} L_\rho(x, y, \lambda)$, we propose to update the primal variables x, y and dual variables alternatively as follows:

$$\textbf{Primal step:} \quad \begin{cases} x_{t+1} = \arg\min_x L_\rho(x, y_t, \lambda_t); \\ y_{t+1} = \arg\min_y L_\rho(x_{t+1}, y, \lambda_t); \end{cases}$$

$$\textbf{Dual step:} \quad \lambda_{t+1} = \arg\max_\lambda L_\rho(x_{t+1}, y_{t+1}, \lambda) - \underbrace{\frac{1}{2\rho}\|\lambda - \lambda_t\|_2^2}_{\text{Proximal term}}.$$

We add a proximal term in the dual step above, because the augmented Lagrangian $L_\rho(x_{t+1}, y_{t+1}, \lambda)$ is linear with respective to λ and thus $\max_\lambda L_\rho(x_{t+1}, y_{t+1}, \lambda) = \infty$. So we need to add a quadratic proximal term to stop λ_{t+1} deviating from λ_t too much.

Plugging the augmented Lagrangian into the above primal and dual steps, we have following algorithm.

The **Alternating Direction Method of Multipliers (ADMM)** solves (14.1) via

$$\begin{aligned} x_{t+1} &= \arg\min_x f(x) + \frac{\rho}{2}\|Ax + By_t - c + \lambda_t/\rho\|_2^2 \\ y_{t+1} &= \arg\min_y g(x) + \frac{\rho}{2}\|Ax_{t+1} + By - c + \lambda_t/\rho\|_2^2 \\ \lambda_{t+1} &= \lambda_t + \rho(Ax_{t+1} + By_{t+1} - c) \end{aligned}$$

It can be shown that if f and g are closed convex functions[4], ADMM algorithm converges.

14.7 Example (Fused Lasso). We apply ADMM to the fused Lasso

$$\min_\beta \frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|D\beta\|_1 \iff \min_{\beta,z} \frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|z\|_1 \text{ s.t. } D\beta - z = 0,$$

and get the updating rule as

$$\begin{aligned} \beta_{t+1} &= \arg\min_\beta \frac{1}{2}\|Y - \mathbb{X}\beta\|_2^2 + \frac{\rho}{2}\|D\beta - z_t + \lambda_t/\rho\|_2^2 \\ z_{t+1} &= \arg\min_z \lambda\|z\|_1 + \frac{\rho}{2}\|D\beta_{t+1} - z + \lambda_t/\rho\|_2^2 \\ \lambda_{t+1} &= \lambda_t + \rho(D\beta_{t+1} - z_{t+1}). \end{aligned}$$

This further yields the algorithm

$$\begin{aligned} \beta_{t+1} &= (\mathbb{X}^\top\mathbb{X} + \rho D^\top D)^{-1}(\mathbb{X}^\top Y + \rho D^\top z_t - D^\top\lambda_t) \\ z_{t+1} &= \text{SoftThreshold}(D\beta_{t+1} + \lambda_t/\rho, \lambda/\rho) \\ \lambda_{t+1} &= \lambda_t + \rho(D\beta_{t+1} - z_{t+1}). \end{aligned}$$

When $D = I$, the above algorithm reduces to the ADMM algorithm for Lasso. From these example, we can also see that ADMM is not a first order method as it involves information of f and g. For Lasso, unlike FISTA, the ADMM algorithm involves the matrix inversion which may be time-consuming when d is large. □

14.8 Example (Graphical Lasso). Given i.i.d. samples $X_1, \ldots, X_n \sim N(0, \Sigma)$, we aim to estimate the precision matrix $\Theta = \Sigma^{-1}$ via the graphical Lasso (see Lecture 9)

$$\begin{aligned} &\min_\Theta -\log\det\Theta + \text{tr}(\Theta^\top\widehat{\Sigma}) + \lambda\|\Theta\|_{1,1} \\ &\qquad\qquad \Updownarrow \\ &\min_{\Theta,\Psi} -\log\det\Theta + \text{tr}(\Theta^\top\widehat{\Sigma}) + \lambda\|\Psi\|_{1,1} \text{ s.t. } \Theta = \Psi \end{aligned}$$

[4]A convex function f is closed if $\{x|f(x) \le \alpha\}$ is a closed set for any α.

Applying the ADMM algorithm we get

$$\begin{aligned}\Theta_{t+1} &= \arg\min_{\Theta} -\log\det\Theta + \mathrm{tr}(\Theta^\top\widehat{\Sigma}) + \frac{\rho}{2}\|\Theta - \Psi_t + \Lambda_t/\rho\|_{\mathrm{F}}^2\\ \Psi_{t+1} &= \arg\min_{\Psi} \lambda\|\Psi\|_{1,1} + \frac{\rho}{2}\|\Theta_{t+1} - \Psi + \Lambda_t/\rho\|_{\mathrm{F}}^2\\ \Lambda_{t+1} &= \Lambda_t + \rho(\Theta_{t+1} - \Psi_{t+1}),\end{aligned}$$

where the Frobenious norm $\|A\|_{\mathrm{F}}^2 = \sum_{jk} A_{jk}^2$ and it can be further simplified as

$$\begin{aligned}\Theta_{t+1} &= \mathcal{F}_\rho(\Psi_t - \Lambda_t/\rho - \widehat{\Sigma}/\rho)\\ \Psi_{t+1} &= \text{SoftThreshold}(\Theta_{t+1} + \Lambda_t/\rho, \lambda/\rho)\\ \Lambda_{t+1} &= \Lambda_t + \rho(\Theta_{t+1} - \Psi_{t+1}),\end{aligned}$$

where for the spectral decomposition $X = UDU^\top$, where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, one has $\mathcal{F}_\rho(X) = U\mathrm{diag}\{\lambda_i + \sqrt{\lambda_i + 4/\rho}\}U^\top$. □

The last example shows that ADMM algorithm can be implemented as a distributed algorithm.

14.9 Example (Consensus optimization). We aim design a divide-and-conquer ADMM for Lasso for massive data when n is ultra-large. The Lasso

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \mathbb{X}_i\beta)^2 + \lambda\|\beta\|_1$$

can be formulated as the general block separable form

$$\min_{x} \sum_{i=1}^{n} f_i(x) \iff \min_{x_i, z} \sum_{i=1}^{n} f_i(x) \text{ s.t. } x_i = z \text{ for all } i = 1, \ldots, n.$$

Applying the ADMM algorithm we get

$$\begin{aligned}\text{Divide} : x_i^{t+1} &= \arg\min_{x_i} f_i(x_i) + \frac{\rho}{2}\|x_i - z^t + \lambda_i^t/\rho\|_2^2, \forall i = 1, \ldots, n\\ \text{Gather} : z^{t+1} &= \frac{1}{n}\sum_{i=1}^{n}(x_i^{t+1} + \lambda_i^t/\rho)\\ \text{Broadcast} : \lambda_i^{t+1} &= \lambda_i^t + \rho(x_i^{t+1} - z^{t+1}), \forall i = 1, \ldots, n.\end{aligned}$$

In the first step, we can update each x_i^{t+1} in parallel, then in the second step, we gather all local iterates and in the third step, we broadcast the updated dual variables back to each core. □

Lecture 15

High Dimensional Inference

15.1 Statistical Inference

We start the fourth part of our course on high dimensional inference. We start with the introducing the problems in statistical inference. Given the statistical model $\{\mathbb{P}_\theta|\theta \in \Theta\}$, we observe $X_1, ..., X_n \overset{iid}{\sim} \mathbb{P}_{\theta^*}$ where θ^* is the truth. Here are the major goals to estimation and inference.

- **Estimation**: We want to find an estimator $\widehat{\theta}$ and we're interested in the estimation rate or the statistical rate $\|\widehat{\theta} - \theta^*\| = O_P(r_n)$ for some rate r_n. In terms of estimation, we'd like to know whether the estimator converges to the truth in probability or how fast it converges to the truth in probability. For example, we show that under certain conditions, Lasso estimator has the ℓ_2-norm rate $O_P(\sqrt{s \log d/n})$.
- **Inference**: we aim to pursue more than a consistent estimator. The inference problems focus on the uncertainty assessment of our estimates. We aim to measure how much certainty we have to our estimates, how significant are the detected signals and how confident are our statistical decisions.

In statistical inference, there are two major problems: confidence interval and hypothesis testing.

- **Confidence interval**: We aim to construct a confidence interval I_α such that the true parameter belongs to the confidence interval with probability larger than $1 - \alpha$:

$$\mathbb{P}(\theta^* \in I_\alpha) \geq 1 - \alpha.$$

- **Hypothesis testing**: We aim to test certain properties of the true parameters, whether the true parameters belong to the null space Θ_0 or it belongs to the alternative Θ_1:

$$H_0 : \theta^* \in \Theta_0 \quad v.s. \quad H_1 : \theta^* \in \Theta_1$$

We aim to propose a test with the significance level $1-\alpha$

$$\psi_\alpha = \begin{cases} 0, & \text{not reject } H_0; \\ 1, & \text{reject } H_0. \end{cases}$$

such that the type-I error $\mathbb{P}_{H_0}(\psi_\alpha = 1) \leq \alpha$ and the type II error $\mathbb{P}_{H_1}(\psi_\alpha = 0) \to 1$ as $n \to \infty$.

15.2 High Dimensional Inference

In high dimensional inference, our parameters of interest $\theta^* \in \mathbb{R}^d$ is typically large, for most cases, larger than the sample sizes n. What will be unique for high dimensional inference?

The first thing is that instead of testing single hypothesis, we will test multiple hypotheses. For example, let each entry of θ^* represent the effect of each gene and we want to select which genes are effective from our data. Instead of testing a single gene, we need to test millions or even trillions of genes. We need to consider the following multiple hypotheses

$$H_{0j} : \theta_j^* \in \Theta_0 \quad v.s. \quad H_{1j} : \theta_j^* \in \Theta_1 \qquad \forall j = 1, ..., N.$$

Conducting the multiple hypotheses is much harder than a single hypothesis due to the multiplicity. Here is a story from John Tukey.

A young psychologist administers 250 hypothesis tests as part of a research project, and finds that, 11 were significant at the 5% level. The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher (Tukey himself?) suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance, because

$$250 \times 5\% = 12.5.$$

In that sense, finding only 11 significant results is actually somewhat disappointing!

Tukey's story tells us the lesson that we should be cautious when conducting multiple hypothesis. The following criterion would be one way to control the multiplicity in the hypotheses.

15.1 Definition (Family-wise error rate). *The family-wise error rate (FWER) is the probability that making at least one type I error in the multiple hypotheses*

$$\textit{FWER} = \mathbb{P}(\textit{ Reject any true } H_{0j}) = \mathbb{P}(\psi_j = 1 \textit{ for some } j \in H_0).$$

Family-wise error rate might be too stringent that we are not allowed to make any type I errors. Another widely used criterion is called the false discovery rate.

15.2 Definition (False discovery rate). *Define the false discovery proportion (FDP) as the proportion of false rejections among all rejections:*

$$\text{FDP} = \sum_{j \in H_0} \frac{\psi_j}{\max\{\sum_{j=1}^d \psi_j, 1\}},$$

where we add 1 to avoid the denominator being 0. The false discovery rate is the expected FDP, i.e., $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$.

We will talk about how to conduct multiple hypotheses in the following lectures.

15.3 Asymptotic Normality of Least Squares

We first review the important theoretical results for inference. The corner stone of inference is the central limit theorem.

15.3 Theorem (Central Limit Theorem). *Let* $X_1, \ldots, X_n, \ldots$ *be i.i.d. random variables with* $\mathbb{E}(X_i) = \mu$ *and Var*$(X_i) = \sigma^2 < \infty$ *for* $i = 1, 2, \ldots$. *The sample mean* $\bar{X}_n = (X_1 + \ldots + X_n)/n$ *converges in distribution to the normal distribution. In particular, we have*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0, 1).$$

A typical inference pipeline for hypothesis: first we construct the statistics T for H_0 and then we want to show that T can be decomposed as

$$T = \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g(X_i) - \mathbb{E}[g(X)])}_{\text{Leading term}} + \underbrace{o_P(1)}_{\text{Remainder}}$$

where the leading term yields the asymptotically normality by CLT and the remainder term is small enough so it is ignorable. The following theorem is useful when we have such decomposition.

15.4 Theorem (Slutsky's Theorem). *If a random variable* $X_n \xrightarrow{d} X$ *and another random variable* $Y_n \xrightarrow{P} c$ *where* c *is a constant, then*

$$X_n + Y_n \xrightarrow{d} X + c, X_n Y_n \xrightarrow{d} cX, \text{ and } X_n / Y_n \xrightarrow{d} X/c.$$

Now we are ready the derive the asymptotic normality of the ordinary least squares.

15.5 Theorem (Asymptotic Normality for OLS). *The linear model* $Y = \mathbb{X}\beta^* + \epsilon$ *has the design matrix* $\mathbb{X} \in \mathbb{R}^{n \times d}$ *with* $\mathrm{rank}(\mathbb{X}) = d \leq n$ *and the i.i.d. centered noise Var*$(\varepsilon_i) = \sigma^2$. *The ordinary least square estimator is*

$$\widehat{\beta}^{\mathrm{LS}} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top Y.$$

Denote sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\mathbb{X}^\top\mathbb{X}$*, then we have*

$$\begin{aligned}\sqrt{n}(\widehat{\beta}^{\mathrm{LS}} - \beta^*) &= \sqrt{n}[(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top(\mathbb{X}\beta^* + \varepsilon) - \beta^*] \\ &= \sqrt{n}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\varepsilon \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\Sigma}^{-1}X_i\varepsilon_i \xrightarrow{d} N(0, \sigma^2\widehat{\Sigma}^{-1}).\end{aligned}$$

Notice that we assume the fixed design so the asymptotic normality results above is conditioned on $\mathbb{X}$*.*

Lecture 16

Debiased Lasso

16.1 Debiased Lasso

In this lecture, we aim to conduct inference for high dimensional linear model. Recall the sparse linear model $Y = \mathbb{X}\beta^* + \varepsilon$, where $\mathbb{X} \in \mathbb{R}^{n\times d}$ and $\|\beta^*\|_0 \le s$. We aim to derive the confidence interval for the Lasso estimator

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

By the first order optimality condition, we have

$$-\frac{1}{n}\mathbb{X}^\top(Y - \mathbb{X}\widehat{\beta}) + \lambda z = 0, \text{ where } z \in \partial\|\widehat{\beta}\|_1. \tag{16.1}$$

Plug the model $Y = \mathbb{X}\beta^* + \varepsilon$ into (16.1) and denote the sample covarianace matrix $\widehat{\Sigma} = \mathbb{X}^\top\mathbb{X}/n$, we have

$$\widehat{\Sigma}(\widehat{\beta} - \beta^*) + \lambda z = \frac{1}{n}\mathbb{X}^\top\varepsilon. \tag{16.2}$$

We expect to "inverse" the sample covariance matrix above, however, when $d > n$, $\widehat{\Sigma}$ is singular. Our idea is that suppose we have a matrix $\widehat{\Theta}$ which is close enough to the inverse, i.e., $\widehat{\Theta}\widehat{\Sigma} \approx I$, we can reveal the parameter in (16.2). Multiplying $\widehat{\Theta}$ on both sides of (16.2) and rearranging the equation, we have

$$\sqrt{n}(\widehat{\beta} - \beta^* + \underbrace{\lambda\widehat{\Theta}z}_{\text{Bias}}) = \underbrace{\frac{1}{\sqrt{n}}\widehat{\Theta}\mathbb{X}^\top\varepsilon}_{\text{Leading term}} + \underbrace{\sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta^*)}_{\text{Remainder}}. \tag{16.3}$$

We can see that the subgradient in the above equation is a bias and we successfully decompose the right hand side into the leading term and remainder term. By (16.1), we have

$\lambda z = \frac{1}{n}\mathbb{X}^\top(Y - \mathbb{X}\widehat{\beta})$ and plug it into (16.3), then

$$\sqrt{n}\Big(\widehat{\beta} + \frac{1}{n}\widehat{\Theta}\mathbb{X}^\top(Y - \mathbb{X}\widehat{\beta}) - \beta^*\Big) = \frac{1}{\sqrt{n}}\widehat{\Theta}\mathbb{X}^\top\varepsilon + \sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta^*). \tag{16.4}$$

We can define the **debiased Lasso**

$$\widehat{\beta}^d = \widehat{\beta} + \frac{1}{n}\widehat{\Theta}\mathbb{X}^\top(Y - \mathbb{X}\widehat{\beta})$$

and then we have

$$\sqrt{n}(\widehat{\beta}^d - \beta^*) = \frac{1}{\sqrt{n}}\widehat{\Theta}\mathbb{X}^\top\varepsilon + \sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta^*).$$

We have following asymptotic normality of the debiased Lasso.

16.5 Theorem (Asymptotic normality of debiased Lasso). *The linear model $Y = \mathbb{X}\beta^* + \varepsilon$ has the i.i.d. random design $\mathbb{X} = (X_1, \ldots, X_n)^\top$ with positive definite covariance $Cov(X_i) = \Sigma$ and the i.i.d. centered noise has $Var(\varepsilon_i) = \sigma^2$. We assume both X_{ij} for all $j = 1, \ldots, d$ and ε_i are subGaussian. Suppose we can find an estimator $\widehat{\Theta}$ of the precision matrix $\Theta = \Sigma^{-1}$ satisfying the following conditions*

C1. $\|\widehat{\beta} - \beta^*\|_1 = O_P(s\sqrt{\log d/n})$;
C2. $\|\widehat{\Sigma}\widehat{\Theta} - I\|_{\max} = O_P(\sqrt{\log d/n})$;
C3. $\max_{1\le j\le d}\|\widehat{\Theta}_j - \Theta_j\|_1 = O_P(s\sqrt{\log d/n})$.

If $s\log d/\sqrt{n} = o(1)$, we have

$$\sqrt{n}(\widehat{\beta}^d_j - \beta^*_j) \xrightarrow{d} N(0, \Theta_{jj}), \text{ for all } j = 1, \ldots, d.$$

Proof. By (16.4), we have

$$\sqrt{n}(\widehat{\beta}^d - \beta^*) = \text{ Leading Term} + \text{Remainder},$$

where Leading Term $= \frac{1}{\sqrt{n}}\widehat{\Theta}\mathbb{X}^\top\varepsilon$ and Remainder $= \sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta^*)$. We first bound the remainder term by

$$\begin{aligned}\|\text{Remainder}\|_\infty &= \|\sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta^*)\|_\infty \\ &\le \sqrt{n}\|\widehat{\Sigma}\widehat{\Theta} - I\|_{\max}\|\widehat{\beta} - \beta^*\|_1 = O_P(s\log d/\sqrt{n}) = o_P(1),\end{aligned}$$

where in the first inequality, we use the matrix Hölder inequality $\|Ax\|_\infty \le \|A\|_{\max}\|x\|_1$ and in the first O_P, we use the conditions C1 and C2. We decompose the leading term as

$$\text{Leading Term } = \frac{1}{\sqrt{n}}\widehat{\Theta}\mathbb{X}^\top\varepsilon = \underbrace{\frac{1}{\sqrt{n}}\Theta\mathbb{X}^\top\varepsilon}_{L_1} + \underbrace{\frac{1}{\sqrt{n}}(\widehat{\Theta}-\Theta)\mathbb{X}^\top\varepsilon}_{L_2}.$$

We have

$$\begin{aligned}\|L_2\|_\infty &= \|\frac{1}{\sqrt{n}}(\widehat{\Theta}-\Theta)\mathbb{X}^\top\varepsilon\|_\infty \\ &\le \sqrt{n}\max_{1\le j\le d}\|\widehat{\Theta}_j-\Theta_j\|_1\|\frac{1}{n}\mathbb{X}^\top\varepsilon\|_\infty = O_P(s\log d/\sqrt{n}) = o_P(1),\end{aligned}$$

where we use Hölder inequality in the first inequality and apply C3 and Corollary 8.6 in Lecture 8. By CLT, the first term

$$L_1 = \frac{1}{\sqrt{n}}\Theta\mathbb{X}^\top\varepsilon \xrightarrow{d} N(0,\Theta).$$

As $\sqrt{n}(\widehat{\beta}^d-\beta^*) = L_1 + L_2 + \text{Remainder}$ and $\|L_2+\text{Remainder}\|_\infty = o_P(1)$, by Slutsky's theorem, we have

$$\sqrt{n}(\widehat{\beta}^d_j-\beta^*_j) \xrightarrow{d} N(0,\Theta_{jj})$$

for all $j=1,\dots,d$. □

There are a few important observations from Theorem 16.5.

- **Regularization**: We did not use the fact that the regularization is the ℓ_1-norm in the proof. Notice that if we change the regularization to other penalties, (16.1) still holds along with the remaining derivations. Theorem 16.5 is true for general penalties.
- **Estimator**: If we change the penalty, the only condition on the new estimator $\widehat{\beta}$ we need to verify is C1 in Theorem 16.5. We will check it for Lasso soon. If the estimator using other penalties satisfies C1, we will still have the asymptotic normality.
- $\widehat{\Theta}$: Notice $\widehat{\Theta}$ is irrelevant to the estimation. We need it only because we want to derive asymptotic normality. So as long as you can find some estimator $\widehat{\Theta}$ satisfying C2 and C3, Theorem 16.5 holds.

We first check C1 in Theorem 16.5 which is related to the rate of Lasso.

16.6 Theorem (ℓ_1-norm rate of Lasso). *Under the same conditions as the Theorem 8.4 in Lecture 8, we have*

$$\|\widehat{\beta}-\beta^*\|_1 = O_P(s\sqrt{\log d/n}).$$

Proof. Theorem 8.4 shows the ℓ_2-norm of Lasso $\|\widehat{\beta} - \beta^*\|_2 = O_P(\sqrt{s \log d/n})$. In the proof of Theorem 8.4, we show that $\Delta = \widehat{\beta} - \beta^* \in \mathbb{C}_3(S) = \{\Delta | \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$. Therefore, we have

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{s}\|\Delta\|_2 = O_P(s\sqrt{\log d/n}).$$

□

16.2 CLIME Estimator

In this section, we will provide a concrete estimator $\widehat{\mathbf{\Theta}}$ satisfying C2 and C3 in Theorem 16.5.

The **constrained ℓ_1-minimization for inverse matrix estimation (CLIME)** estimates the j-th column of the precision matrix Θ_j via

$$\widehat{\Theta}_j = \arg\min_{\beta} \|\beta\|_1 \text{ s.t. } \|\widehat{\Sigma}\beta - e_j\|_\infty \leq \lambda \tag{16.7}$$

for $j = 1, \ldots, d$, where λ is the tuning parameter, $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ is the j-th canonical basis.

In history, the CLIME estimator was proposed to estimate a sparse precision matrix, i.e., the Gaussian graphical model (See Section 9.2.2 in Lecture 9). The ℓ_1-norm objective function finds a sparse column of $\widehat{\Theta}$ and the constraint is based on $\widehat{\Sigma}\widehat{\Theta}_j \approx e_j$ as we expect $\widehat{\Theta} \approx \Theta = \Sigma^{-1}$ and $\widehat{\Theta}\widehat{\Sigma} \approx I$.

In order to make sure that (16.7) is a reasonable estimator, we first need to check if the constraint is feasible.

16.8 Lemma (Feasibility of CLIME). *If $\max_j \|\Theta_j\|_1 < \infty$ and we choose $\lambda = C\sqrt{\log d/n}$ for some sufficiently large constant C, then $\max_j \|\widehat{\Sigma}\Theta_j - e_j\|_\infty \leq \lambda$ with probability at least $1 - 1/d$, i.e., Θ_j is a feasible solution of the constraint of CLIME for all $1 \leq j \leq d$.*

Proof. By definitions, we have $\Sigma\Theta = I$ and thus

$$\begin{aligned} \max_j \|\widehat{\Sigma}\Theta_j - e_j\|_\infty &= \max_j \|(\widehat{\Sigma} - \Sigma)\Theta_j\|_\infty \\ &\leq \|\widehat{\Sigma} - \Sigma\|_{\max} \max_j \|\Theta_j\|_1 = O_P(\sqrt{\log d/n}), \end{aligned} \tag{16.9}$$

where in the last equality we use the condition $\max_j \|\Theta_j\|_1 < \infty$ and the result $\|\widehat{\Sigma} - \Sigma\|_{\max} = O_P(\sqrt{\log d/n})$ showed in the proof of Proposition 8.7 in Lecture 8. □

By the definition of CLIME, $\widehat{\Theta}_j$ must be a feasible solution to (16.7) for all $j = 1, \ldots, d$. Therefore, if we choose $\lambda = C\sqrt{\log d/n}$ for some sufficiently large constant C, we have C2 checked that

$$\|\widehat{\Sigma}\widehat{\Theta} - I\|_{\max} = \max_j \|\widehat{\Sigma}\widehat{\Theta}_j - e_j\|_\infty \le \lambda = O_P(\sqrt{\log d/n}).$$

What remained to be shown is C3 in Theorem 16.5 and it follows from the following theorem on the ℓ_1-norm rate of CLIME.

16.10 Theorem (ℓ_1-norm rate of CLIME). *If $\max_j \|\Theta_j\|_1 < \infty$ and $s = \max_j \|\Theta_j\|_0$ and we choose $\lambda = C\sqrt{\log d/n}$ for some sufficiently large constant C, then*

$$\max_j \|\widehat{\Theta}_j - \Theta_j\|_1 = O_P(s\sqrt{\log d/n}).$$

Proof. Define $\Delta_j = \widehat{\Theta}_j - \Theta_j$ and S is the support of Θ_j. As we show in Lemma 16.8 that Θ_j is a feasible solution to (16.7), so we have

$$0 \le \|\Theta_j\|_1 - \|\widehat{\Theta}_j\|_1 = \|\Theta_{jS}\|_1 - \|\widehat{\Theta}_{jS}\|_1 - \|\widehat{\Theta}_{jS^c}\|_1 \le \|\Delta_{jS}\|_1 - \|\Delta_{jS^c}\|_1.$$

Therefore, $\|\Delta_{jS^c}\|_1 \le \|\Delta_{jS}\|_1$. The proof is exactly same as how we prove the Lasso error vector belongs to the cone. Now we first derive the ℓ_∞-norm of CLIME. By Lemma 16.8, we choose sufficiently large C for $\lambda = C\sqrt{\log d/n}$, so we have

$$\max_j \|\Theta_j\|_1 \|\widehat{\Sigma} - \Sigma\|_{\max} \le \lambda \tag{16.11}$$

Therefore, we have

$$\|\widehat{\Sigma}(\widehat{\Theta} - \Theta)\|_{\max} \le \|\widehat{\Sigma}\widehat{\Theta} - I\|_{\max} + \|\widehat{\Sigma}\Theta - I\|_{\max} \le 2\lambda, \tag{16.12}$$

where the second inequality is due to (16.9) and the definition of CLIME. We further have

$$\begin{aligned}
\|\Sigma(\widehat{\Theta} - \Theta)\|_{\max} &\le \|\widehat{\Sigma}(\widehat{\Theta} - \Theta)\|_{\max} + \|(\widehat{\Sigma} - \Sigma)(\widehat{\Theta} - \Theta)\|_{\max} \\
&\le 2\lambda + \max_j \|\widehat{\Theta}_j - \Theta_j\|_\infty \|\widehat{\Sigma} - \Sigma\|_{\max} \\
&\le 2\lambda + 2\max_j \|\Theta_j\|_\infty \|\widehat{\Sigma} - \Sigma\|_{\max} \le 4\lambda,
\end{aligned}$$

where the second inequality is by (16.12) and Hölder inequality, the third inequality is by the triangle inequality and and $\|\Theta_j\|_1 \ge \|\widehat{\Theta}_j\|_1$ and the last inequality is by (16.11).

So we have the ℓ_∞-norm rate

$$\|\widehat{\Theta} - \Theta\|_{\max} \le \max_j \|\Theta_j\|_\infty \|\Sigma(\widehat{\Theta} - \Theta)\|_{\max} \le 4\max_j \|\Theta_j\|_\infty \lambda = O_P(\sqrt{\log d/n}).$$

As $\|\Delta_{jS^c}\|_1 \leq \|\Delta_{jS}\|_1$ for all $j = 1, \ldots, d$, we have

$$\begin{aligned}\max_j \|\Delta_j\|_1 &= \max_j \|\Delta_{jS}\|_1 + \max_j \|\Delta_{jS^c}\|_1 \\ &\leq \max_j 2\|\Delta_{jS}\|_1 \leq 2s \max_j \|\Delta_{jS}\|_\infty = 2s\|\widehat{\Theta} - \Theta\|_{\max} = O_P(s\sqrt{\log d/n}),\end{aligned}$$

which completes the proof. □

16.2.1 Comparison of Lasso and Debiased Lasso

In order to have the asymptotic normality for high dimensional linear model, we need to impose much stronger assumptions than the estimation. We need (1) the precision matrix of the design $\Theta = \Sigma^{-1}$ being sparse and (2) the scaling condition $\frac{s \log d}{\sqrt{n}} = o(1)$, neither of which is required to guarantee the consistency of Lasso.

16.3 General M-Estimator

In this section, we generalize the debiasing method to general high dimension M-estimators. For the loss $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, X_i)$, we assume that $\mathbb{E}[\nabla_\theta \ell(\theta^*, X)] = 0$, i.e., the truth θ^* is the minimizer of the population loss. We first consider the unregularized M-estimator

$$\widehat{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \ell(\theta, X_i).$$

For example, for the logistic regression,

$$\ell(\theta, (Y, X)) = \log\big(1 + \exp(-YX^\top\theta)\big).$$

Applying the first order optimality conditions, we would have

$$0 = \nabla_\theta \mathcal{L}_n(\widehat{\theta}) \approx \nabla_\theta \mathcal{L}_n(\theta^*) + \nabla^2_\theta \mathcal{L}_n(\theta^*)(\widehat{\theta} - \theta^*)\widehat{\theta} - \theta^* \approx -[\nabla^2_\theta \mathcal{L}_n(\theta^*)]^{-1}\nabla_\theta \mathcal{L}_n(\theta^*)$$

Therefore, we have

$$\sqrt{n}(\widehat{\theta} - \theta^*) \approx -\frac{1}{\sqrt{n}}\sum_{i=1}^n [\nabla^2_\theta \mathcal{L}_n(\theta^*)]^{-1}\nabla_\theta \ell(\theta, X_i) \xrightarrow{d} N(0, ABA),$$

where $A = \mathbb{E}[\nabla^2_\theta \ell(\theta^*, X)]^{-1}$, and $B = \mathbb{E}[\nabla_\theta \ell(\theta^*, X)\nabla_\theta \ell(\theta^*, X)^\top]$. By the property of Fisher information, we have $B^{-1} = -A$ and therefore

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, \mathbb{E}[\nabla^2_\theta \ell(\theta^*, X)]^{-1}).$$

When $\ell(\theta, X)$ is the log-likelihood, the asymptotic covariance $\mathbb{E}[\nabla^2_\theta \ell(\theta^*, X)]^{-1}$ is the Fisher information.

Now we turn to the regularized M-estimator

$$\widehat{\theta} = \arg\min_{\theta} \mathcal{L}_n(\theta) + \lambda\|\theta\|_1. \tag{16.13}$$

By the first order optimality conditions,

$$\nabla_\theta \mathcal{L}_n(\widehat{\theta}) + \lambda z = 0.$$

We expand $\nabla_\theta \mathcal{L}_n(\widehat{\theta})$ like above and get

$$\nabla_\theta \mathcal{L}_n(\theta^*) + \nabla^2_\theta \mathcal{L}_n(\theta^*)(\widehat{\theta} - \theta^*) + \lambda z = 0.$$

Following the same strategy of Lasso, we aim to approximate the inversion of $\nabla^2_\theta \mathcal{L}_n(\theta^*)$ via $\widehat{\Theta}$ and

$$\sqrt{n}(\widehat{\theta} - \widehat{\Theta}\nabla_\theta \mathcal{L}_n(\widehat{\theta}) - \theta^*) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\Theta}\nabla_\theta \ell(\theta, X_i) + \sqrt{n}(I - \widehat{\Theta}\nabla^2_\theta \mathcal{L}_n(\theta^*))(\widehat{\theta} - \theta^*).$$

The debiased M-estimator of (16.13) is

$$\widehat{\theta}^d = \widehat{\theta} - \widehat{\Theta}\nabla_\theta \mathcal{L}_n(\widehat{\theta}),$$

where we estimate the j-th column of the inverse of Hessian matrix via

$$\widehat{\Theta}_j = \arg\min_{\beta} \|\beta\|_1 \text{ s.t. } \|\nabla^2_\theta \mathcal{L}_n(\widehat{\theta})\beta - e_j\|_\infty \le \lambda,$$

for $j = 1, \ldots, d$. Under the assumption that the Fisher information $\mathbb{E}[\nabla^2_\theta \ell(\theta^*, X)]^{-1}$ is sparse and some other regularity conditions, we have

$$\sqrt{n}(\widehat{\theta}^d_j - \theta^*_j) \xrightarrow{d} N(0, (\mathbb{E}[\nabla^2_\theta \ell(\theta^*, X)]^{-1})_{jj}).$$

The assumption that the Fisher information $\mathbb{E}[\nabla^2_\theta \ell(\theta^*, X)]^{-1}$ is sparse is very hard to check in practice (for example in logistic regression), an open problem in high dimensional inference is that if it is possible to derive the asymptotic normality of general regularized M-estimator without such sparsity assumption.

Lecture 17

Multiple Hypotheses

17.1 Conformal Inference

The conformal inference aims to build confidence intervals for predictions without any unnecessary assumptions, especially those about models. Given i.i.d. random pairs $(X_i, Y_i) \sim P$, our goal is to build a confidence interval $\widehat{C}_\alpha$ for the prediction given the new pair (X_{n+1}, Y_{n+1}), which has

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha$$

In order to explain the idea of conformal inference, we start with Y_i's themselves. Denote

$$\widehat{q}_n = \text{Quantile}(1 - \alpha, \{Y_i\}_{i=1}^n)$$
$$\mathbb{P}(Y_{n+1} \leq \widehat{q}_n) \approx 1 - \alpha$$

By symmetricity, we could notice that the rank of Y_{n+1} among $Y_1, ..., Y_{n+1}$ follows a uniform distribution Unif$\{1, 2, ..., n+1\}$. We could construct

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n} 1\{Y_i \leq y\} + \frac{1}{n+1}$$

And by the symmetry, we have

$$\mathbb{P}\left(\pi(Y_{n+1}) \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1}\right) \geq 1 - \alpha$$

We can invert the $\pi(\cdot)$ above and solve Y_{n+1} as

$$\mathbb{P}(Y_{n+1} \leq \widetilde{q}_n) \geq 1 - \alpha, \text{ where } \widetilde{q}_n = \text{Quantile}(1 - \alpha, \{Y_i\}_{i=1}^n \bigcup \{\infty\}).$$

However, the above confidence interval has not involved X_{n+1}. Let $\widehat{f}(x)$ be some estimator of $\mathbb{E}[Y|X = x]$ using the data $\{(X_i, Y_i)\}_{i=1}^n$.

Denote the absolute residual $R_i = |Y_i - \widehat{f}(X_i)|$ and we may use a similar idea as above by considering

$$\widetilde{q}_n = \text{Quantile}(1-\alpha, \{R_i\}_{i=1}^n \bigcup \{\infty\})$$

One possible confidence interval could be

$$\widehat{C}_n(x) = [\widehat{f}(x) - \widehat{q}_n, \widehat{f}(x) + \widehat{q}_n].$$

However, such confidence interval is not honest. The reason is that R_i depends on $\widehat{f}$, which further depends on $\{(X_i, Y_i)\}_{i=1}^n$. As a result, R_i could not be independent and identically distributed. To handle this issue, we would estimate $\widehat{f}_{n,(x,y)}$ using the data

$$\{(X_1, Y_{n+1}), (X_2, Y_{n+1}), ..., (X_n, Y_{n+1}), (x, y)\}$$

Under this framework, we define the new residuals

$$\begin{aligned} R_i^{(x,y)} &= |Y_i - \widehat{f}_{n,(x,y)}(X_i)| \text{ if } i = 1, 2, ..., n, \\ R_{n+1}^{(x,y)} &= |y - \widehat{f}_{n,(x,y)}(x)|. \end{aligned}$$

By symmetry, now we have the rank of $R_{n+1}^{(x,y)}$ among $R_i^{(x,y)}, i \in \{1, 2, ..., n+1\}$ follows Unif$(1, 2, ..., n+1)$. So we can construct the confidence interval via

$$\begin{aligned} \widetilde{q}_n &= \text{Quantile}(1-\alpha, \{R_i^{(x,y)}\}_{i=1}^n \bigcup \{\infty\}) \\ \widehat{C}_n(x) &= [\widehat{f}(x) - \widehat{q}_n, \widehat{f}(x) + \widehat{q}_n] \end{aligned}$$

We can show that

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

The reason why we do so is that (1) we would not need any unnecessary assumptions, (2) we would not need to use any asymptotic properties, (3) we could avoid overfitting since for most of the times the naive CI would undercover, and (4) such R_i could be generalized in several more sophisticated frameworks.

17.2 Multiple Hypotheses

Here we are discussing the hypotheses $\{H_{0i}\}_{i=1}^N$, with

$$H_{0i} : \theta_i^* = \theta_{i0}; \quad H_{1i} : \theta_i^* \neq \theta_{i0}.$$

In some practical scenarios the number of assumptions, N could be very large. N may be on the level of several millions on GWAS when our goal is to detect significant genes. We would hope to find a way to control the family-wise error rate (FWER), which could be written as

$$\text{FWER} = \mathbb{P}(\text{There is at least one type-I error}) \leq \alpha.$$

17.2.1 P-Values

For a single null hypothesis, we could either construct a statistic T and a corresponding confidence region such that the rejection region R_α holds that

$$P_0(T \in R_\alpha) = \alpha.$$

We can define the p-value as

$$p(T) = \inf_\alpha P_0(T \in R_\alpha)$$

Since $p(T) \sim \text{Unif}[0,1]$ under the null hypothesis, we will tend to reject the null hypothesis H_0 if $p(T)$ is smaller than the significance level α.

17.2.2 Bonferroni Correction

Here we use the notations given above and suppose that we have p_i as the p-value of a null hypothesis H_{i0}. For a significance level α, the **Bonferroni correction** rejects H_{i0} if $p_i \leq \frac{\alpha}{N}$. In this case, we would have

$$\begin{aligned}\text{FWER} &= \mathbb{P}(\text{There is at least one type-I error}) \\ &\leq \mathbb{P}\left(\bigcup_{i=1}^N \{\psi_i = 1, i \in H_0\}\right) \leq \sum_{i=1}^N \mathbb{P}(\psi_i = 1, i \in H_0) = \sum_{i=1}^N \frac{\alpha}{N} = \alpha.\end{aligned}$$

The Bonferroni is usually **too conservative**, as the step we use union bound could make the inequality very loose. In fact, we have not utilized the dependency of p_i in Bonferroni correction.

For example, we could consider a debiased Lasso model $Y = \mathbb{X}\beta^* + \epsilon$, and have $H_{0j} : \beta_j^* = 0$ as our multiple hypotheses. According to the previous lecture, we have that given some regularization conditions, the debiased Lasso estimator $\widehat{\beta}^d$ would have

$$\sqrt{n}(\widehat{\beta}_j^d - \beta_j^*) \xrightarrow{d} N(0, \Theta_{jj})$$

for any arbitrary j, and Θ is the precision matrix. If we consider Bonferroni correction, we could calculate p_j as

$$p_j = 1 - 2\Phi\left(\frac{\sqrt{n}|\widehat{\beta}_j^d|}{\sqrt{\widehat{\Theta}_{jj}}}\right)$$

And reject H_{0j} if $p_j \leq \frac{\alpha}{d}$. This will be conservative when the dimension d is large.

17.2.3 Maximal Statistic

An alternative to Bonferroni correction is that we calculate the maximal statistic as

$$T = \max_{1 \leq j \leq d} T_j = \max_{1 \leq j \leq d} \frac{\sqrt{n}|\widehat{\beta}_j^d|}{\sqrt{\widehat{\Theta}_{jj}}}$$

If we can estimate the quantile of the maximal statistic $\widehat{C}(1-\alpha)$ such that

$$\mathbb{P}(T > \widehat{C}(1-\alpha)) = \alpha$$

We would reject H_{0j} if $T_j > \widehat{C}(1-\alpha))$. Then the FWER becomes

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(\text{There is at least one Type I error}) \\
&\leq \mathbb{P}(\bigcup_{j=1}^{d} \{T_j > \widehat{C}(1-\alpha)\}) \\
&= 1 - \mathbb{P}(T_j \leq \widehat{C}(1-\alpha) \text{ for any } j) = 1 - \mathbb{P}(\max_j T_j \leq \widehat{C}(1-\alpha)) = \alpha
\end{aligned}$$

Now we could save much p-values by using the maximal statistic compared to Bonferroni. In the next lecture, we are going to discuss how to estimate the quantile of maximal statistics.

Lecture 18

False Discovery Rate

18.1 Gaussian Multiplier Bootstrap

We will continue the discussion on control family-wise error rate via the maximal statistic. Given hypotheses $\{H_{0i}\}_{i=1}^N$, for each H_{0j}, we have a statistic T_j such that we for a single hypothesis, we will reject H_{0j} if $T_j \geq q_\alpha$ where $q_\alpha = \arg\min_t \mathbb{P}_{H_0}(|T_j| > t) \leq \alpha$. In the previous lecture, we introduced the maximal statistic $T = \max_j |T_j|$. If we can estimate the quantile of T:

$$\widehat{C}(1-\alpha) = \text{Quantile}(\max_j |T_j|, 1-\alpha), \tag{18.1}$$

we reject H_{0j} if $T_j \geq \widehat{C}(1-\alpha)$ for all $1 \leq j \leq N$. We showed that such maximal statistic approach obtains FWER $< \alpha$. Here, we consider how to estimate the $(1-\alpha)\times 100\%$ quantile of the $\max_j T_j$. Let us start with a simple case: the statistics $T_1, \ldots, T_N$ are i.i.d. $N(0,1)$, then we solve the following equation

$$\mathbb{P}\big(\max_j |T_j| > t\big) = \mathbb{P}\big(\bigcup_{j=1}^d \{|T_j| > t\}\big) = 1 - \mathbb{P}\big(\bigcap_{j=1}^d \{|T_j| \leq t\}\big) = 1 - \prod_{j=1}^d [1 - 2\Phi(1-t)] = \alpha,$$

where Φ denotes the cdf of the standard normal distribution. So the $(1-\alpha)\times 100\%$ quantile of the $T = \max_j T_j$ is

$$\widehat{C}(1-\alpha) = 1 - \Phi^{-1}\big([1-(1-\alpha)^{1/d}]/2\big).$$

The case where the T_j are dependent and only asymptotically normal is more challenging. Typically, such T_j can be decomposed as

$$T_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f_j(X_i) - \mathbb{E}[f_j(X)]\big) + o_P(1), \tag{18.2}$$

where $f_j : \mathbb{R}^d \to \mathbb{R}$ is some deterministic function and the first term on the right hand side is the leading term contributing to asymptotic normality of the statistic. For example, in

Lecture 16, we showed that the debiased Lasso has the decomposition

$$\sqrt{n}(\widehat{\beta}_j^d - \beta_j^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Theta_j^\top X_i \varepsilon_i + o_P(1).$$

So we can conduct the hypotheses $H_{0j} : \beta_j^* = 0$ by the statistic $\sqrt{n}|\widehat{\beta}_j^d|$[5] In order to estimate the quantile of $\max_j T_j$, our idea is to esimate the quantile of the maximal of the leading term:

$$T_0 = \max_j \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f_j(X_i) - \mathbb{E}[f_j(X)]\big).$$

Denote $f(X) = (f_1(X), \ldots, f_d(X))^\top \in \mathbb{R}^d$. By central limiting theorem, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f(X_i) - \mathbb{E}[f(X)]\big) \xrightarrow{d} N(0, \text{Cov}(f(X))).$$

Let $Z \sim N(0, \text{Cov}(f(X)))$ and we will expect that if a random vector converge in distribution to a normal vector, then the distribution of their maximal statistics should be close as well, i.e.,

$$T_0 = \max_j \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f_j(X_i) - \mathbb{E}[f_j(X)]\big) \overset{d}{\approx} \max_j Z_j.$$

So the quantile $\widehat{C}(1-\alpha)$ in (18.1) has

$$\widehat{C}(1-\alpha) \approx \text{Quantile}(\max_j |T_{0j}|, 1-\alpha) \approx \text{Quantile}(\max_j |Z_j|, 1-\alpha), \tag{18.3}$$

where we can estimate the last term as long as we can sample the multivariate normal $N(0, \text{Cov}(f(X)))$. However, in most cases, we do not know the covariance. For example, in the debiased Lasso,

$$f(X, \varepsilon) = \Theta X \varepsilon \text{ and } \text{Cov}(f(X, \varepsilon)) = \sigma^2 \Theta,$$

where $\text{Var}(\varepsilon) = \sigma^2$. One solution is to plug in the CLIME estimator $\widehat{\Theta}$ and residual variance estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^2 (Y_i - X_i\widehat{\beta})^2$, where $\widehat{\beta}$ is the Lasso estimator. We can simulate $\widetilde{Z} \sim N(0, \widehat{\sigma}^2\widehat{\Theta})$ and estimate the quantile $\widehat{C}(1-\alpha) \approx \max_j \text{Quantile}(\max_j |\widetilde{Z}_j|, 1-\alpha)$. However, it is not always possible to have a closed form the covariance $\text{Cov}(f(X))$. We will provide a more general approach to estimate the quantile of T_0 via the so called Gaussian multiplier bootstrap.

Let $\xi_1, \ldots, \xi_n$ be i.i.d. samples of $N(0,1)$, we define the **Gaussian multiplier bootstrap** statistic as

$$T_\xi = \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f(X_i) - \mathbb{E}[f(X)]\big)\xi_i.$$

[5] We do not standardize the statistic only for notation simplicity. The standardized t-statistc $\sqrt{n}\widehat{\Theta}_{jj}^{-1/2}|\widehat{\beta}_j^d|$ also has the decomposition (18.2).

As $T_\xi \big| \{X_i\}_{i=1}^n \sim N(0, \widehat{\text{Cov}}(f(X)))$, where the sample covariance

$$\widehat{\text{Cov}}(f(X)) = \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - \mathbb{E}[f(X)]\big)\big(f(X_i) - \mathbb{E}[f(X)]\big)^\top \approx \text{Cov}(f(X)).$$

we will expect for the multivariate normal $Z \sim N(0, \text{Cov}(f(X)))$, there is

$$\arg\min_t \{\mathbb{P}(\max_j |T_{\xi,j}| \geq t | \{X_i\}_{i=1}^n) \leq \alpha\} \approx \text{Quantile}(\max_j |Z_j|, 1-\alpha).$$

Combining with (18.3), we estimate the quantile of T_0 via

$$\widehat{C}(1-\alpha) \approx \arg\min_t \{\mathbb{P}(\max_j |T_{\xi,j}| \geq t | \{X_i\}_{i=1}^n) \leq \alpha\}.$$

However, recall that for the debiased Lasso, $f(X, \varepsilon) = \Theta X \varepsilon$, but we do not know Θ and ε. The following part summarizes the procedure to estimate the quantile when $f(X)$ is unknown.

Let $\xi_1, \ldots, \xi_n$ be i.i.d. samples of $N(0,1)$ and $\widehat{f}(X)$ is a consistent estimate of $f(X) - \mathbb{E}[f(X)]$. We construct the statistic

$$\widehat{T}_\xi = \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{f}(X_i)\xi_i.$$

The quantile of T_0 can be estimated as

$$\widehat{C}(1-\alpha) \approx \arg\min_t \{\mathbb{P}(\max_j |\widehat{T}_{\xi,j}| \geq t | \{X_i\}_{i=1}^n) \leq \alpha\}. \tag{18.4}$$

For the debiased Lasso, we can estimate $f(X_i, \varepsilon_i) = \Theta X_i \varepsilon_i$ via $\widehat{f}(X_i, \varepsilon_i) = \widehat{\Theta} X_i \widehat{\varepsilon}_i$, where $\widehat{\Theta}$ is the CLIME estimator and $\widehat{\varepsilon}_i = Y_i - X_i\widehat{\beta}$. We construct the multiplier Gaussian bootstrap statistic

$$\widehat{T}_\xi = \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\Theta} X_i \widehat{\varepsilon}_i \xi_i.$$

Let $\widehat{C}(1-\alpha)$ be the estimator of the quantile of $\max_j \sqrt{n}(\widehat{\beta}_j^d - \beta_j^*)$ via (18.4). We can reject $H_{0j} : \beta_j^* = 0$ if $\sqrt{n}|\widehat{\beta}_j^d| \geq \widehat{C}(1-\alpha)$ for all $1 \leq j \leq d$ such that FWER $\leq \alpha$.

For more details of Gaussian multiplier bootstrap and the maximal statistics, we refer to a seminal work by Chernozhukov et al. (2013).

18.2 False Discovery Rate: Independent P-values

As we have commented in Lecture 16, the family-wise error rate might be too stringent that we are not allowed to make any type I errors and we define the false discovery rate to make our inference more liberal.

Let $\psi_i = 1$ if we reject H_{0i} and $\psi_i = 0$ otherwise for $1 \le i \le N$. Recall that the false discovery proportion (FDP) and the false discovery rate are

$$\text{FDP} = \frac{\#\text{False Positive}}{\#\text{Rejected Hypotheses}} = \frac{\sum_{i \in H_0} \psi_i}{\max\{\sum_{i=1}^d \psi_i, 1\}} \text{ and } \text{FDR} = \mathbb{E}[\text{FDP}].$$

We aim to find such tests $\{\psi_i\}_{i=1}^N$ such that $\text{FDR} = \mathbb{E}[\text{FDP}] \le \alpha$. Let p_i be the p-value H_{0i} and we first consider the case the p-values $p_1, \dots, p_N$ are independent. We have the following procedure to control the FDR.

We order the p-values as $p_{(1)} \le \cdots \le p_{(N)}$. Let $i_{\max} = \arg\max\{i : p_{(i)} \le \frac{i\alpha}{N}\}$. The **Benjamini-Hochberg** (BH_α) procedure rejects $H_{0(i)}$ if $i \le i_{\max}$.

Two examples of the BH_α procedure are given in Figure 18.1. In the example in the right panel of Figure 18.1, notice that there is an i such that $p_{(i)} > \frac{\alpha i}{N}$ but $i < i_{\max}$. Therefore, we would still reject this H_{0i} based on this procedure. Notice that the BH_α is more liberal than the Bonferroni correction as there is an additional i in the α/N term.

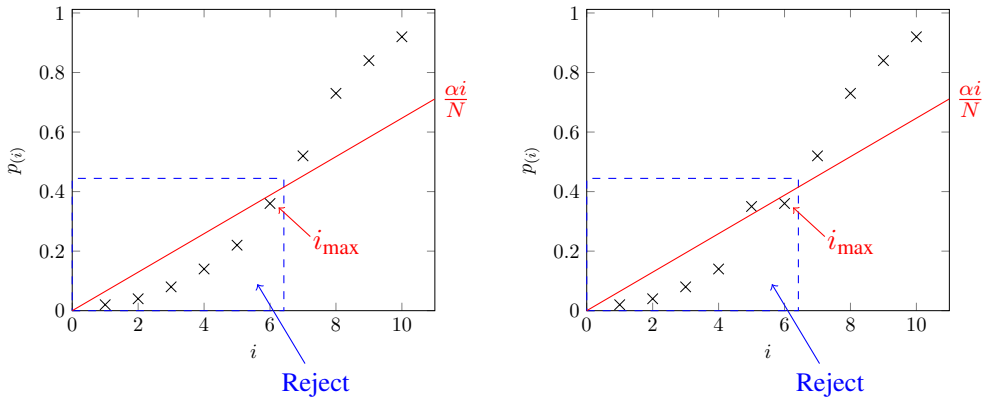


Figure 18.1. *Two examples of the Benjamini-Hochberg procedure. The red line illustrates the threshold for $p_{(i)}$. The blue rectangle illustrates the region rejection.*

The following theorem shows that Benjamini-Hochberg procedure can control FDR for independent p-values.

18.5 Theorem. *Suppose that $p_1, \dots, p_N$ are independent. Then BH_α controls the FDR as*

$$\text{FDR} = \mathbb{E}[\text{FDP}_{\text{BH}_\alpha}] = \frac{N_0}{N}\alpha \le \alpha,$$

where N_0 denotes the number of true H_{0i}'s.

Proof. Denote the number of rejections $R = \sum_{j=1}^{N} \psi_j$, then

$$\mathrm{FDP} = \sum_{i \in H_0} \frac{\psi_i}{R \vee 1}.$$

We claim that

$$\mathbb{E}\big[\frac{\psi_i}{R \vee 1}\big] = \frac{\alpha}{N}, \tag{18.6}$$

as it then follows that

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] = \sum_{i \in H_0} \mathbb{E}\bigg[\frac{\psi_i}{R \vee 1}\bigg] = \frac{N_0 \alpha}{N} \leq \alpha,$$

which completes the proof. So what remains is the prove the claim (18.6). We decompose the ratio inside of the expectation as a summation

$$\frac{\psi_i}{R \vee 1} = \frac{\psi_i}{\big(\sum_{j=1}^{N} \psi_j\big) \vee 1} = \sum_{k=1}^{N} \frac{\psi_i \mathbb{1}(R = k)}{k \vee 1},$$

where for $k = 0$, the summand is zero.

We make the two key observations. First, When $R = k$, $\psi_i = 1$ if and only if $p_i \leq \frac{\alpha k}{N}$ because $p_{(k)} \leq \frac{\alpha k}{N}$ and we reject H_{0i} if $p_i \leq p_{(k)}$.

A challenge in evaluating $\mathbb{E}[\psi_i \mathbb{1}(R = k)]$ is that R and ψ_i are not independent. Consider the following leave-one-out approach. If $p_i \leq \frac{\alpha k}{N}$, then we set p_i to 0, denoted by $p_i \to 0$. Let $R(p_i \to 0)$ denote the number of rejected hypotheses by BH_α using the p-values: $p_1, \ldots, p_{i-1}, 0, p_{i+1}, \ldots, p_N$. From Figure 18.2, our second key observation is that if $p_i \leq \frac{\alpha k}{N}$, then we set p_i to 0 and the number of rejections is the same as R. Namely, $\psi_i \mathbb{1}(R = k) = \psi_i \mathbb{1}(R(p_i \to 0) = k)$.

Let $\mathcal{F}_i = \{p_1, p_2, \ldots, p_{i-1}, p_{i+1}, \ldots, p_N\}$. Then, using the two observations above, we have

$$\begin{aligned}
\mathbb{E}\Big[\frac{\psi_i}{R \vee 1}\Big|\mathcal{F}_i\Big] &= \mathbb{E}\bigg[\sum_{k=1}^{N} \frac{\psi_i \mathbb{1}(R = k)}{k \vee 1}\Big|\mathcal{F}_i\bigg] \\
&= \sum_{k=1}^{N} \mathbb{E}\bigg[\frac{\psi_i \mathbb{1}(R = k)}{k}\Big|\mathcal{F}_i\bigg] \\
&= \sum_{k=1}^{N} \mathbb{E}\bigg[\frac{\mathbb{1}(p_i \leq \frac{\alpha k}{N}) \mathbb{1}(R(p_i \to 0) = k)}{k}\Big|\mathcal{F}_i\bigg] \\
&= \sum_{k=1}^{N} \frac{\frac{\alpha k}{N} \mathbb{1}(R(p_i \to 0) = k)}{k} = \frac{\alpha}{N} \sum_{k=1}^{N} \mathbb{1}(R(p_i \to 0) = k) = \frac{\alpha}{N},
\end{aligned}$$

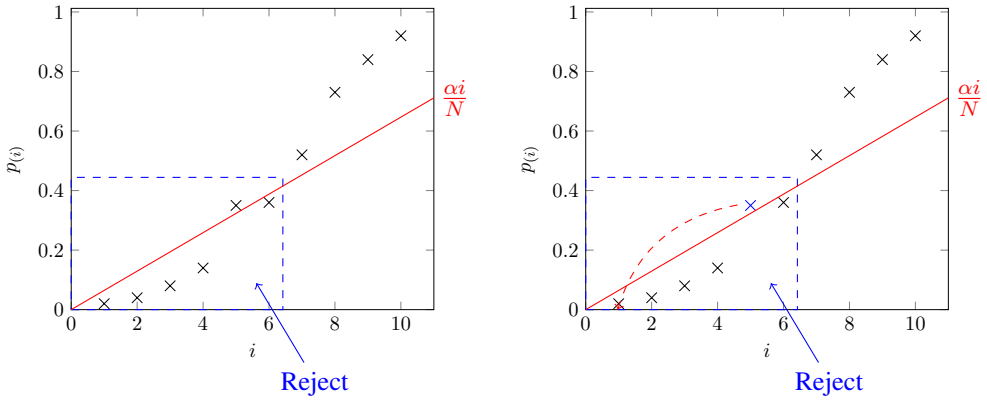


Figure 18.2. *Two examples of the Benhamini-Hochberg procedure. The panel in the right illustrates how setting $p_{(5)}$ to 0 does not affect the rejection region.*

where in the third equality, we use the two key observations and in the fourth equality we use $R(p_i \to 0)$ is irrelevant to p_i. Therefore, by tower property, we have

$$\mathbb{E}\Big[\frac{\psi_i}{R \vee 1}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[\frac{\psi_i}{R \vee 1}\Big|\mathcal{F}_i\Big]\Big] = \frac{\alpha}{N}$$

and we prove the claim (18.6). □

Lecture 19

Knock-Off

19.1 False Discovery Rate: Dependent P-values

We will continue the discussion of controlling the false discovery rate (FDR). When testing null hypotheses $\{H_{0j}\}_{j=1}^d$, recall that the FDR is defined as

$$\text{FDR} = \mathbb{E}\left(\frac{\#\text{False Positives}}{\#\text{Rejected Hypotheses}}\right).$$

Last time, we discussed the case where the p-values corresponding to the $\{H_{0j}\}_{j=1}^d$ were independent. Here, we consider the more challenging case where the p-values are dependent.

Throughout, we consider the following framework. Given a response Y (e.g., phenotype) and features $X_1, \ldots, X_d$ (e.g., SNPs), our goal is to select features X_j's related to Y. Since the X_j's may be correlated, we consider testing the null hypothesis $Y \perp\!\!\!\perp X_j | X_{-j}$. In the example of the linear model $Y = \mathbb{X}\beta + \epsilon$ where $\mathbb{X} \in \mathbb{R}^{n \times d}$, we are interested in testing the null hypothesis $\beta_j = 0$.

19.2 Permutation Test

One popular approach for controlling the FDR in this context is the **permutation test**. Consider the Lasso estimator which we denote as $\widehat{\beta} \leftarrow \text{Lasso}(\mathbb{X}, Y)$. Let $\mathbb{X}_\pi$ denote the matrix obtained by permuting the rows of $\mathbb{X}$ and consider the Lasso estimator based on $\mathbb{X}_\pi$, denoted by $\widetilde{\beta} \leftarrow \text{Lasso}(\mathbb{X}_\pi, Y)$.

The idea behind the permutation test is that we expect $\widehat{\beta} \overset{d}{=} \widetilde{\beta}$. However, one must be careful because this is not always true, which can be illustrated in the following counterexample. Suppose that $Y = X_1 + \epsilon$ where $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$, $\text{Var}(X_1) = \text{Var}(X_2) = 1$ and $\text{Corr}(X_1, X_2) = 1/2$. A straightforward calculation shows that $\mathbb{E}(YX_1) = 1$ and $\mathbb{E}(YX_2) = 0.5$. Letting $X_{2,\pi}$ denote the vector obtained by permuting the entries of X_2, it can be shown that $\mathbb{E}(YX_{2,\pi}) = 0$ which implies $YX_2 \overset{d}{\neq} YX_{2,\pi}$.

19.3 Knock-Off

As the permutation test is not always valid, we may instead consider the **knock-off** approach for controlling the FDR, which is a new and active area of research. Suppose that we have a test statistic Z_j that depends on $\mathbb{X}$ and Y, such as the t-statistic of β_j in Lasso. The idea behind this approach is that we construct dummy variables $\widetilde{\mathbb{X}}$ such that $Z_j(\mathbb{X}, Y) \stackrel{d}{=} Z_j(\widetilde{\mathbb{X}}, Y)$.

Suppose that we wish to select from d candidate features. Denote

$$[\underbrace{Z_1, \ldots, Z_d}_{\text{Original}}, \underbrace{\widetilde{Z}_1, \ldots, \widetilde{Z}_d}_{\text{Knock-Off}}] = Z([\mathbb{X}, \widetilde{\mathbb{X}}], Y)$$

As a prototypical example, we consider a linear model where the outcome is Y and the design matrix is $[\mathbb{X}, \widetilde{\mathbb{X}}]$, as illustrated in Figure 19.1.

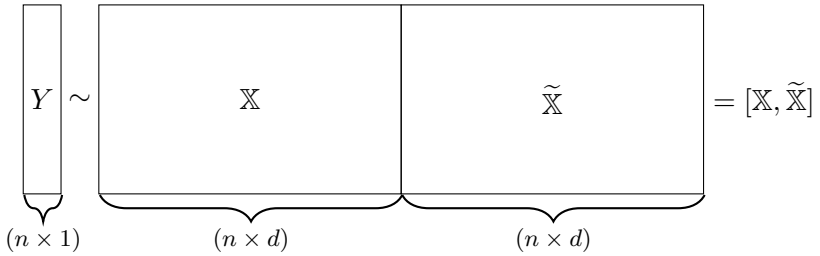


Figure 19.1. *A linear model where the outcome is Y and design matrix is $[\mathbb{X}, \widetilde{\mathbb{X}}]$.*

If $j \in H_{0j}$, then we expect that $(Z_j, \widetilde{Z}_j) \stackrel{d}{=} (\widetilde{Z}_j, Z_j)$ (i.e., exchangeability holds). This notion can be generalized as follows. Let $(\widetilde{Z}, Z)_{\text{swap}(T)}$ denote the vector obtained by swapping the set $T \subset \{1, \ldots, d\}$ of the entries of Z with the corresponding entries of $\widetilde{Z}$ in $(Z, \widetilde{Z})$, as depicted in Figure 19.2. We expect that for any $T \subset H_0$, $(Z, \widetilde{Z}) \stackrel{d}{=} (\widetilde{Z}, Z)_{\text{swap}(T)}$. In order to find an $\widetilde{\mathbb{X}}$ such that $(Z, \widetilde{Z}) \stackrel{d}{=} (\widetilde{Z}, Z)_{\text{swap}(T)}$ is satisfied, we note that $\mathbb{X} \stackrel{d}{=} \widetilde{\mathbb{X}}$ implies $(Z, \widetilde{Z}) \stackrel{d}{=} (\widetilde{Z}, Z)_{\text{swap}(T)}$.

19.1 Definition (Knock-Off Score). *The knock-off score is a statistic $W_j = w_j(Z_j, \widetilde{Z}_j)$ such that*

1. *The w_j are antisymmetric, i.e. $w_j(x, y) = -w_j(y, x)$ for $1 \leq j \leq d$*
2. *If $j \in H_0$, then $W_j \stackrel{d}{=} -\widetilde{W}_j$*
3. *The sign$(W_j)$$|W_1|, \ldots, |W_d|$ are i.i.d. Ber$(1/2)$*

For instance, it can be easily verified that $W_j = Z_j - \widetilde{Z}_j$ is a valid knock-off score if $(Z_j, \widetilde{Z}_j) \stackrel{d}{=} (\widetilde{Z}_j, Z_j)$.

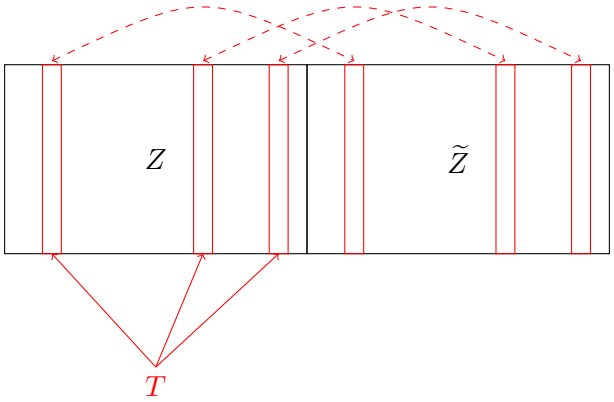


Figure 19.2. *The construction of* $(\widetilde{Z}, Z)_{swap(T)}$*, where* $Z, \widetilde{Z} \in \mathbb{R}^d$ *and* $T \subset \{1, \dots, d\}$.

Consider a procedure where we reject H_{0j} if $W_j \geq t$ for some threshold level t. By definition,

$$\text{FDP}(t) = \frac{\#\{j \in H_0 : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1}.$$

To estimate $\text{FDP}(t)$, it follows from Condition 2 in Definition 19.1 that

$$\text{FDP}(t) \approx \frac{\#\{j \in H_0 : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \underbrace{\frac{\#\{j : W_j \leq -t\} + 1}{\#\{j : W_j \geq t\} \vee 1}}_{\widehat{\text{FDP}}(t)}.$$

Letting

$$S^+(t) = \{j : |W_j| \geq t, \text{sign}(W_j) = 1\}$$
$$S^-(t) = \{j : |W_j| \geq t, \text{sign}(W_j) = -1\},$$

we can express $\widehat{\text{FDP}}(t)$ more simply by

$$\widehat{\text{FDP}}(t) = \frac{\#S^-(t) + 1}{\#S^+(t) \vee 1}.$$

The knock-off procedure (illustrated in Figure 19.3) can then be described as follows:

Let $t^* = \arg\min_t\{\widehat{\text{FDP}}(t) \leq \alpha\}$. The knock-off procedure rejects H_{0j} if and only if $j \in S^+(t^*)$.

The validity of the knock-off procedure is proven in Theorem 19.3, which requires some results on martingales that can be summarized as follows. A sequence of random variables

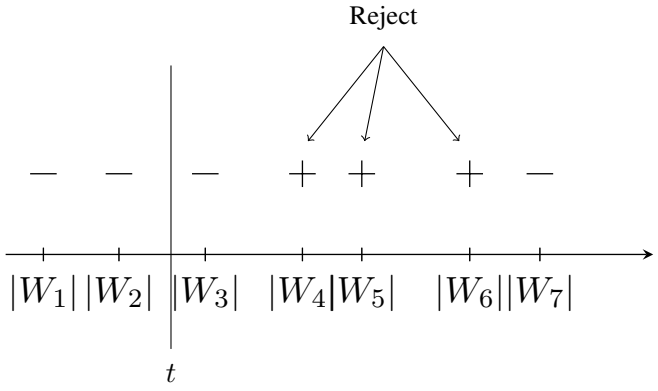


Figure 19.3. *Schematic of the knock-off procedure. The line indexed by t illustrates the threshold level. The sign above $|W_j|$ indicates the sign of W_j. According to the procedure, we reject null hypotheses corresponding to the test statistics W_4, W_5, and W_6.*

$\{X_i\}_{i=1}^{\infty}$ is called a martingale if $\mathbb{E}(X_{n+1}|X_1,\ldots,X_n) = X_n$ almost surely for all $n \geq 1$ and is called a supermartingale if $\mathbb{E}(X_{n+1}|X_1,\ldots,X_n) \leq X_n$ almost surely for all $n \geq 1$. For instance, if $\xi_i = \pm 1$ with probability $1/2$, then $X_n = \sum_{i=1}^{n} \xi_i$ is a martingale. In the context of a gambling system, one's income is a martingale in a fair game and is a supermartingale in an unfair game. Lastly, we will need the following result on martingales and supermartingales in the proof of Theorem 19.3.

19.2 Theorem (Optimal Stopping Theorem). *Suppose that the stopping time τ only depends on the current information. If X_t is a martingale, then $\mathbb{E}(X_\tau) = \mathbb{E}(X_0)$. If X_t is supermartingale, then $\mathbb{E}(X_\tau) \leq \mathbb{E}(X_0)$.*

19.3 Theorem (Knock-Off). *Suppose that we reject H_{0j} if and only if $j \in S^+(t^*)$ where $t^* = \arg\min_t\{\widehat{FDP}(t) \leq \alpha\}$. Then,*

$$\mathbb{E}\left(\frac{\#\textit{False positives}}{\#\textit{Rejected Hypotheses} \vee 1}\right) = \mathbb{E}(\widehat{FDP}(t)) \leq \alpha$$

Proof. We have that

$$\begin{aligned}
\text{FDP}(t^*) &= \frac{\#\{j \in H_0 : j \in S^+(t^*)\}}{\#\{j : j \in S^+(t^*)\}}\left(\frac{1+\#\{j \in H_0 : j \in S^-(t^*)\}}{1+\#\{j \in H_0 : j \in S^-(t^*)\}}\right) \\
&\leq \widehat{\text{FDP}}(t^*)\frac{\#\{j \in H_0 : j \in S^+(t^*)\}}{1+\#\{j \in H_0 : j \in S^-(t^*)\}} \\
&\leq \alpha\frac{V_+(t^*)}{1+V_-(t^*)}
\end{aligned}$$

where $V_\pm(t^*) = \#\{j \in H_0 : j \in S^\pm(t^*)\}$.

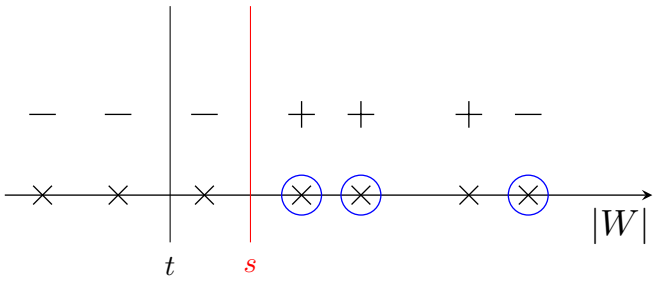


Figure 19.4. *Schematic of the knock-off procedure. The lines indexed by t and s illustrate two threshold levels. The sign above $|W_j|$ indicates the sign of W_j. The blue circles indicate the null hypotheses that are true.*

It suffices to show that $\mathbb{E}\big(\frac{V_+(t^*)}{1+V_-(t^*)}\big) \leq 1$. A key observation is that $V^+(t)|V^-(t) + V^+(t)$ follows a hypergeometric distribution, as the $\text{sign}(W_j)|W$ are i.i.d. $\text{Ber}(1/2)$. Our goal then is to show that for $s \geq t$,

$$\mathbb{E}\left(\frac{V_+(s)}{1+V_-(s)}|\mathcal{F}_t\right) \leq \frac{V_+(t)}{1+V_-(t)}$$

where $\mathcal{F}_t$ contains information if $j \in H_0$ for $|W_j| > t$ and $V_-(t) + V_+(t)$. See Figure 19.4 for an illustration. Since $\frac{V_+(t^*)}{1+V_-(t^*)}$ is supermartingale, it follows from Theorem 19.2 that

$$\mathbb{E}\left(\frac{V_+(t^*)}{1+V_-(t^*)}\right) \leq \mathbb{E}\left(\frac{V_+(0)}{1+V_-(0)}\right) \leq 1$$

which completes the proof. □

References

BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011a). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

BELLONI, A., CHERNOZHUKOV, V. ET AL. (2011b). ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39** 82–130.

CAI, T. and LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association* **106** 1566–1577.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819.

CHIQUET, J., GRANDVALET, Y., CHARBONNIER, C. ET AL. (2012). Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics* **6** 795–830.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360.

FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical Foundations of Data Science*. CRC press.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research* **16** 559–616.

MEINSHAUSEN, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52** 374–393.

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 1009–1030.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645.

WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics* **42** 2164.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101** 1418–1429.