High-Dimensional Statistics

Yanbo Tang

Imperial College London
17th of Feb. 2025

1 Introduction

Update weekly, hopefully with less typos every week. We construct models based on a phenomena of interest, naturally for more complex phenomena we would expect to use a more complex model. We also base model complexity on the amount of information that is available to us, as we are able to collect more information, we are expected to include more information into the statistical model. A curse and blessing of the 21th century has been the amount of information that we now have access too, and the tension this creates between available computational resources and theoretical guarantees for our proposed prediction and inference procedures.

These notes are a convex combination of existing works Wainwright (2019); Vershynin (2018); Boucheron et al. (2013); Rigollet and Hütter (2023) and are a condensed version of the material therein; you should think of this as notes on existing notes rather than a new piece of work. This was created as a reference for a 5 week module for a postgraduate course, so it is more condensed than a classical treatment of the material and is meant to highlight and contrast different aspects of high-dimensional statistics. Some exercises are directly taken from the notes cited above as well.

1.1 Prerequisites

Basic knowledge of undergraduate statistics is assumed, students should be aware of basic concepts of convergence in distribution, convergence in probability and the central limit theorem. Some knowledge of linear regression is assumed as well, otherwise the material will aim to be self-contained as much as possible.

2 Concentration

We will think of a concentration inequality for a sequence of random variable X_n as a bound of the type:

$$P(|X_n - E[X_n]| > t) \leq f_n(t)$$

for some function that is increasing in t, this provides us with an idea of the distribution of its tails. For now, let us imagine that n takes the role of sample size and X_n is an estimator for a quantity of interest, for example an empirical average, it is also reasonable to then expected that $f_n(t)$ is a decreasing function in n. The key feature of these kinds of bounds is that they are valid for all n and therefore avoid asymptotic arguments and allows for more precise control over the behaviour of the tails of these random variables.

We would also wish for the functions $f_n(t)$ to decay as quickly as possible with n and t, we see in what follows that the best rate that one can usually hope for will be roughly $\exp(-nt^2/2)$, but of course in general worse rates are possible.

2.1 Motivation for non-asymptotic results

In our undergraduate studies, we studied various notion of convergence. Amongst them, convergence in probability and convergence in distribution are of primary interest to statisticians. We review these concepts but also contrasts them with the finite sample centric approach we will take in this course. Convergence in probability states that:

$$\forall_{\epsilon>0} \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) \to 0,$$

assuming that these random variables live on a common probability space. This statement does give us some notion of concentration, for example if we were to use the weak law of large numbers, for a sequence of IID random variable with common mean μ and finite first absolute moment we have:

$$\sum_{i=1}^{n} \frac{X_i}{n} \xrightarrow{p} \mu,$$

this tells us that the empirical mean will be expected to be close to the true mean given enough samples.

But this alone does not tell us anything about the rate at which this is happening, and without this, it is difficult to apply a limit theorem to a practical setting. More specifically, we do not know what "enough samples" means in our context. In fact, one can construct arbitrary slowly converging sequences of random variable, as this simple example shows.

Example 1. Let $U \sim \text{uniform}(0,1)$, and $M_n \downarrow 0$ with $M_0 < 1$. Consider the sequence of random variable random variable $X_n = U\mathbb{I}[u \in (0, M_n)]$, then $X_n \xrightarrow{p} 0$ and for all $0 < \epsilon < 1$

$$\mathbb{P}[|X_n| > \epsilon] \leq M_n.$$

We can take arbitrary slow sequences M_n, for example nested logarithms $M_n = \log(\log(\dots(n)))$ and if we want $M_n < 1/2$ we would need a n that is a towering exponential function.

We say that $X_n \xrightarrow{D} X$ if for every continuity point t of the random variable X

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

where $F_n(t)$ are the cumulative functions of X_n. One common way to show convergence in distribution is with the central limit theorem, which states that for a sequence of IID random variables with common mean $E[X_1] = \mu$, $E[(X_1 - \mu)^2] = \sigma^2 < \infty$:

$$\frac{\sum_{i=1}(X_i - \mu)}{\sigma n^{1/2}} \xrightarrow{D} Z,$$

for a standard normal random variable Z. Typically we pretend that this result is "exact" and ignore the fact that it holds only in the limit to construct confidence statements. One exercises you might have encountered is the following:

Example 2. Consider an IID sequence $X_i \sim Bernouilli(p)$. Then by the CLT

$$\frac{\sum_{i=i}^{n}(X_i - p)}{\sqrt{np(1-p)}} \xrightarrow{D} Z,$$

for a standard normal distribution Z.

We are guaranteed that in the limit our results are exact, but we don't know the performance of these confidence intervals in the finite sample setting. This motivates the use of concentration inequalities which are able to provide a stronger guarantee in finite samples settings.

Another goal which we will set for ourselves is the to include dimensionality into the rate of concentration. Classical asymptotic results usually assumes that the ambient dimension of the random variable is fixed. However, if the dimension of the random variable is allowed to vary with the sample size, then some well established results may no longer hold. The following explores such a case with the weak law of large numbers:

Example 3. Consider a p-dimensional multivariate normal distribution $X_{n,p} \sim N(0, I_p)$ where I_p is the p dimension identity matrix. Then if the dimension p is fixed by the weak law of large numbers:

$$\frac{\sum_{i=1}^n X_{n,p}}{n} \xrightarrow{p} 0.$$

However, if we let p increase with n such that $p/n \to c > 0$ then

$$\left\| \frac{\sum_{i=1}^n X_{n,p}}{n} \right\|_2^2 \sim \frac{\chi_p^2}{n},$$

whose variance does not tend to 0, therefore this never concentrates around its expectation of 0.

A statement like the weak law of large number is not well defined if we let the dimension increase, but we understand that the implicit idea of getting closer to the truth with more samples is no longer valid in this case. Therefore, in letting the dimensionality of the underlying problem vary, we are able to more clearly identity the effect of dimension. Many other classical tests, such as the commonly used likelihood ratio tests, fails in high-dimensions even for relatively simple logistic models, see He et al. (2021).

2.2 Sub-Gaussian Concentration

It begins with the humble Markov's inequality.

Theorem 1. *For a positive random variable X with $E[X] < \infty$:*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

The limitation that the random variable has to be non-negative seems restrictive, but this can be circumvented by transforming the initial random variable. One common transformation is to use $|X - \mu|^k$:

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

this inequality is useful so long as $E|X - \mu|^k < \infty$ otherwise it will be vacuous (although still technically valid). The case of $k = 2$ is the well known Chebyschev's inequality. Indeed it is possible to refine this result, given that this bound technically holds for any choice of k:

$$\mathbb{P}(|X - \mu| \geq t) \leq \min_{k=1,\ldots} \frac{\mathbb{E}[|X - \mu|^k]}{t^k},$$

although in practice this requires us to either compute or upper bound the centralized moments and depending on the value of t being considered, it may be possible for the minima to be realized

at different values of k. For any monotonically increasing transformation $f(x) : \mathbb{R} \to \mathbb{R}^+$, the following sequence of inequality holds:

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(f(X - \mu) \geq f(t)) \leq \frac{E[f(X - \mu)]}{f(t)}.$$

A common choice is the function $f(x) = \exp(\lambda x)$ gives us the following:

$$\begin{aligned}\mathbb{P}(X - \mu \geq t) &\leq \inf_{\lambda \in [0,b]} \mathbb{E}[\exp(\lambda(X - \mu) - \lambda t)] \\ &= \exp\left\{ \inf_{\lambda \in [0,b]} \log(E[\exp(\lambda X)]) - \lambda t \right\},\end{aligned}$$

where the bound needs to be optimized for each t and b is the largest value for which the expectation $\mathbb{E}\exp(\lambda X) < \infty$. An apparent weakness of the approach is that the moment generating function must be known, which is typically not the case for most complex random variables. But we will see that an upper bound on the moment generating function suffices to obtain a tail bound.

Chernoff's approach if often used because moment generating functions are very well behaved under convolutions for independent random variables with mean μ_i as it would only involve the following:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu \geq t \right) &\leq \inf_{\lambda \in [0,b]} \prod_{i=1}^n M_{X_i - \mu_i}(\lambda t) \exp(-\lambda t) \\ &= \exp\left\{ \inf_{\lambda \in [0,b]} \log(M_{X_i - \mu_i}(\lambda t) - \lambda t \right\},\end{aligned}$$

where $M_{X_i - \mu_i}(\cdot)$ is the moment generating function for the random variable $X_i - \mu_i$. Using the moment approach with Markov's inequality would involve us having to compute the all of the k-th order cross terms which is tedious. Of course, in the case of IID random variables, both cases simplify considerably.

Remark 1. *As we saw, Chernoff's inequalities are often better than what one can obtain with any single application of Markov's inequality, we do need to essentially assume that all moments of the random variable exists, whereas with Markov's inequality we only need this up to some order k. Most of the presented inequalities are sharp, meaning that there exists a random variable which realizes the $\leq$ with equality for certain values of t. Therefore we can only hope to trade in better rates of concentration through stricter assumptions.*

Applying this to the Gaussian distribution with mean μ and σ gives us the following bound on the tails of the Gaussian:

$$\mathbb{P}[X > \mu + t] \leq \exp\left\{ \inf_{\lambda \in [0,b]} \log(E[\exp(\lambda X)]) - \lambda t \right\} = \exp\left\{ \inf_{\lambda \in [0,\infty)} \frac{\lambda^2 \sigma^2}{2} - \lambda t \right\} = \exp\left(\frac{-t^2}{2\sigma^2} \right),$$

where the value of λ which solves $\inf_{\lambda \in [0,\infty)} \frac{\lambda^2 \sigma^2}{2} - \lambda t$ can be obtained by differentiating with respect to λ and solving the equation:

$$\lambda \sigma^2 - t = 0,$$

justifiable as the function is strongly convex and infinitely differentiable. By symmetry of the Gaussian distribution (consider $-X$), we have the following lower tail bound:

$$\mathbb{P}[X < \mu - t] \leq \exp\left(\frac{-t^2}{2\sigma^2} \right).$$

we can then combine these two bounds into a two sided inequality by a union bound (let $A_1 = \{X - \mu > t\}$, $A_2 = \{X - \mu < -t\}$, then $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$):

$$\mathbb{P}[|X - \mu| > t] \leq \exp\left(\frac{-t^2}{2\sigma^2}\right).$$

These bounds are not optimal for Gaussian random variables, in fact they are off by at least a polynomial factor $\frac{1}{t}$, see exercise 2 for a sharper bound.

But this result is useful as we can obtain these types bounds so long as the random variable has "Gaussian like" tails, and this can be quantified through a bound on the moment generating function of the random variable.

Definition 1. *A random variable X is called sub-Gaussian with proxy variance σ^2 if there exists a $\sigma^2 > 0$:*

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

for all $\lambda \in \mathbb{R}$. A random vector $X \in \mathbb{R}^d$ is sub-Gaussian with proxy variance σ^2 if $u^\top X$ is sub-Gaussian with proxy variance σ^2 for all $u \in S^{d-1}$.

In this text σ^2 will be refereed to as the proxy variance, it is not an unique value but a smaller value of σ^2 will provide a better bound. The requirement that this inequality holds for all real numbers requires that the moment generating function exists for all $\lambda \in \mathbb{R}$, which is stronger than the requirement for the generic Chernoff approach. Immediately we have:

Proposition 1. *-Gaussian random variable X with proxy variance σ satisfies:*

$$\mathbb{P}[X - \mu > t] \leq \exp\left(\frac{-t^2}{2\sigma^2}\right),$$
$$\mathbb{P}[X - \mu < -t] \leq \exp\left(\frac{-t^2}{2\sigma^2}\right),$$
$$\mathbb{P}[|X - \mu| > t] \leq 2\exp\left(\frac{-t^2}{2\sigma^2}\right).$$

The class of sub-Gaussian random variables is relatvely broad and includes many useful random variables. Indeed by Hoeffding's lemma, we immediately have that all bounded random variables are sub-Gaussian with proxy variance $\sigma^2 = (b-a)^2/4$.

Lemma 1. ***Hoeffding Lemma**: Let X be any random variable such that $a < X < b$ almost surely. Then for all $\lambda \in \mathbb{R}$:*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2(b-a)^2/8)$$

Proof. Without loss of generality assume that $\mathbb{E}[X] = 0$. By convexity of $\exp(\lambda x)$:

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{a-x}{b-a}e^{\lambda a}.$$

Therefore,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \frac{b - \mathbb{E}[X]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[X] - a}{b-a}e^{\lambda b} = \frac{b}{b-a}e^{\lambda a} + \frac{-a}{b-a}e^{\lambda b} =: e^{\mathcal{L}(\lambda(b-a))},$$

for all $x \in [a,b]$. where $\mathcal{L}(h) = \frac{ha}{b-a} + \ln\left(1 + \frac{a-e^h a}{b-a}\right)$. The function

$$\mathcal{L}(0) = \mathcal{L}'(0) = 0 \quad \text{and} \quad \mathcal{L}''(h) = -\frac{abe^h}{(b-ae^h)^2}.$$

From the AMGM inequality we thus see that $\mathcal{L}''(h) \leq \frac{1}{4}$ for all h. By a second order Taylor expansion with the mean value form of the remainder, there is some $0 \leq \theta \leq 1$ such that

$$\mathcal{L}(h) = \mathcal{L}(0) + h\mathcal{L}'(0) + \frac{1}{2}h^2\mathcal{L}''(h\theta) \leq \frac{1}{8}h^2.$$

Thus, $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2(b-a)^2/8}$. □

Sub-Gaussian bounds are preserved under convolution, which are useful for studying the behavior of averages and sums.

Proposition 2. ***Exercise 2.13 in Wainwright (2019)*** *Suppose that X_1 and X_2 are 0 mean sub-Gaussian random variables with proxy variances of σ_1^2 and σ_2^2*

- *If they are independent, then $X_1 + X_2$ is sub-Gaussian with proxy variance $\sigma_1^2 + \sigma_2^2$*
- *In general, $X_1 + X_2$ sub-Gaussian with proxy variance $(\sigma_1 + \sigma_2)^2$*
- *For $c \in \mathbb{R}$, cX_1 is subGaussian with proxy variance $c^2\sigma_1^2$.*

This gives us the following concentration inequality for averages of sub-Gaussian random variables:

Theorem 2. ***Hoeffding bound for averages:*** *Let X_i for $i = 1, \ldots, n$ be a sequence of IID random variables with proxy variances σ^2, then:*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i/n - \mu\right| \geq t\right) \leq \exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

This theorem can be thought of as providing a qualitative bound for the weak law of large numbers. But the strength of these inequalities can be seen when they are able to account for the behavior of the dimension of the problem in greater detail. We consider an application of this bound to the problem of Monte Carlo estimation for volumes.

Suppose we are interested in calculating the unknown volume (Lebesgue measure) of a set F which is contained within a set F' with known finite volume. Then if we can sample X_i's from the uniform distribution supported on F', it is possible to approximate the volume of F by:

$$\text{Vol}(F) \approx \text{Vol}(F') \sum_{i=1}^{n} \frac{\mathbb{I}[X_i \in F]}{n}.$$

As the random variable $\text{Vol}(F')\mathbb{I}[X_i \in F]$ only takes the value of 0 or $\text{Vol}(F')$,

Example 4. Let $F = \{x \in R^p : \|x\|_2 \leq 1\}$ and let $F' = \{x \in R^p : \|x\|_\infty \leq 1\}$, where F is a hypersphere of dimension p with radius 1, while F' is the hypercube centered at 0 with sides of length 2. In this case we know that $\text{Vol}(F) = \pi^{p/2}/\Gamma(p/2+1)$ and $\text{Vol}(F') = 2^p$, thus

$$\mathbb{P}\left[\left|\mathrm{Vol}(F')\sum_{i=1}^{n}\mathbb{I}[X_i\in F]/n-\mathrm{Vol}(F)\right|>\delta\right]\leq 2\exp\left(-\frac{\delta^2 n}{2^{2p-1}}\right),$$
$$\mathbb{P}\left[\left|\frac{\mathrm{Vol}(F')\sum_{i=1}^{n}\mathbb{I}[X_i\in F]-\mathrm{Vol}(F)}{\mathrm{Vol}(F)}\right|>\delta\right]\leq 2\exp\left(-\frac{\delta^2\pi^p n}{2^{2p-1}\Gamma(p/2+1)^2}\right),$$

The absolute error is decaying extremely quickly, as the volume of an unit sphere is exponentially decaying to 0 as its dimension increases, therefore the chance of hitting the sphere by sampling from the unit cube is also exponentially decaying to 0. Our estimate will be essentially 0, but this is quite close to the volume of an unit sphere in high dimensions. However, in order for the relative error to tend to 0 we require that:

$$n(p)=\exp[\omega\{p\log(p)\}],$$

by Stirling's approximation, to achieve a measure of relative consistency we need a exponentially increasing number of samples in the dimension of the sphere.

2.3 Maxima of sub-Gaussians

It is of interest to control the maximum or supremum of a collection of random variables, this is commonly used in empirical process theory or learning theory for example and we will see this used in Section 3. The fast rate of decay in the tail of sub-Gaussian random variables is also very useful for controlling the maximum of independent sub-Gaussian random variables.

Proposition 3. *Let $X_1,\ldots,X_n$ be a sequence of sub-Gaussian random variables with common proxy variance σ^2 then*

$$\mathbb{E}[\max_{1\leq n}X_i]\leq\sigma\sqrt{2\log(n)},$$
$$\mathbb{P}(\max_{1\leq n}X_i>t)\leq N\exp\left(\frac{-t^2}{\sigma^2}\right).$$

Note that independence is not needed.

Parts of the following proof generalizes to other random variables, for example sub-exponential random variables which will be introduced in the next section.

Proof. By Jensen's inequality

$$\exp\left(\lambda\mathbb{E}\max_{i=1,\ldots,N}Z_i\right)\leq\mathbb{E}\exp\left(\lambda\max_{i=1,\ldots,N}Z_i\right)=\mathbb{E}\max_{i=1,\ldots,n}e^{\lambda Z_i},$$

using the fact that $\max_{i=1,\ldots,n}a_i\leq\sum_{i=1}^{n}a_i$ for positive a_i,

$$\mathbb{E}\max_{i=1,\ldots,N}e^{\lambda Z_i}\leq\sum_{i=1}^{N}\mathbb{E}e^{\lambda Z_i}\leq Ne^{\lambda^2\sigma^2/2}.$$

Taking logarithms on both sides, we have

$$\mathbb{E}\max_{i=1,\ldots,N}Z_i\leq\frac{\log N}{\lambda}+\frac{\lambda\sigma}{2}.$$

The upper bound is minimized for $\lambda = \sqrt{2\log N/\sigma^2}$ which yields

$$\mathbb{E} \max_{i=1,\dots,N} Z_i \leq \sqrt{2\sigma \log N}.$$

For the second statement it follows from an application of the union bound:

$$\mathbb{P}\left(\max_{1\leq i\leq n} X_i > t\right) = \mathbb{P}\left(\bigcup_{1\leq i\leq n} \{X_i > t\}\right) \leq \sum_{1\leq i\leq n} \mathbb{P}(X_i > t) \leq ne^{-\frac{t^2}{2\sigma^2}}.$$

□

Two sided bounds can also be obtained by considering $-X_i$ and using the union bound. The independent case is actually the worst case scenario for the growth of the maximum. For some intuition behind this, imagine a sequence of perfectly correlated standard Gaussian random variables with $X_1 = X_2 = \cdots = X_n$, then the expectation of the maximum will simply be 0.

What if we wanted to control a maximum or supremum over an infinite set? For example, consider the unit ℓ_2 ball in $\mathbb{R}^d$ which is defined as the set of vectors with Euclidean norm $\|u\|_2$ at most 1. Formally,

$$\mathcal{B}_2 = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i^2 \leq 1 \right\},$$

and we are interested in controlling for:

$$E\big[\sup_{\theta\in\mathcal{B}_2} \theta^\top X\big],$$

where X follows a sub-Gaussian distribution. We will try to write the maximum over $\mathcal{B}_2$ as a maximum over some finite set along with some "approximation error", to do so, we introduce the idea of covering numbers and ϵ-nets.

Definition 2. *Fix $K \subset \mathbb{R}^d$ and $\varepsilon > 0$. A set $\mathcal{N}$ is called an ε-net of K with respect to a distance $d(\cdot,\cdot)$ on $\mathbb{R}^d$, if $\mathcal{N} \subset K$ and for any $z \in K$, there exists $x \in \mathcal{N}$ such that $d(x,z) \leq \varepsilon$.*

If $\mathcal{N}$ is an ε-net of K with respect to a norm $\|\cdot\|$, then every point of K is at distance at most ε from a point in $\mathcal{N}$. If K is a compact set, then it is always possible to find an ϵ covering, we are however after an efficient covering, so we need to find a good upper bound for the number of points needed.

Lemma 2. *For any $\varepsilon \in (0,1)$, the unit Euclidean ball $\mathcal{B}_2$ has an ε-net $\mathcal{N}$ with respect to the Euclidean distance of cardinality $|\mathcal{N}| \leq (3/\varepsilon)^d$.*

Proof. Consider the following iterative construction of the ε-net. Choose $x_1 = 0$. For any $i \geq 2$, take x_i to be any $x \in \mathcal{B}_2$ such that $|x - x_j|_2 > \varepsilon$ for all $j < i$. If no such x exists, then we are done. Clearly, this creates an ε-net of the unit ball. We now control its size.

Observe that since $|x - y|_2 > \varepsilon$ for all $x, y \in \mathcal{N}$, the Euclidean balls centered at $x \in \mathcal{N}$ and with radius $\varepsilon/2$ are disjoint. Moreover,

$$\bigcup_{z\in\mathcal{N}} \{z + \frac{\varepsilon}{2}\mathcal{B}_2\} \subset (1 + \frac{\varepsilon}{2})\mathcal{B}_2$$

where $\{z + \varepsilon\mathcal{B}_2\} = \{z + \varepsilon x, x \in \mathcal{B}_2\}$. Thus, measuring the volumes of these sets, we get

$$\mathrm{vol}\left((1+\frac{\varepsilon}{2})\mathcal{B}_2\right) \geq \mathrm{vol}\left(\bigcup_{z\in\mathcal{N}}\{z+\frac{\varepsilon}{2}\mathcal{B}_2\}\right) = \sum_{z\in\mathcal{N}}\mathrm{vol}\left(\{z+\frac{\varepsilon}{2}\mathcal{B}_2\}\right).$$

This is equivalent to

$$\left(1+\frac{\varepsilon}{2}\right)^d \geq |\mathcal{N}|\left(\frac{\varepsilon}{2}\right)^d.$$

Therefore, we get the following bound

$$|\mathcal{N}| \leq \left(1+\frac{2}{\varepsilon}\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d.$$

□

Theorem 3. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2. Then*

$$\mathbb{E}\big[\sup_{\theta\in B_2}\theta^T X\big] = \mathbb{E}\big[\sup_{\theta\in B_2}|\theta^T X|\big] \leq 4\sigma\sqrt{d}.$$

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

$$\sup_{\theta\in B_2}\theta^T X = \sup_{\theta\in B_2}|\theta^T X| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

Proof. Let $\mathcal{N}$ be a $1/2$-net of B_2 with respect to the Euclidean norm which satisfies $|\mathcal{N}| \leq 6^d$ by Lemma 2. Observe that for every $\theta \in B_2$, there exists $z \in \mathcal{N}$ and x such that $|x|_2 \leq 1/2$ and $\theta = z + x$. Therefore,

$$\max_{\theta\in B_2}\theta^T X \leq \max_{z\in\mathcal{N}} z^T X + \max_{x\in\frac{1}{2}B_2} x^T X.$$

But

$$\max_{x\in\frac{1}{2}B_2} x^T X = \frac{1}{2}\max_{x\in B_2} x^T X.$$

Therefore,

$$\mathbb{E}[\max_{\theta\in B_2}\theta^T X] \leq 2\mathbb{E}[\max_{z\in\mathcal{N}} z^T X] \leq 2\sigma\sqrt{2\log(|\mathcal{N}|)d} \leq 4\sigma\sqrt{d}.$$

The bound with high probability follows as

$$\mathbb{P}\big(\max_{\theta\in B_2}\theta^T X > t\big) \leq \mathbb{P}\big(2\max_{z\in\mathcal{N}} z^T X > t\big) \leq |\mathcal{N}|e^{-\frac{t^2}{8\sigma^2}} \leq 6^d e^{-\frac{t^2}{8\sigma^2}}.$$

To conclude the proof, we find t such that

$$e^{-\frac{t^2}{8\sigma^2}+d\log(6)} \leq \delta \iff t^2 \geq 8\log(6)\sigma^2 d + 8\sigma^2\log(1/\delta).$$

Therefore, it is sufficient to take

$$t = \sqrt{8\log(6)\sigma^2\sqrt{d} + 2\sigma^2\sqrt{2\log(1/\delta)}}.$$

Remark 2. ***Exercise 2.3 Wainwright (2019)*** *The Chernoff method can be sub-optimal. If a positive random variable X has a moment generating function whose value is finite for an interval around 0 then there exists a t such that:*

$$\inf_{k=0,1,\ldots} \frac{\mathbb{E}[|X|^k]}{t^k} \leq \inf_{\lambda>0} \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda t)},$$

which in turn implies that:

$$\mathbb{P}(X \geq t) \leq \inf_{k=0,1,\ldots} \frac{\mathbb{E}[|X|^k]}{t^k} \leq \inf_{\lambda>0} \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda t)}.$$

This shows that a well optimized moment bound is never worst than a Chernoff bound.

Remark 3. *For the problem of volume estimation, generally pure Monte Carlo approaches does not perform well in high-dimensions and the example was purely illustrative. The estimation of the volume of convex high-dimensional figures has a rich history, see [I'll find it eventually!] for a summary of some of the results from MCMC type approaches.*

Exercise 1. *Show that is the moment generating function $M_X(t)$ exists for some values of $|t| < \delta$, then all moments $E[X^k]$ exists. Show that the converse is not true.*

Exercise 2. *Show that for a standard normal random variable Z*

$$\left(\frac{1}{z} - \frac{1}{z^3}\right) \leq P[Z \geq z] \leq \frac{1}{z}\phi(z) \text{ for } z > 0,$$

where $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$, the density of a standard normal distribution.

Exercise 3. *Show the difference in the α level quantiles implies by the sub-Gaussian bound and the Mill's ratio.*

2.4 Sub-Exponential Concentration

The Gaussian tail bounds decay roughly of order $\exp(-nt^2)$, which is quite rapid, however, these tails are extremely light, so it is worth thinking about other classes of random variable which shows slower but still exponential decay.

Definition 3. *A random variable with mean X with $\mu = E[X]$ is sub-exponential if there are non-negative parameters (ν, α) such that*

$$E\left[\exp(\lambda(X-\mu))\right] \leq \exp\left(\frac{\nu^2\lambda^2}{2}\right) \text{ for all } |\lambda| < \frac{1}{\alpha}.$$

A good example of a commonly used distribution which is sub-exponential is the exponential distribution with rate parameter 1, the centralized random variable $X-1$ has moment generating function is $M_{X-1}(\lambda) = \exp(-\lambda)(1-\lambda)^{-1}$ if $\lambda < 1$, note that this random variable is not sub-Gaussian as the moment generating function does not exist everywhere.

Proposition 4. *Suppose that X is a sub-exponential distribution with parameters (ν, α) then:*

$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}} & \text{for } t > \frac{\nu^2}{\alpha}. \end{cases}$$

Proof. Without loss of generality assume that $\mu = 0$. We use the Chernoff-type approach as was done for the Gaussian

$$\mathbb{P}[X \geq t] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}] \leq \underbrace{\exp\left(-\lambda t + \frac{\lambda^2\nu^2}{2}\right)}_{g(\lambda,t)},$$

which is valid for all $\lambda \in [0, \alpha^{-1}]$.

To complete the proof, it remains to compute, for each fixed $t \geq 0$,

$$g^*(t) := \inf_{\lambda\in[0,\alpha^{-1}]} g(\lambda, t).$$

Note that the unconstrained minimum of the function $g(\lambda, t)$ occurs at $\lambda^* = t/\nu^2$. However, if $0 \leq t < \frac{\nu^2}{\alpha}$, then this unconstrained optimum corresponds to the constrained minimum as well, so that

$$g^*(t) = -\frac{t^2}{2\nu^2}$$

over this interval.

Otherwise, we may assume that $t \geq \frac{\nu^2}{\alpha}$. In this case, since the function $g(\cdot, t)$ is monotonically decreasing in the interval $[0, \lambda^*]$, the constrained minimum is achieved at the boundary point of α^{-1}, and we have

$$g^*(t) = g(\alpha^{-1}, t) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{\nu^2}{\alpha} \leq -\frac{t}{2\alpha},$$

where we used the fact that $\frac{\nu^2}{\alpha} \leq t$. □

Sometimes it is difficult to compute the moment generating function, a commonly used sufficient condition to get sub-exponential bounds is the Bernstein condition:

Definition 4. ***Bernstein condition****. Given a random variable X with mean μ and variance σ^2, we say that Bernstein's condition with parameter b holds if*

$$\mathbb{E}[(X-\mu)^k] \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \textit{for all } k \in \mathbb{N}.$$

Proposition 5. *For any random variable satisfying the Bernstein condition with parameter b*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\lambda\sigma^2}{2-2|\lambda|b}\right) \qquad \textit{for all } |\lambda| < \frac{1}{\nu},$$

and, moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2\exp\left(\frac{-t^2}{2(\sigma^2+bt)}\right) \qquad \textit{for all } t \geq 0.$$

Proposition 6. ***Preservation of sub-exponential property.*** *For a sequence of independent random variables X_i for $i = 1, \ldots, n$ which are sub-exponential (ν_i, α_i), the sum*

$$\sum_{i=1}^{n}(X_i - E(X_i)),$$

is sub-exponential with parameters (ν_*, α_*) where $\alpha_* = \max_{i=1,\ldots,n}\alpha_i$ and $\nu_\star = \sqrt{\sum_{i=1}^n \nu_i^2}$.

Similar to the sub-Gaussian case, we can show the following for the concentration of averages of independent sub-exponential distribution:

Proposition 7. *For a sequence of independent random variables X_i for $i = 1, \ldots, n$ which are sub-exponential (ν_i, α_i),*

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n}(X_k - \mu_k) \geq t\Big] \leq \begin{cases} \exp\left(-\frac{n^2t^2}{2\nu_*}\right) & \text{for } 0 \leq t \leq \frac{\nu_*^2}{n\alpha_*}, \\ \exp\left(-\frac{nt}{2\alpha_*}\right) & \text{for } t > \frac{\nu_*^2}{n\alpha_*}, \end{cases}$$

Example 5. **Concentration of Chi-squared random variables**. Consider a chi-squared random variable with n degrees of freedom, denoted by $Y \sim \chi_n^2$, by properties of gamma distribution (of which the chi-squared belongs to) we can write

$$Y = \sum_{k=1}^{n} Z_k^2$$

where $Z_k \sim \mathcal{N}(0,1)$ are i.i.d. variates. The random variable Z_k^2 is sub-exponential with parameters $(2, 4)$ (show this!). Consequently, since the random variables $\{Z_k\}_{k=1}^n$ are independent, the χ^2-variate Y is sub-exponential with parameters $(\nu, \alpha) = (2\sqrt{n}, 4)$, and provides us with the following tail bound

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\right| \geq t\right] \leq 2e^{-nt^2/8}, \quad \text{for all } t \in (0,1).$$

We will now see an important application of sub-exponential concentration: the Johnson-Lindenstrauss lemma. Suppose that we have a set of very high-dimensional vectors $\{u_1, \ldots, u_N\}$ of dimension d, which we cannot properly store on our computers due to memory constraints. We would ideally like to compress the data using some function $F : \mathbb{R}^d \to \mathbb{R}^m$ in a way to preserve some important feature of this set of vectors, in this example suppose that we are interested in preserving the pairwise Euclidean distance.

More precisely, we want a mapping F such that for some error tolerance $\delta \in (0,1)$

$$(1-\delta) \leq \frac{\|F(u_i) - F(u_j)\|_2}{\|u_i - u_j\|_2} \leq (1+\delta),$$

for all $i \neq j$. It turns out that there is a very easy way of doing this with a random projection.

Form a random matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ filled with independent $\mathcal{N}(0,1)$ entries, and use it to define a linear mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ via $u \mapsto \mathbf{X}u/\sqrt{m}$. We now verify that F satisfies our requirement with high probability. Let $x_i \in \mathbb{R}^d$ denote the ith row of $\mathbf{X}$, and consider some fixed $u \neq 0$. Since x_i is a standard normal vector, the variable $\langle x_i, u/\|u\|_2\rangle$ follows a $\mathcal{N}(0,1)$ distribution, and hence the quantity

$$Y := \frac{\|\mathbf{X}u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle x_i, u/\|u\|_2\rangle^2,$$

follows a χ^2 distribution with m degrees of freedom, using the independence of the rows. Therefore, applying the tail bound for chi-squared random variables, we find that

$$\mathbb{P}\left[\left|\frac{\|\mathbf{X}u\|_2^2}{m\|u\|_2^2} - 1\right| \geq \delta\right] \leq 2e^{-m\delta^2/8}, \quad \text{for all } \delta \in (0,1).$$

Rearranging and recalling the definition of F yields the bound

$$\mathbb{P}\left[\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [(1-\delta),(1+\delta)]\right] \leq 2e^{-m\delta^2/8}, \quad \text{for any fixed } 0 \neq u \in \mathbb{R}^d.$$

Noting that there are $\binom{N}{2}$ distinct pairs of data points, we apply the union bound to conclude that

$$\mathbb{P}\left[\frac{\|F(u^i-u^j)\|_2^2}{\|u^i-u^j\|_2^2} \notin [(1-\delta),(1+\delta)] \text{ for some } u^i \neq u^j\right] \leq 2\binom{N}{2}e^{-m\delta^2/8}.$$

For any $\epsilon \in (0,1)$, this probability can be driven below ϵ by choosing $m > \frac{16}{\delta^2}\log(N/\epsilon)$.

Exercise 4. *Prove Proposition 6.*

Exercise 5. *Prove Proposition 7.*

2.5 Functional concentration

So far we have only dealt with bounds on averages of random variables, ideally we would like to extend to concentration for functions of random variables. The question becomes what assumptions will we then need on these functions to be able to provide exponential type concentration?

To obtain rapid concentration we need our function to not vary wildly with different potential inputs, otherwise it would be difficult to control their potential outputs. The first type of functions with this stability type behaviour are function with bounded differences. Suppose we have a function $f : \mathbb{R}^d \to \mathbb{R}$ for all $x_1, x_2, \ldots, x_d, x_1', x_2', \ldots, x_d' \in \mathbb{R}$

$$\begin{aligned}
|f(x_1', x_2, \ldots, x_j, \ldots, x_d) - f(x_1, x_2, \ldots, x_j, \ldots, x_d)| &\leq L_1 \\
&\vdots \\
|f(x_1, x_2, \ldots, x_j', \ldots, x_d) - f(x_1, x_2, \ldots, x_j, \ldots, x_d)| &\leq L_j \\
&\vdots \\
|f(x_1, x_2, \ldots, x_j, \ldots, x_d') - f(x_1, x_2, \ldots, x_j, \ldots, x_d)| &\leq L_d,
\end{aligned}$$

meaning that if we switch any of the single j-th inputs the function will not change by more than L_j. In this way the function is very *stable* and we obtain the following:

Proposition 8. ***(Bounded differences inequality/McDiaramids)*** *Suppose that f satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$ and that the random vector*

$$X = (X_1, X_2, \ldots, X_n)$$

has independent components. Then

$$\mathbb{P}\big[|f(X) - \mathbb{E}[f(X)]| \geq t\big] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}} \quad \textit{for all } t \geq 0.$$

This has been used in a variety of settings, for instance it has been used to show that stable learning algorithm generalize well to unseen inputs. We will look at two uses of this inequality, one which involves U-statistics and another involving Erdos-Renyi random graphs.

Example 6. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a bounded symmetric function of its arguments (say $\|g\|_\infty \leq b$). Given an IID sequence X_k, $k \geq 1$, of random variables, the quantity

$$U := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k) \tag{1}$$

is as a pairwise *U-statistic*. For instance, if $g(s,t) = |s-t|$, then U is an unbiased estimator of the mean absolute pairwise deviation $\mathbb{E}[|X_1 - X_2|]$. While U is not a sum of independent random variables, the dependence is relatively weak. Viewing U as a function $f(x) = f(x_1, \ldots, x_n)$, for any given coordinate k, we have

$$\begin{aligned} |f(x_1, \ldots, x'_j, \ldots, x_n) - f(x_1, \ldots, x_j, \ldots, x_n)| &\leq \frac{1}{\binom{n}{2}} \sum_{i \neq j} |g(x_i, x_j) - g(x_i, x'_j)| \\ &\leq \frac{(n-1)(2b)}{\binom{n}{2}} = \frac{4b}{n}, \end{aligned}$$

so that the bounded differences property holds with parameter $L_j = \frac{4b}{n}$ in each coordinate. Thus, we conclude that

$$\mathbb{P}(|U - \mathbb{E}[U]| \geq t) \leq 2e^{-\frac{nt^2}{8b^2}}.$$

This tail inequality implies that U is a consistent estimate of $\mathbb{E}[U]$, and provides a finite sample guarantee for its performance. Similar techniques can be used to obtain tail bounds on U-statistics of higher order, involving sums over k-tuples of variables. Note that is the random variables were bounded instead of the function $g(\cdot, \cdot)$ the same bound would hold.

Example 7. **(Clique number in random graphs)** An undirected graph is a pair $G = (V, E)$, composed of a vertex set $V = \{1, \ldots, d\}$ and an edge set E, where each edge $e = (i, j)$ is an unordered pair of distinct vertices ($i \neq j$). A graph clique C is a subset of vertices such that $(i, j) \in E$ for all $i, j \in C$.

The clique number $C(G)$ of the graph is the cardinality of the largest clique—note that $C(G) \in [1, d]$. When the maximum clique is of size 1, the graph will be fully disconnected, while a maximum clique of size d means every node is connected with every other node. If the edges E of the graph are drawn according to some random process, then the clique number $C(G)$ is a random variable, and we can study its concentration around its mean $\mathbb{E}[C(G)]$.

The *Erdös–Rényi* ensemble of random graphs is one of the most well-studied and simplest model. For each $i < j$ (this is so you don't include the same edge twice), introduce a Bernoulli *edge-indicator variable* X_{ij} with parameter $p \in (0, 1)$, where $X_{ij} = 1$ means that edge (i, j) is included in the graph, and $X_{ij} = 0$ means that it is not included.

The $\binom{d}{2}$-dimensional random vector $Z := \{X_{ij}\}_{i<j}$ specifies the edge set; thus, we may view the clique number $C(G)$ as a function $Z \mapsto f(Z)$. Let Z' denote a vector in which a single coordinate of Z has been changed, and let G' and G be the associated graphs. Then $C(G')$ can differ from $C(G)$ by at most 1, so that $|f(Z') - f(Z)| \leq 1$. Thus, the function $C(G) = f(Z)$ satisfies the bounded difference property in each coordinate with parameter $L_j = 1$, so

$$\mathbb{P}\left[\frac{1}{n}|C(G) - \mathbb{E}[C(G)]| \geq \delta\right] \leq 2e^{-2n\delta^2}.$$

Consequently, the clique number of an Erdös–Rényi random graph is very sharply concentrated around its expectation (although we have not calculated its expectation, but you can try it!).

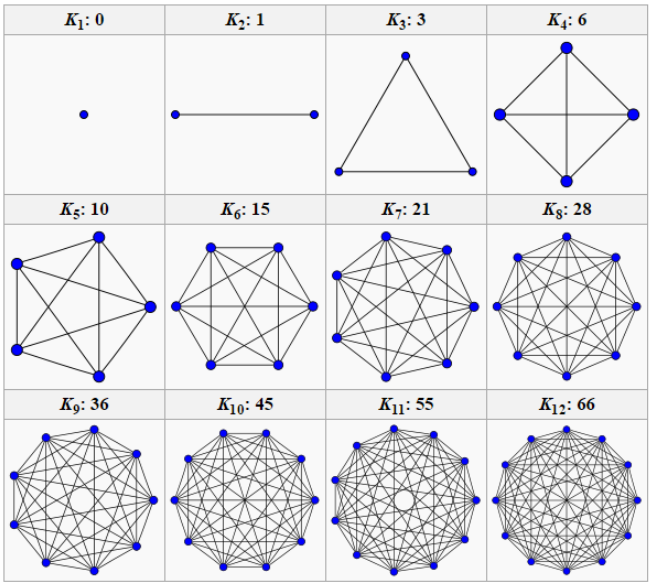

Figure 1: Illustration of what different clique sizes looks like, clique sizes are used as a descriptive statistic for graphs. The title of each sub-plot: $K_i : j$ indicate the number of nodes (i) and the number of total edges (j). You can think of them as groups of friends or perhaps more realistically as enemies.

The other form of "smoothness" that is commonly used is Lipschitz continuity of a function. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm if:

$$|f(x) - f(y)| \leq L\|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^n.$$

This condition controls how much the function varies with different inputs, but contrary to the bounded difference assumption, this function can now be unbounded. Also recall that by Rademacher's theorem Lipschitz functions are differentiable almost everywhere.

Theorem 4. *Let $(X_1, \dots X_n)$ be a vector of IID standard Gaussian random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz function with respect to the Euclidean norm. Then $f(X) - E[f(X)]$ is sub-Gaussian with parameter at most L and*

$$\mathbb{P}[|f(X) - E[f(X)]| \geq t] \leq 2\exp\left(\frac{-t^2}{2L^2}\right) \text{ for all } t \geq 0.$$

Note this bound is dimension free and the concentration only depends on L, but it is possible for L to increase with n however.

To use this result, let us consider a *random matrix* of standard Gaussians, in particular we are interested in the singular values of such matrices. We saw in the Johnson-Lindenstrauss example that these matrices are interesting objects of study.

As a reminder for a real matrix $A \in \mathbb{R}^{n \times d}$, the singular value decomposition is:

$$A = \sum_{i=1}^{r} s_i(A) u_i v_i^\top, \text{ where } r = rank(A).$$

The non negative numbers $s_i(A)$ are called the singular values of A, the vectors $u_i \in \mathbb{R}^n$ are the left singular vectors of A, and $v_i \in \mathbb{R}^d$ are the right singular vectors of A. The singular values are the square root of the singular values of the matrix $A^\top A$ or equivalently $AA^\top$, specifically if $\lambda_i(A)$ denotes the i-th largest eigenvalue of a real symmetric matrix:

$$s_i(A) = \sqrt{\lambda_i(A^\top A)} = \sqrt{\lambda_i(AA^\top)}.$$

If we have a square symmetric real matrix, then the singular values are simply the absolute values of the eigenvalues.

The following lemma is quite useful in bounding the effect of a small perturbation on the singular values of a matrix:

Lemma 3. ***Weyl's lemma.*** *Given two matrices X and Y in $\mathbb{R}^{n\times d}$, we have*

$$\max_{i=1,\dots,d} |s_k(X) - s_k(Y)| \leq s_1(X-Y) \leq \|X-Y\|_F,$$

where $\|\cdot\|_F$ is the Frobenius norm of a $\mathbb{R}^{n\times d}$ matrix:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} a_{ij}^2} = \sqrt{\mathit{Trace}(A^\top A)} = \sqrt{\sum_{i=1}^{\min(n,d)} s_i(A)}.$$

You can think of the Frobenius norm as being a vectorized L^2 norm of a matrix.

Example 8. **(Singular values of Gaussian random matrices)** For integers $n > d$, let $X \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and let

$$s_1(X) \geq s_2(X) \geq \cdots \geq s_d(X)$$

denote its ordered singular values. Let us think us s_k as functions which maps $\mathbb{R}^{n\times d} \to \mathbb{R}^+$. By Weyl's lemma, given another matrix $Y \in \mathbb{R}^{n\times d}$, we have

$$\max_{k=1,\dots,d} |s_k(X) - s_k(Y)| \leq \|X-Y\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}(x_{ij}-y_{ij})^2},$$

which shows that each singular value $s_k(X)$ is a 1-Lipschitz function of the random matrix, so that by Theorem 4, for each $k = 1,\dots,d$,

$$\mathbb{P}(|s_k(X) - \mathbb{E}[s_k(X)]| \geq \delta) \leq 2e^{-\frac{\delta^2}{2}} \quad \text{for all } \delta \geq 0.$$

Consequently, we are guaranteed that the expectations are representative of the typical behavior of the random singular values. It turns out that characterizing the distribution of the expectations of these singular values is much more difficult as we need to consider the structures of these random matrices much more carefully.

Random matrix theory is quite rich and interesting (Terrance Tao has a nice set of notes on this) and we consider some additional results related to the concentration of sums of matrices when we consider covariance estimation.

Finally, what if we don't have a Gaussian distribution? Well, for a general class of *strongly log-concave* distributions a similar type of concentration exists as well:

Definition 5. *A distribution supported in $\mathbb{R}^n$ with density $p(x) = \exp(-\psi(x))$ is said to be γ strongly log concave if there exists a $\gamma > 0$ such that:*

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|x-y\|_2^2,$$

for all $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^n$.

Theorem 5. *Let $\mathbb{P}$ be any strongly log-concave distribution with parameter $\gamma > 0$. Then for any L-Lipschitz function with respect to the Euclidean norm:*

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq t\right] \leq 2\exp\left(\frac{-\gamma t^2}{4L^2}\right).$$

A Gaussian distribution with non-singular covariance function is strongly log-concave, so we can think of this result as a generalization of the result for IID Gaussians with a slightly worse rate. The log-concave and strongly log-concave assumption is commonly used to obtain fast rates of convergence in the literature, see Saumard and Wellner (2014) for a review on the subject.

An entire course could be made on concentration inequalities, but we will stop here for now and revisit additional concentration result for matrices down the line. But if these types of result are of interest to you, please see Boucheron et al. (2013) where a large collection such such bounds are proved and documented.

3 Linear Regression

3.1 Introduction

Linear regression is one of the first models with covariates (or features) that you have seen in your undergraduate studies. We revisit it here with a (potentially) slightly different perspective.

Most regression models can be written in the form of:

$$Y_i = f(x_i) + \epsilon_i, i = 1, \ldots, n,$$

where $f(\cdot)$ is some functional relationship and ϵ_i are some centred error terms. In this chapter we assumed $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ follows some sub-Gaussian distribution with proxy variance σ^2 and $E[\epsilon_i] = 0$, and that $f(x) = x^\top\theta$ for some $\theta \in \mathbb{R}^d$. Specifically we assume the data generating model is:

$$Y_i = x_i^\top\theta^\star + \epsilon_i, i = 1, \ldots, n,$$

Note that the sub-Gaussian assumption doesn't require the errors to be independent or identically distributed (but the dependence cannot be too large or else σ^2 won't be constant in n), so the results we will show are direct extensions of the traditional analysis performed with Gaussian errors.

We also assume that we a *fixed design* meaning that our covariates are deterministic and not random, you can think of this as equivalently conditioning the statistical analysis to the observed values of X.

3.2 Bounds on MSE

We first consider the performance of our estimated models in terms of the *Mean Square Error* (MSE), for a general regression problem this is:

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

for us this will simply to

$$MSE(X\hat{\theta}) = \frac{1}{n}\|X(\hat{\theta}_n - \theta^\star)\|_2^2,$$

where $\hat{\theta}_n$ is some estimated value for the regression parameter and $\theta^\star$ is the true data-generating value of θ.

3.2.1 Unconstrained least squares

We define the *least squares estimator* $\hat{\theta}^{LS}$ to be any vector which satisfies:

$$\hat{\theta}^{LS} \in \arg\min_{\theta\in\mathbb{R}^d} \|Y - X\theta\|_2^2,$$

this solution may or may not be unique depending on the design matrix X . You may have seen some version of the least squares solution involving an inverse of the kind $(X^\top X)^{-1}$, but even when this matrix is non-invertible we can always define *a* solution through the Moore-Penrose pseudoinverse of the matrix $X^\top X$. We denote the pseudoinverse of a matrix $A \in \mathbb{R}^{m\times n}$ as $A^\dagger$, this pseudoinverse can be thought of as providing an approximate solution to this system of equation:

$$Ax = b$$

with the property that for all $x \in \mathbb{R}^n$ $\|Ax - b\|_2 \geq \|Az - b\|_2$ for $z = A^\dagger b$; this can be thought of as providing the least squares solution to this system of equations when it cannot be solved exactly.

In the simplest scenario with no constraints, the following proposition characterizes the least squares estimator for θ:

Proposition 9. *The least squares estimator $\hat{\theta}^{LS} \in \mathbb{R}^d$ satisfies*

$$X^\top X\hat{\theta}^{LS} = X^\top Y.$$

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

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

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

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

$$\nabla_\theta |Y - X\theta|_2^2 = 0,$$

where ∇_θ is the gradient operator. We have

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

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

$$X^\top X\theta = X^\top Y.$$

□

From here on out, we use $\lesssim$ symbol to mean $<$ with all constants independent of dimensions and sample size being ommited. For example, $f(n,p) = 10\log(p)n \lesssim \log(p)n$.

Theorem 6. *Assume that the linear model holds where $\varepsilon \sim subG_n(\sigma^2)$. Then the least squares estimator $\hat{\theta}^{LS}$ satisfies*

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

where $r = rank(X^\top X)$. Moreover, for any $\delta > 0$, with probability at least $1-\delta$,

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

Proof. By definition of the least squares estimator

$$|Y - X\hat{\theta}^{\mathrm{LS}}|_2^2 \leq |Y - X\theta^*|_2^2 = |\varepsilon|_2^2.$$

Moreover,

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

Therefore,

$$|X\hat{\theta}^{LS} - X\theta^*|_2^2 \leq 2\varepsilon^\top X(\hat{\theta}^{LS} - \theta^*) = 2|X\hat{\theta}^{LS} - X\theta^*|_2\frac{\varepsilon^\top X(\hat{\theta}^{LS} - \theta^*)}{|X(\hat{\theta}^{LS} - \theta^*)|_2},$$

$$\text{and therefore: } |X\hat{\theta}^{LS} - X\theta^*|_2 \leq 2\frac{\varepsilon^\top X(\hat{\theta}^{LS} - \theta^*)}{|X(\hat{\theta}^{LS} - \theta^*)|_2}.$$

It is difficult to control

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

as $\hat{\theta}^{LS}$ depends on ε and this dependency may be complicated. To remove this dependency, we can "sup-out" $\hat{\theta}^{LS}$, note that the vector $X(\hat{\theta}^{LS} - \theta^*)/|X(\hat{\theta}^{LS} - \theta^*)|_2$ lives on a unit sphere of dimension n and we could immediate use our results of Theorem 3 to bound this quantity, but this will give us a very crude upper bound.

First we reduce the dimensionality of the problem, let $\Phi = [\phi_1, \ldots, \phi_r] \in \mathbb{R}^{n\times r}$ be an orthonormal basis of the column span of X. In particular, there exists $\nu \in \mathbb{R}^r$ such that $X(\hat{\theta}^{LS} - \theta^*) = \Phi\nu$. This gives

$$\frac{\varepsilon^\top X(\hat{\theta}^{LS} - \theta^*)}{|X(\hat{\theta}^{LS} - \theta^*)|_2} = \frac{\varepsilon^\top\Phi\nu}{|\Phi\nu|_2} = \frac{\varepsilon^\top\Phi\nu}{|\nu|_2} = \tilde{\varepsilon}^\top\frac{\nu}{|\nu|_2} \leq \sup_{u\in B_2}\tilde{\varepsilon}^\top u,$$

where B_2 is the unit ball of $\mathbb{R}^r$ and $\tilde{\varepsilon} = \Phi^\top\varepsilon$. Thus

$$\mathbb{E}[|X\hat{\theta}^{LS} - X\theta^*|_2^2] \leq \mathbb{E}[4\sup_{u\in B_2}(\tilde{\varepsilon}^\top u)^2],$$

Note that, $\tilde{\varepsilon} \sim \mathrm{subG}_r(\sigma^2)$ as well (show this as an exercise). Therefore to conclude the bound in expectation, observe that Exercise 6 yields

$$4\mathbb{E}[\sup_{u\in B_2}(\tilde{\varepsilon}^\top u)^2] = 4\sum_{i=1}^r\mathbb{E}[\tilde{\varepsilon}_i^2] \leq 16\sigma^2 r.$$

Although in the proof of the expectation bound we did not directly use Theorem 3, for the bound in probability we will need the the last step in the proof of Theorem 3 that

$$\sup_{u\in B_2}(\tilde{\varepsilon}^\top u)^2 \leq 8\log(6)\sigma^2 r + 8\sigma^2\log(1/\delta),$$

with probability $1-\delta$, □

Exercise 6. ***Moments of sub-Gaussian random variables.*** *Let X be any random variable such that*

$$\mathbb{P}[|X| > t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

then for any positive integers $k \geq 1$,

$$E[|X|^k] \leq (2\sigma^2)^{k/2} k\Gamma(k/2).$$

3.2.2 Constrained Estimation

Sometimes it is more efficient to work with constrained a *constrained* estimation problem rather than the full parameter set if we have additional information on possible solutions. A convenient and useful choice are $K \subset \mathbb{R}^d$ symmetric convex sets. If we know a priori that $\theta^* \in K$, we may prefer a constrained least squares estimator $\hat{\theta}_K^{\mathrm{LS}}$ defined by

$$\hat{\theta}_K^{\mathrm{LS}} \in \operatorname*{argmin}_{\theta \in K} |Y - \mathbb{X}\theta|_2^2.$$

The fundamental inequality used in the proof of the unconstrained estimator would still hold and the bounds on the MSE may be smaller. Indeed, we have

$$|\mathbb{X}\hat{\theta}_K^{\mathrm{LS}} - \mathbb{X}\theta^*|_2^2 \leq 2\varepsilon^\top \mathbb{X}(\hat{\theta}_K^{\mathrm{LS}} - \theta^*) \leq 2 \sup_{\theta \in K-K} (\varepsilon^\top \mathbb{X}\theta),$$

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

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

where $\mathbb{X}K = \{\mathbb{X}\theta : \theta \in K\} \subset \mathbb{R}^n$.

We consider the estimation problem constrained to an L^1 ball. This will involve controlling the complexity of the L^1 ball, similar to what we have done in Theorem 3:

Theorem 7. *Let P be a polytope with N vertices $v^{(1)}, \ldots, v^{(N)} \in \mathbb{R}^d$ and let $X \in \mathbb{R}^d$ be a random vector such that $[v^{(i)}]^\top X, i = 1, \ldots, N$, are sub-Gaussian random variables with variance proxy σ^2. Then*

$$\mathbb{E}\left[\max_{\theta \in P} \theta^\top X\right] \leq \sigma\sqrt{2\log(N)},$$

and

$$\mathbb{E}\left[\max_{\theta \in P} |\theta^\top X|\right] \leq \sigma\sqrt{2\log(2N)}.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\left(\max_{\theta \in P} \theta^\top X > t\right) \leq Ne^{-\frac{t^2}{2\sigma^2}},$$

and

$$\mathbb{P}\left(\max_{\theta \in P} |\theta^\top X| > t\right) \leq 2Ne^{-\frac{t^2}{2\sigma^2}}.$$

The proof is omitted as it is similar to some of the previous results. Recall that the L^1 ball (of radius 1) is defined by

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

and it has exactly $2d$ vertices $\mathcal{V} = \{e_1, -e_1, \dots, e_d, -e_d\}$, where e_j is the j-th vector of the canonical basis of $\mathbb{R}^d$. This implies that the set $\mathbb{X}K = \{\mathbb{X}\theta, \theta \in K\} \subset \mathbb{R}^n$ is also a polytope with at most $2d$ vertices that are in the set $\mathbb{X}\mathcal{V} = \{\mathbb{X}_1, -\mathbb{X}_1, \dots, \mathbb{X}_d, -\mathbb{X}_d\}$ where $\mathbb{X}_j$ is the j-th column of $\mathbb{X}$. Indeed, $\mathbb{X}K$ is obtained by rescaling and embedding (resp. projecting) the polytope K when $d \leq n$ (resp., $d \geq n$).

Theorem 8. *Let $\mathcal{B}_1$ be the unit ℓ_1 ball of $\mathbb{R}^d$, $d \geq 2$ and assume that $\theta^* \in \mathcal{B}_1$. Moreover, assume the conditions of Theorem 6 and that the columns of $\mathbb{X}$ are normalized such that $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$. Then the constrained least squares estimator $\hat{\theta}^{LS}_{\mathcal{B}_1}$ satisfies*

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

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

$$MSE(\mathbb{X}\hat{\theta}^{LS}_{B_1}) \lesssim \sigma\sqrt{\frac{\log(d/\delta)}{n}}.$$

The rate in the decay of MSE in the number of samples is now $\sqrt{n}$ rather than n which is worse than the unconstrained problem, however the dimension dependency is now logarithmic instead of linear; we can think of the rank of $X^\top X$ as d if the matrix is invertible.

Proof. From the same steps as in the proof of Theorem 6, we eventually arrive at

$$|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{B_1} - \mathbb{X}\theta^*|_2^2 \leq 4 \sup_{v \in \mathbb{X}K} (\varepsilon^\top v).$$

Observe now that since $\varepsilon \sim \mathrm{subG}_n(\sigma^2)$, for any column $\mathbb{X}_j$ such that $|\mathbb{X}_j|_2 \leq \sqrt{n}$, the random variable $\varepsilon^\top \mathbb{X}_j \sim \mathrm{subG}(n\sigma^2)$. Therefore, applying Theorem 7, we get the bound on $E[|\mathrm{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K)|]$ and for any $t \geq 0$,

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

To conclude the proof, we find t such that

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

which shows the second statement of the Theorem. □

Note that the proof of Theorem 6 also applies to $\hat{\theta}^{\mathrm{LS}}_{B_1}$ (exercise!) so that $\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{B_1}$ benefits from the best of both rates,

$$\mathbb{E}[|\mathrm{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{B_1})|] \lesssim \min\left(\sigma^2\frac{r}{n}, \sigma\sqrt{\frac{\log d}{n}}\right).$$

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

This type of constrained estimator may appear similar to the familiar LASSO by the duality of the optimization problem, although in this case we haven't quite estimate the "correct" radius of the constraint so this is slightly worst than the rate for LASSO which we present in the next sub-section.

3.2.3 LASSO

Sparsity can take different forms, but a popular one is to consider a vector $\theta \in \mathbb{R}^d$ with only k non-zero coordinates. Heuristically we can link this to the principal of parsimony where a less complex explanation is typically preferred over a overly complex one. We call the number of non-zero coefficients of a vector $\theta \in \mathbb{R}^d$ its ℓ_0 "norm":

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

We call a vector θ with $\ell_0 << d$ a sparse vector. More precisely, if $|\theta|_0 \leq k$, we say that θ is a k-sparse vector. We call

$$\text{supp}(\theta) = \{j \in \{1, \dots, d\} : \theta_j \neq 0\},$$

the support of θ.

Denote by $B_0(k)$ the ℓ_0 "ball" of $\mathbb{R}^d$, i.e., the set of k-sparse vectors, defined by

$$B_0(k) = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}.$$

Our goal is to control the MSE of $\hat{\theta}^{IS}_K$ when $K = B_0(k)$. Note that computing $\hat{\theta}^{IS}_{B_0(k)}$ defined as:

$$\hat{\theta}^{\text{LS}}_{B_0(k)} \in \underset{\theta \in B_0(k)}{\text{argmin}} |Y - \mathbb{X}\theta|_2^2.$$

but this would require computing $\binom{d}{k}$ least squares estimators (ask yourself why don't we need to compute $\sum_{i=1}^{k} \binom{d}{i}$ estimators) since this loss is no longer smooth due to the constraint; in fact, this number is exponentially growing in k. In practice this will be hard (or even impossible) but it is interesting to use the bounds obtained for this constrained problem as a benchmark for the LASSO and other penalized regressions.

Theorem 9. *Fix a positive integer $k \leq d/2$. Let $K = B_0(k)$ be set of k-sparse vectors of $\mathbb{R}^d$ and assume that $\theta^* \in B_0(k)$. Moreover, assume the conditions of Theorem 6. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds*

$$MSE(\mathbb{X}\hat{\theta}^{IS}_{B_0(k)}) \lesssim \frac{k\sigma^2}{n} \log\left(\frac{ed}{2k}\right) + \log(6)\frac{\sigma^2 k}{n} + \frac{\sigma^2}{n}\log(1/\delta).$$

The proof can be found in Rigollet and Hütter (2023), The rate obtained here is roughly of order $k\log(k/n)/n$ which is $n^{-1/2}$ faster than the L^1 constraint, the dependency is also logarithm d and further this is divided by k. Should we know the true sparsity level, this rate corresponds to essentially the optimal rate that we can hope for in practice.

The LASSO is the convex relaxation of the l_0 constraint problem, searching over reasonable radius of the L^1 ball to penalize and we see this will give us something closer to the optimal rate in n. Specifically the LASSO estimator is defined as:

$$\arg\min_{\theta \in \mathbb{R}^d} \frac{|Y - \mathbb{X}\theta|_2^2}{n} + \lambda_n \|\theta\|_1,$$

for some penalty τ. In practice λ is almost always chosen by cross validation, but in our theorem we will use a theoretical optimal value which depends on the unknown variance σ^2.

To obtain the "fast rate" with a scaling of order n in the MSE (rather than $n^{1/2}$), we require some additional assumptions on the design matrix.

Assumption 1. ***Assumption INC(k)*** *The design matrix $\mathbb{X}$ has incoherence k for some integer $k > 0$ if*

$$\left| \frac{\mathbb{X}^T \mathbb{X}}{n} - I_d \right|_\infty \leq \frac{1}{32k}$$

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

1. *For all $j = 1, \ldots, d$,*

$$\left| \frac{\|\mathbb{X}_j\|_2^2}{n} - 1 \right| \leq \frac{1}{32k}.$$

2. *For all $1 \leq i, j \leq d, i \neq j$, we have*

$$\frac{|\mathbb{X}_i^T \mathbb{X}_j|}{n} \leq \frac{1}{32k}.$$

We will give more intuition on why this Assumption is necessary in the subsequent section, but it has to do with the geometry of when it is possible to recover the sparse signal.

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

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

In particular $|\theta|_1 = |\theta_S|_1 + |\theta_{S^c}|_1$. The following Lemma will help us bound some key quantities that will appear in the main proof.

Lemma 4. *Fix a positive integer $k \leq d$ and assume that $\mathbb{X}$ satisfies assumption INC(k). Then, for any $S \in \{1, \ldots, d\}$ such that $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ that satisfies the cone condition*

$$|\theta_{S^c}|_1 \leq 3|\theta_S|_1,$$

it holds that

$$|\theta|_2^2 \leq 2\frac{|\mathbb{X}\theta|_2^2}{n}.$$

Theorem 10. *Fix $n \geq 2$. Assume that the linear model holds where $\varepsilon \sim subG_n(\sigma^2)$. Moreover, assume that $\|\theta^*\|_0 \leq k$ and that $\mathbf{X}$ satisfies assumption INC(k). Then the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter defined by*

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

satisfies

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

and

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}. \tag{3}$$

with probability at least $1 - \delta$.

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

$$\frac{1}{n}\|Y - \mathbf{X}\hat{\theta}^{\mathcal{L}}\|_2^2 \leq \frac{1}{n}\|Y - \mathbf{X}\theta^*\|_2^2 + \lambda_n\|\theta^*\|_1 - \lambda_n\|\hat{\theta}^{\mathcal{L}}\|_1.$$

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

$$\|\mathbf{X}\hat{\theta}^{\mathcal{L}} - \mathbf{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 2\varepsilon^\top\mathbf{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + \frac{n\lambda_n}{2}\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + n\lambda_n\|\theta^*\|_1 - n\lambda_n\|\hat{\theta}^{\mathcal{L}}\|_1.$$

Applying Hölder's inequality

$$\begin{aligned}\varepsilon^\top\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) &\leq |\varepsilon^\top\mathbb{X}|_\infty|\hat{\theta}^{\mathcal{L}} - \theta^*|_1 \\ &\leq \frac{n\lambda_n}{4}|\hat{\theta}^{\mathcal{L}} - \theta^*|_1,\end{aligned}$$

as

$$\mathbb{P}(|\mathbb{X}^\top\varepsilon|_\infty \geq t) = \mathbb{P}(\max_{1\leq j\leq d}|\mathbb{X}_j^\top\varepsilon| > t) \leq 2de^{-\frac{t^2}{4n\sigma^2}}$$

where used the fact that $|\mathbb{X}_j|_2^2 \leq n+1/(32k) \leq 2n$ and took $t = \sigma\sqrt{2n\log(2d)}+\sigma\sqrt{2n\log(1/\delta)} = n\lambda_n/2$, which implies this inequality holds with probability at least $1-\delta$. Therefore, taking $S = \text{supp}(\theta^*)$ to be the support of θ^*, we get

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

In particular, it implies that

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

so that $\theta = \hat{\theta}^{\mathcal{L}} - \theta^*$ satisfies the cone condition in Lemma 4. Using now the Cauchy-Schwarz inequality and Lemma 4 respectively, we get, since $|S| \leq k$,

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

Combining this result with Equation 4, we find

$$|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \leq 8nk\lambda_n^2.$$

This concludes the proof of the bound on the MSE. To prove the upper bound 3, we use Lemma 4 once again to get

$$|\hat{\theta}^{\mathcal{L}} - \theta^*|_2^2 \leq 2\,\text{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) \leq 16k\lambda_n^2.$$

□

3.2.4 SLOPE

By comparing the rates for LASSO and the L^0 constraint estimation problem the logarithm looks like $\log(ed/2k)$ whereas in LASSO we are missing the division by k in the logarithm. In practice this may not change things for small k, but ideally we would like to match that of the "optimal" (we haven't shown optimality officially) rate.

Instead of penalizing every coordinate with the same weight, we will now aim to make the penalty proportional to the signal size, and this will turn out to be the key insight to obtain a new optimal procedure but we need to assume the error distributions is an isotropic Gaussian.

Definition 6. *(Slope estimator). Let $\lambda = (\lambda_1, \dots, \lambda_d)$ be a non-increasing sequence of positive real numbers, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. For $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$, let $(\theta_1^*, \dots, \theta_d^*)$ be a non-increasing rearrangement of the modulus of the entries, $|\theta_1|, \dots, |\theta_d|$. We define the sorted ℓ_1 norm of θ as*

$$|\theta|_* = \sum_{j=1}^{d} \lambda_j \theta_j^*,$$

or equivalently as

$$|\theta|_* = \max_{\phi \in S_d} \sum_{j=1}^{d} \lambda_j |\theta_{\phi(j)}|.$$

The Slope estimator is then given by

$$\hat{\theta}^S \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\theta\|_2^2 + \tau |\theta|_* \right\}$$

for a choice of tuning parameters λ and $\tau > 0$.

Slope stands for Sorted L-One Penalized Estimation, and is motivated by the quest for a penalized estimation procedure that could offer a control of false discovery rate (FDR) for the hypotheses $H_{0,j} : \theta_j^* = 0$. We should check that $|\cdot|_*$ is indeed a norm and that $\hat{\theta}^S$ can be computed efficiently, for example by proximal gradient algorithms; we will have a small subsection on this in the lecture time permitting.

In what follows, we use

$$\lambda_j = \sqrt{\log(2d/j)}, \quad j = 1, \dots, d.$$

Theorem 11. *Fix $n \geq 2$. Assume that the linear model holds where $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$. Moreover, assume that $|\theta^*|_0 \leq k$ and that $\mathbb{X}$ satisfies assumption INC(k') with $k' \geq 4k \log(2de/k)$. Then the Slope estimator $\hat{\theta}^S$ with regularization parameter defined by*

$$\tau = 4\sqrt{2}\sigma \sqrt{\frac{\log(1/\delta)}{n}} \tag{5}$$

satisfies

$$MSE(\mathbb{X}\hat{\theta}^S) = \frac{1}{n} \|\mathbb{X}\hat{\theta}^S - \mathbb{X}\theta^*\|_2^2 \lesssim \sigma^2 \frac{k \log(2d/k\delta)}{n} \tag{6}$$

and

$$\|\hat{\theta}^S - \theta^*\|_2^2 \lesssim \sigma^2 \frac{k \log(2d/k) \log(1/\delta)}{n}. \tag{7}$$

with probability at least $1 - \delta$.

Therefore up to constants we see that SLOPE recovers the behaviour of best subset selection, although we need to assume a specific (isotropic and homoskedastic) error model. See Rigollet and Hütter (2023) for a proof with a suboptimal rate (but it is easier to digest than the original).

Remark 4. *If you have seen Gauss-Markov's theorem then you may be wondering how come LASSO and SLOPE beats out the unconstrained least squares estimate when it is "optimal". It's worth recalling that optimality was defined as the best linear unbiased estimator and that both LASSO and SLOPE are biased.*

3.3 Sparse set recovery

When LASSO is first introduced, it is usually seen as a method of recovering the sparse set of active coordinates and not necessarily as a way of reducing the MSE of the prediction problem. In this section we aim to prove that under assumptions on the fixed design matrix and on the minimal size of the signal, we can recover the active coordinate set with high-probability.

But first let us think of why sparse recovery is even desirable. The most immediate answer is that parsimony is a desired quality in any model, but let us consider the following example taken from Wainwright (2019) for a more exotic use of linear models:

Example 9. (Selection of Gaussian graphical models) A zero-mean Gaussian random vector $(Z_1, \ldots, Z_d)$ with a non-degenerate covariance matrix has a density of the form

$$p_{\Theta^*}(z_1, \ldots, z_d) = \frac{1}{\sqrt{(2\pi)^d \det((\Theta^*)^{-1})}} \exp\left(-\frac{1}{2} z^T \Theta^* z\right),$$

where $\Theta^* \in \mathbb{R}^{d \times d}$ is the inverse covariance matrix, also known as the precision matrix. For many interesting models, the precision matrix is sparse, with relatively few non-zero entries. The problem of Gaussian graphical model selection is to infer the non-zero entries in the matrix Θ^*.

It turns out this problem can be reduced to a sparse linear regression problem. For a given index $s \in V := \{1, 2, \ldots, d\}$, suppose that we are interested in recovering its neighborhood, meaning the subset

$$\mathcal{N}(s) := \{t \in V \mid \Theta^*_{st} \neq 0\}.$$

In order to do so, imagine performing a linear regression of the variable Z_s on the $(d-1)$-dimensional vector $Z_{\setminus\{s\}} := \{Z_t, t \in V \setminus \{s\}\}$. We can write

$$\underbrace{Z_s}_{\text{response } y} = \langle \underbrace{Z_{\setminus\{s\}}}_{\text{predictors}}, \theta^* \rangle + w_s,$$

where w_s is a zero-mean Gaussian variable, independent of the vector $Z_{\setminus\{s\}}$. Moreover, the vector $\theta^* \in \mathbb{R}^{d-1}$ has the same sparsity pattern as the sth off-diagonal row

$$(\Theta^*_{st}, t \in V \setminus \{s\})$$

of the precision matrix.

Before we begin with the sparse recovery of the noisy version of the problem, let us think about the deterministic version of the problem. Suppose that you are asked to solve the following system of equations for θ:

$$X\theta = Y,$$

with $X \in \mathbb{R}^{n \times d}$, $\theta \in \mathbb{R}^d$ and $Y \in \mathbb{R}^n$ with $d > n$, then it is impossible to find a unique solution to this system of equation. But if you are certain that the θ vector is k-sparse, for some $k < n$, then you have some hope of having an unique solution. The brute force approach would be to manually try to solve the following constrained system

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \text{ such that } X\theta = Y,$$

we instead consider the convex relaxation of this problem by using:

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \text{ such that } X\theta = Y,$$

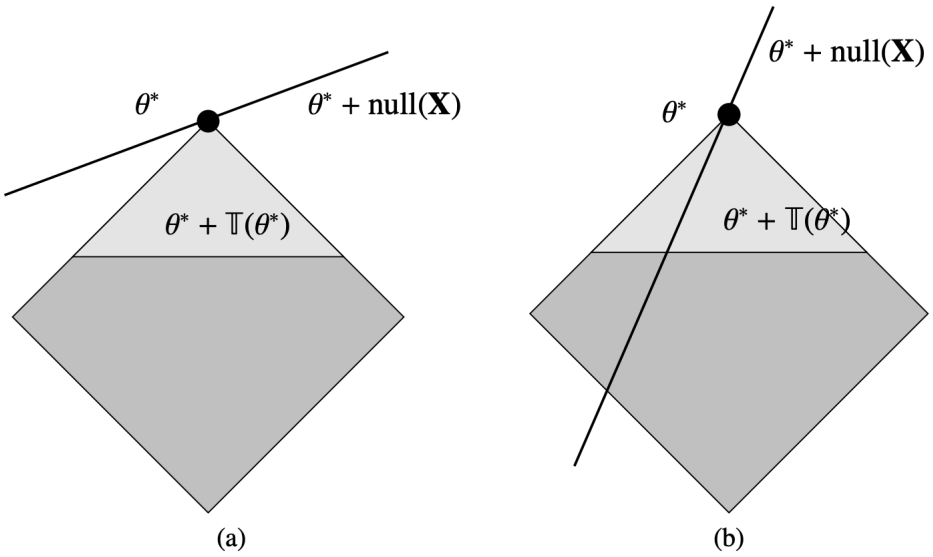

Figure 2: Figure taken from Wainwright (2019) Chapter 7. (a) shows the favorable case where the set $\theta^\star + null(X)$ only intersects the tangent cone at $\theta^\star$, so that any other solution will have greater L^1 norm. (b) shows the unfavorable case as the minimum L^1 solution will lie in the interior of the tangent cone meaning it won't be $\theta^\star$, nor will it necessarily have the correct sparsity pattern.

this is called *basis pursuit* and it is an example of a Linear Program (LP).

If the true $\theta^\star$ has support S, then intuitively the success of basis pursuit should depend on how the nullspace of $\mathbf{X}$ is related to this support, as well as the geometry of the ℓ_1-ball. The nullspace of $\mathbf{X}$ is given by $\text{null}(\mathbf{X}) := \{\Delta \in \mathbb{R}^d \mid \mathbf{X}\Delta = 0\}$. Since $\mathbf{X}\theta^* = y$ by assumption, any vector $\theta^* + \Delta$ for some $\Delta \in \text{null}(\mathbf{X})$ is s solution to the basis pursuit program. Consider the *tangent cone* of the ℓ_1-ball at θ^*, given by

$$\mathbb{T}(\theta^*) = \{\Delta \in \mathbb{R}^d \mid \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}. \tag{8}$$

As illustrated in Figure 2, this set captures the set of all directions relative to θ^* along which the ℓ_1-norm remains constant or decreases. The set $\theta^* + \text{null}(\mathbf{X})$, drawn with a solid line in Figure 2, corresponds to the set of all vectors that are feasible for the basis pursuit LP. Consequently, if θ^* is the unique optimal solution of the basis pursuit LP, then it must be the case that the intersection of the nullspace $\text{null}(\mathbf{X})$ with this tangent cone contains only the zero vector. This favorable case is shown in Figure 2(a), whereas Figure 2(b) shows the non-favorable case, in which θ^* need not be optimal.

Using Equation 8 let us derive a condition under which recovery is possible; a condition which will not rely on knowledge of the true value of $\theta^\star$. For any $t > 0$

$$\|\theta^*\|_1 - \|t\Delta_S\|_1 + \|t\Delta_{S^c}\| \leq \|\theta^* + t\Delta_S\|_1 + \|t\Delta_{S^c}\|_1 = \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1,$$
$$\text{therefore } \|\Delta_{S^c}\| \leq \|\Delta_S\|_1$$

would imply that Equation 8 holds; let us define the set:

$$C(S) := \{\delta \in \mathbb{R}^d : \|\Delta_{S^c}\| \leq \|\Delta_S\|_1\}$$

Definition 7. *The matrix X satisfies the restricted nullspace property with respect to S if $C(S) \cap null(X) = \{0\}$.*

The restricted nullspace property is equivalent to the success of the basis pursuit LP in the following sense:

Theorem 12. *The following two properties are equivalent:*

(a) For any vector $\theta^* \in \mathbb{R}^d$ with support S, the basis pursuit program applied with $y = X\theta^*$ has unique solution $\hat{\theta} = \theta^*$.

(b) The matrix X satisfies the restricted nullspace property with respect to S.

Proof. We first show that (b) $\Rightarrow$ (a). Since both $\hat{\theta}$ and θ^* are feasible for the basis pursuit program, and since $\hat{\theta}$ is optimal, we have $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$. Defining the error vector $\tilde{\Delta} := \hat{\theta} - \theta^*$, we have

$$\|\theta^*_S\|_1 = \|\theta^*\|_1 \geq \|\theta^* + \tilde{\Delta}\|_1 = \|\theta^*_S + \tilde{\Delta}_S\|_1 + \|\tilde{\Delta}_{S^c}\|_1 \geq \|\theta^*_S\|_1 - \|\tilde{\Delta}_S\|_1 + \|\tilde{\Delta}_{S^c}\|_1,$$

where we have used the fact that $\theta^*_{S^c} = 0$, and applied the triangle inequality. Rearranging this inequality, we conclude that the error $\tilde{\Delta} \in C(S)$. However, by construction, we also have $X\tilde{\Delta} = 0$, so $\tilde{\Delta} \in \text{null}(X)$ as well. By our assumption, this implies that $\tilde{\Delta} = 0$, or equivalently that $\hat{\theta} = \theta^*$.

In order to establish the implication (a) $\Rightarrow$ (b), it suffices to show that, if the ℓ_1-relaxation succeeds for all S-sparse vectors, then the set $\text{null}(X) \setminus \{0\}$ has no intersection with $C(S)$. For a given vector $\theta^* \in \text{null}(X) \setminus \{0\}$, consider the basis pursuit problem

$$\min_{\beta \in \mathbb{R}^d} \|\beta\|_1 \quad \text{such that } X\beta = X \begin{bmatrix} \theta^*_S \\ 0 \end{bmatrix}. \tag{7.11}$$

By assumption, the unique optimal solution will be $\bar{\beta} = [\theta^*_S \quad 0]^T$. Since $X\theta^* = 0$ by assumption, the vector $[0 \quad -\theta^*_{S^c}]^T$ is also feasible for the problem, and, by uniqueness, we must have $\|\theta^*_S\|_1 < \|\theta^*_{S^c}\|_1$, implying that $\theta^* \notin C(S)$ as claimed. □

Checking the restricted null space condition is difficult as it requires knowledge of the correct sparse set S, so in general easier to verify sufficient conditions are needed.

Proposition 10. *If*

$$\left| \frac{\mathbb{X}^T \mathbb{X}}{n} - I_d \right|_\infty \leq \frac{1}{3k},$$

then the restrict null space condition holds for all subsets of cardinality at most k.

Note that this guarantees that the restricted nullspace condition holds uniformly for all possible sets of cardinality k or less, meaning that it doesn't require specific knowledge of the correct set S. The above is the incoherence condition that we used in the proof of the LASSO MSE guarantees; the constant changed from 32 to 3 due to the noise in the estimation problem. Historically the incoherence condition was the first sufficient condition which was introduced. Unfortunately the $|\cdot|_\infty$ (maximum of the entries) is not a submultiplicative norm, which makes it harder to work with, another similar condition is the restricted isometry property:

Definition 8. *For an integer k, $X \in \mathbb{R}^{n \times d}$ satisfies a restricted isometry property of order k with constant $\delta_k(X) > 0$ if*

$$\left\| \frac{X_S^\top X_S}{n} - I_k \right\|_{op} \leq \delta_k(X)$$

for all subsets of size at most k.

Proposition 11. *If the RIP constant of order $2k$ is bounded as $\delta_{2k} < 1/3$, then the restricted null space condition holds for any subset S of cardinality $|S| \leq k$.*

For a proof of this please see Wainwright (2019) Chapter 7 (proposition 7.11). In general there are many more conditions on the design which are sufficient to show the restricted null space condition, all of whom are not necessary. On a last note, the conditions needed for the noisy version of the estimation needs to be stronger than that of the deterministic version. See Wainwright (2019) Chapter 7 for discussion on other conditions such as the restricted eigenvalue condition and others.

Now let us consider when it is possible recover the correct set of active coordinates for LASSO; we will see one more condition on the design matrix along the way.

Assumption 2. *The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below:*

$$\gamma_{\min}\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right) \geq c_{\min} > 0.$$

Assumption 3. *There exists some $\alpha \in [0, 1)$ such that*

$$\max_{j \in S_c} ||(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T X_j||_1 \leq \alpha.$$

Assumption 2 is very mild: in fact, it would be required in order to ensure that the model is identifiable, even if the support set S were known a priori.

Assumption 3 (mutual incoherence) is a more difficult to interpret. Suppose that we want to predict the column vector X_j using a linear combination of the columns of $\mathbf{X}_S$. The best weight vector $\tilde{\omega} \in \mathbb{R}^{|S|}$ is

$$\tilde{\omega} = \arg\min_{\omega \in \mathbb{R}^{|S|}} ||X_j - \mathbf{X}_S\omega||_2^2 = (\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T X_j.$$

The mutual incoherence condition bounds $||\tilde{\omega}||_1$. If the column space of $\mathbf{X}_S$ were orthogonal to X_j, then the optimal weight vector $\tilde{\omega}$ would be identically zero. In general, we cannot expect this orthogonality to hold, but the mutual incoherence condition imposes a type of approximate orthogonality. Let

$$\Pi_{S^\perp}(\mathbf{X}) = I_n - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T,$$

a type of orthogonal projection matrix.

Theorem 13. *Consider an S-sparse linear regression model for which the design matrix satisfies Assumptions 2 and 3. Then for any choice of regularization parameter such that*

$$\lambda_n \geq \frac{2}{1-\alpha}\left\|X_S^T\Pi_{S^\perp}(\mathbf{X})\frac{\epsilon}{n}\right\|_\infty, \tag{9}$$

the Lasso program has the following properties:

(a) ***Uniqueness:*** *There is a unique optimal solution $\hat{\theta}$.*

(b) ***No false inclusion:*** *This solution has its support set $\hat{S}$ contained within the true support set S.*

(c) ***ℓ_∞-bounds:*** *The error $\hat{\theta} - \theta^*$ satisfies*

$$\|\hat{\theta}_S - \theta_S^*\|_\infty \leq \underbrace{\left\|\left(\frac{X_S^TX_S}{n}\right)^{-1}X_S^T\frac{\epsilon}{n}\right\|_\infty + \left\|\left(\frac{X_S^TX_S}{n}\right)^{-1}\right\|_\infty\lambda_n}_{B(\lambda_n,\mathbf{X})},$$

where $\|A\|_\infty = \max_{i=1,\ldots,s}\sum_j |A_{i,j}|$ is the matrix ℓ_∞-norm.

(d) ***No false exclusion:*** *The Lasso includes all indices $i \in S$ such that $|\theta_i^*| > B(\lambda_n; \mathbf{X})$, and hence is variable selection consistent if $\min_{i \in S} |\theta_i^*| > B(\lambda_n; \mathbf{X})$.*

As ϵ is a random vector, we need to bound the probability of it satisfying the requirements of Theorem 13 if we want a meaningful statistical statement.

Corollary 1. *For a S-sparse linear model based on a noise vector ϵ with zero-mean i.i.d. σ-sub-Gaussian entries, and a deterministic design matrix $\mathbf{X}$ that satisfies assumptions 2 and 3, as well as the C-column normalization condition $\max_{j=1,\ldots,d} \|\mathbf{X}_j\|_2/\sqrt{n} \leq C$. Suppose that we solve the Lasso program with regularization parameter*

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2\log(d-k)}{n}} + \delta\right\}$$

for some $\delta > 0$. Then the optimal solution $\widehat{\theta}$ is unique with its support contained within S, and satisfies the ℓ_∞-error bound

$$\|\widehat{\theta}_S - \theta_S^*\|_\infty \leq \frac{\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{2\log s}{n}} + \delta\right) + \left\|\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right)^{-1}\right\|_\infty \lambda_n, \tag{10}$$

all with probability at least $1 - 4e^{-n\delta^2/2}$.

Proof. We first verify that the given choice of regularization parameter satisfies the bound (9) with high probability. It suffices to bound the maximum absolute value of the random variables

$$Z_j := X_j^T\left[\mathbf{I}_n - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T\right]\left(\frac{\epsilon}{n}\right), \quad \text{for } j \in S^c.$$

Since $\Pi_{S^\perp}(\mathbf{X}) = \mathbf{I}_n - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T$ is an orthogonal projection matrix, we have

$$\|\Pi_{S^\perp}(\mathbf{X})X_j\|_2 \leq \|X_j\|_2 \leq C\sqrt{n},$$

by the column normalization assumption. Therefore, each variable Z_j is sub-Gaussian with parameter at most $C^2\sigma^2/n$. From standard sub-Gaussian tail bounds, we have

$$\mathbb{P}\left(\max_{j \in S^c}|Z_j| \geq t\right) \leq 2(d-s)e^{-nt^2/2C^2\sigma^2},$$

from which we see that our choice of λ_n ensures that the bound holds with the claimed probability. The only remaining step is to simplify the ℓ_∞-bound. The second term in this bound is a deterministic quantity, so we focus on bounding the first term. For each $i = 1, \ldots, s$, consider the random variable

$$\widetilde{Z}_i := e_i^T\left(\frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S\right)^{-1}\mathbf{X}_S^T\epsilon/n.$$

Since the elements of the vector w are i.i.d. σ-sub-Gaussian, the variable $\widetilde{Z}_i$ is zero-mean and sub-Gaussian with parameter at most

$$\frac{\sigma^2}{n}\left\|\left(\frac{1}{n}\mathbf{X}_S^T\mathbf{X}_S\right)^{-1}\right\|_2 \leq \frac{\sigma^2}{c_{\min}n}$$

where we have used the eigenvalue condition in Assumption 2. Consequently, for any $\delta > 0$, we have

$$\mathbb{P}\left(\max_{i=1,\ldots,s}|\widetilde{Z}_i| > \frac{\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{2\log s}{n}} + \delta\right)\right) \leq 2e^{-n\delta^2/2},$$

from which the claim follows. □

Let us now prove Theorem 13. To do so we will use the Primal Dual Witness method which we elaborate on below. But first there are some technical details to review. We need to work in terms of the subdifferential of the ℓ_1-norm given that it is not differentiable at 0. Given a convex function $f\colon \mathbb{R}^d \to \mathbb{R}$, we say that $z \in \mathbb{R}^d$ is a subgradient of f at θ, denoted by $z \in \partial f(\theta)$, if we have

$$f(\theta + \Delta) \geq f(\theta) + \langle z, \Delta \rangle \quad \text{for all } \Delta \in \mathbb{R}^d.$$

When $f(\theta) = ||\theta||_1$, it can be seen that $z \in \partial||\theta||_1$ if and only if $z_j = \text{sign}(\theta_j)$ for all $j = 1, 2, \ldots, d$. Here we allow sign(0) to be any number in the interval $[-1, 1]$. For the LASSO program, we say that a pair $(\overline{\theta}, \overline{z}) \in \mathbb{R}^d \times \mathbb{R}^d$ is primal-dual optimal if $\overline{\theta}$ is a minimizer and $\hat{z} \in \partial||\overline{\theta}||_1$. Any such pair must satisfy the zero-subgradient condition

$$\frac{1}{n} X^T (X\overline{\theta} - y) + \lambda_n \hat{z} = 0,$$

which is the analog of a zero-gradient condition in the non-differentiable setting.

The primal-dual witness method constructs a pair $(\overline{\theta}, \hat{z})$ satisfying the zero-subgradient condition, and such that $\overline{\theta}$ has the correct signed support. When this procedure succeeds, the constructed pair is primal-dual optimal, and acts as a witness for the fact that the Lasso has a unique optimal solution with the correct signed support.

Definition 9. ***Primal–dual witness (PDW) construction:***

1. *Set* $\widehat{\theta}_{S^c} = 0$.
2. *Determine* $(\widehat{\theta}_S, \widehat{z}_S) \in \mathbb{R}^s \times \mathbb{R}^s$ *by solving the oracle subproblem*

$$\widehat{\theta}_S \in \arg\min_{\theta_S \in \mathbb{R}^s} \left\{ \underbrace{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}_S \theta_S\|_2^2 + \lambda_n \|\theta_S\|_1}_{=: f(\theta_S)} \right\}, \tag{11}$$

   and then choosing $\widehat{z}_S \in \partial\|\widehat{\theta}_S\|_1$ *such that* $\nabla f(\theta_S)\big|_{\theta_S = \widehat{\theta}_S} + \lambda_n \widehat{z}_S = 0$.

3. *Solve for* $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ *via the zero-subgradient equation, and check whether or not the strict dual feasibility condition* $\|\widehat{z}_{S^c}\|_\infty < 1$ *holds.*

Note that the vector $\widehat{\theta}_{S^c} \in \mathbb{R}^{d-s}$ is determined in step 1, whereas the remaining three subvectors are determined in steps 2 and 3. By construction, the subvectors $\widehat{\theta}_S$, $\widehat{z}_S$, and $\widehat{z}_{S^c}$ satisfy the zero-subgradient condition. Using the fact that $\widehat{\theta}_{S^c} = \theta^*_{S^c} = 0$ and writing out this condition in block matrix form, we obtain

$$\frac{1}{n}\begin{bmatrix} \mathbf{X}_S^T \mathbf{X}_S & \mathbf{X}_S^T \mathbf{X}_{S^c} \\ \mathbf{X}_{S^c}^T \mathbf{X}_S & \mathbf{X}_{S^c}^T \mathbf{X}_{S^c} \end{bmatrix} \begin{bmatrix} \widehat{\theta}_S - \theta^*_S \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \mathbf{X}_S^T w \\ \mathbf{X}_{S^c}^T w \end{bmatrix} + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{12}$$

We say that the PDW construction succeeds if the vector $\widehat{z}_{S^c}$ constructed in step 3 satisfies the strict dual feasibility condition. The following result shows that this success acts as a witness for the Lasso in the sense that the solution obtained from the PDW construction matches that obtained from the LASSO program without knowledge of S:

Lemma 5. *If the lower eigenvalue condition (A3) holds, then success of the PDW construction implies that the vector* $(\widehat{\theta}_S, 0) \in \mathbb{R}^d$ *is the unique optimal solution of the Lasso.*

Proof. When the PDW construction succeeds, then $\widehat{\theta} = (\widehat{\theta}_S, 0)$ is an optimal solution with associated subgradient vector $\widehat{z} \in \mathbb{R}^d$ satisfying $\|\widehat{z}_{S^c}\|_\infty < 1$, and $\langle \widehat{z}, \widehat{\theta} \rangle = \|\widehat{\theta}\|_1$. Now let $\widetilde{\theta}$ be any other optimal solution. If we introduce the shorthand notation $F(\theta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\theta\|_2^2$, then we are guaranteed that

$$F(\widehat{\theta}) + \lambda_n \langle \widehat{z}, \widehat{\theta} \rangle = F(\widehat{\theta}) + \lambda_n \|\widehat{\theta}\|_1,$$

and hence

$$F(\widehat{\theta}) - \lambda_n \langle \widehat{z}, \widetilde{\theta} - \widehat{\theta} \rangle = F(\widetilde{\theta}) + \lambda_n (\|\widetilde{\theta}\|_1 - \langle \widehat{z}, \widetilde{\theta} \rangle).$$

But by the zero-subgradient conditions, we have $\lambda_n \widehat{z} = -\nabla F(\widehat{\theta})$, which implies that

$$F(\widehat{\theta}) + \langle \nabla F(\widehat{\theta}), \widetilde{\theta} - \widehat{\theta} \rangle - F(\widetilde{\theta}) = \lambda_n (\|\widetilde{\theta}\|_1 - \langle \widehat{z}, \widetilde{\theta} \rangle).$$

By convexity of F, the left-hand side is negative, which implies that $\|\widetilde{\theta}\|_1 \leq \langle \widehat{z}, \widetilde{\theta} \rangle$. But since we also have $\langle \widehat{z}, \widetilde{\theta} \rangle \leq \|\widetilde{\theta}\|_1$, we must have $\|\widetilde{\theta}\|_1 = \langle \widehat{z}, \widetilde{\theta} \rangle$. Since $\|\widehat{z}_{S^c}\|_\infty < 1$, this equality can only occur if $\widetilde{\theta}_j = 0$ for all $j \in S^c$.

Thus, all optimal solutions are supported only on S, and hence can be obtained by solving the oracle subproblem. Given the lower eigenvalue condition, this subproblem is strictly convex, and so has a unique minimizer. □

To prove Theorem 13, it suffices to show that the vector $\widetilde{z}_{sc} \in \mathbb{R}^{d-s}$ constructed in step 3 of the PDW approach satisfies the strict dual feasibility condition. Using the zero-subgradient conditions, we can solve for the vector $\widetilde{z}_{S^c} \in \mathbb{R}^{d-s}$

$$\widehat{z}_{S^c} = -\frac{1}{\lambda_n n} X_{S^c}^T X_S (\widehat{\theta}_S - \theta_s^*) + X_{S^c}^T \left(\frac{\epsilon}{\lambda_n n} \right). \tag{13}$$

Similarly, using the assumed invertibility of $X_S^T X_S$ in order to solve for the difference $\widehat{\theta}_S - \theta_S^*$ yields

$$\widehat{\theta}_S - \theta_s^* = (X_S^T X_S)^{-1} X_S^T \epsilon - \lambda_n n (X_S^T X_S)^{-1} \widehat{z}_S.$$

Substituting this expression back into equation Equation 13 and simplifying yields

$$\widehat{z}_{S^c} = \underbrace{X_{S^c}^T X_S (X_S^T X_S)^{-1} \widehat{z}_S}_{\mu} + \underbrace{X_{S^c}^T [I - X_S (X_S^T X_S)^{-1} X_S^T] \left(\frac{\epsilon}{\lambda_n n} \right)}_{V_{S^c}}.$$

By the triangle inequality, $\|\widehat{z}_{S^c}\|_\infty \leq \|\mu\|_\infty + \|V_{S^c}\|_\infty$. By the mutual incoherence condition $\|\mu\|_\infty \leq \alpha$. By our choice of regularization parameter, $\|V_{S^c}\|_\infty \leq \frac{1}{2}(1 - \alpha)$. Combining these bounds, we conclude that $\|\widehat{z}_{S^c}\|_\infty \leq \frac{1}{2}(1 + \alpha) < 1$, which establishes the strict dual feasibility condition.

It remains to establish a bound on the ℓ_∞-norm of the error $\widehat{\theta}_S - \theta_s^*$. From our expression for $\widehat{\theta}_S - \theta_s^*$ and the triangle inequality, we have

$$\|\widehat{\theta}_S - \theta_s^*\|_\infty \leq \left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} X_S^T \frac{\epsilon}{n} \right\|_\infty + \left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right\|_\infty \lambda_n,$$

which completes the proof.

3.4 Proximal Gradient Descent

The reason why we choose these convex losses is that they are easier to handle from an optimization point of view; it is worth introducing some (or at least one) of these optimization methods. One of the first proposed numerical procedure to obtain the SLOPE estimator is to use proximal gradient descent which differs from the usual approached used for LASSO, this is worth discussing in case you have not seen it. [...]

4 Matrix Concentration

Matrices contains additional structure compared to vectors and scalars and specialized techniques are needed to obtain concentration for sums of random matrices. The most immediate application of matrix concentration will be covariance estimation but there are other less direct uses such as spectral clustering for graphs.

4.1 Basic concentration

Recall that we have shown a random sub-Gaussian matrix concentrates towards its expectation (as the singular value are 1-Lipschitz functions of the matrix), but we have yet to show how large this expectation can be. In this section we will show a bound on the singular value and use this to perform *spectral clustering* on a stochastic block model. Let us review some linear algebra first. We can characterize singular values and eigenvalues of matrices in a variational manner via the Courant Fisher min-max theorem, which states that:

$$\lambda_i(A) = \max_{dim(E)=i} \min_{x \in S(E)} x^\top A x$$

where the maximum is taken over all i-dimensional subspaces E of $\mathbb{R}^n$ Using this we can characterize the operator norm or the maximum singular value as follows:

$$\|A\| := \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in S^{n-1}} \|Ax\|_2.$$

Equivalently, the operator norm of A can be computed by maximizing the quadratic form $\langle Ax, y \rangle$ over all unit vectors x, y:

$$\|A\| = \max_{x \in S^{n-1},\, y \in S^{m-1}} \langle Ax, y \rangle.$$

Since the largest singular value or operator norm is a maximum over an uncountable collection, we will need to use some familiar tricks to reduce this problem into a finite one through ϵ-net arguments.

Lemma 6. *Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any ε-net $\mathcal{N}$ of the sphere S^{n-1}, we have*

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1-\varepsilon} \sup_{x \in \mathcal{N}} \|Ax\|_2.$$

Proof. The lower bound in the conclusion is simple as $\mathcal{N} \subset S^{n-1}$. To prove the upper bound, fix the unit vector $x \in S^{n-1}$ which achieves the maximum $\max_{x \in S^{n-1}} \|Ax\|_2$, i.e.

$$\|A\| = \|Ax\|_2$$

and choose $x_0 \in \mathcal{N}$ that approximates x so that

$$\|x - x_0\|_2 \leq \varepsilon.$$

By the definition of the operator norm, this implies

$$\|Ax - Ax_0\|_2 = \|A(x - x_0)\|_2 \leq \|A\|\|x - x_0\|_2 \leq \varepsilon\|A\|.$$

Using the triangle inequality, we find that

$$\|Ax_0\|_2 \geq \|Ax\|_2 - \|Ax - Ax_0\|_2 \geq \|A\| - \varepsilon\|A\| = (1 - \varepsilon)\|A\|.$$

Dividing both sides of this inequality by $1 - \varepsilon$, we complete the proof. □

You would think this Lemma would be sufficient, but it turns our that it is much better to work with the following formulation instead:

Lemma 7. *Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any ε-net $\mathcal{N}$ of the sphere S^{n-1} and any ε-net $\mathcal{M}$ of the sphere S^{m-1}, we have*

$$\sup_{x \in \mathcal{N}} \sup_{y \in \mathcal{M}} y^\top Ax \leq \|A\|_{op} \leq \frac{1}{1 - 2\varepsilon} \sup_{x \in \mathcal{N}} \sup_{y \in \mathcal{M}} y^\top Ax.$$

Think about why Lemma 6 would have been a bad idea in the proof to follow. The following theorem states that the norm of an $m \times n$ random matrix A with independent sub-gaussian entries satisfies

$$\|A\|_{op} \lesssim \sqrt{m} + \sqrt{n}$$

with high probability, so roughly the square root of max of n or m. It is worth remembering that for a deterministic matrix, the best general upper bound we can hope for is:

$$\|A\|_{op} \leq \|A\|_F = O(\sqrt{nm}),$$

which is growing faster than that of a random matrix. Although these are upper bounds, if they are taken as tight, then in general a random matrix is "smaller" than a fixed matrix.

Theorem 14. *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, σ-sub-gaussian random variables. Then, for any $t > 0$ we have*

$$\|A\| \leq C\sigma\left(\sqrt{m} + \sqrt{n} + t\right)$$

with probability at least $1 - 2\exp(-t^2)$ for some constant $C > 0$.

Proof. This proof is based on an ε-net argument. We need to control $\langle Ax, y\rangle$ for all vectors x and y on their respective unit spheres. To this end, we discretize each sphere using a net (approximation step), establish a tight control of $\langle Ax, y\rangle$ for fixed vectors x and y from the net (concentration step), and finish by taking a union bound over all x and y in the net.

Step 1: Approximation. Choose $\varepsilon = 1/4$. We can use Lemma 2 to find an ε-net $\mathcal{N}$ of the sphere S^{n-1} and an ε-net $\mathcal{M}$ of the sphere S^{m-1} with cardinalities

$$|\mathcal{N}| \leq 9^n \quad \text{and} \quad |\mathcal{M}| \leq 9^m.$$

By Lemma 7, the operator norm of A can be bounded using these nets as follows:

$$\|A\| \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y\rangle.$$

Step 2: Concentration. Fix $x \in \mathcal{N}$ and $y \in \mathcal{M}$. Then the quadratic form

$$\langle Ax, y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}x_i y_j$$

is a sum of independent, sub-gaussian random variables states that the sum is sub-gaussian, and has proxy variance:

$$\leq C\sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{m} x_i^2 y_j^2 = C\sigma^2 \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{j=1}^{m} y_j^2\right) = C\sigma^2.$$

Recalling the usual sub-Gaussian bound, we can restate this as the tail bound

$$\mathbb{P}\left\{\langle Ax, y\rangle \geq u\right\} \leq 2\exp(-u^2/C\sigma^2), \quad u \geq 0.$$

Step 3: Union bound. Next, we unfix x and y using a union bound. Suppose the event $\max_{x\in\mathcal{N}, y\in\mathcal{M}}\langle Ax, y\rangle \geq u$ occurs. Then there exist $x \in \mathcal{N}$ and $y \in \mathcal{M}$ such that $\langle Ax, y\rangle \geq u$. Thus the union bound yields

$$\mathbb{P}\left\{\max_{x\in\mathcal{N}, y\in\mathcal{M}}\langle Ax, y\rangle \geq u\right\} \leq \sum_{x\in\mathcal{N}, y\in\mathcal{M}} \mathbb{P}\left\{\langle Ax, y\rangle \geq u\right\}.$$

Using the tail bound obtained in step 2 and the estimate on the sizes of $\mathcal{N}$ and $\mathcal{M}$, we bound the probability above by

$$9^{n+m} \cdot 2\exp(-u^2/C\sigma^2).$$

Choose

$$u = C\sigma(\sqrt{m} + \sqrt{n} + t).$$

Then $u^2 \geq C^2\sigma^2(n + m + t^2)$, and if the constant C is chosen sufficiently large, the exponent in is large enough, say $u^2/C\sigma^2 \geq 3(n + m) + t^2$. Thus

$$\mathbb{P}\left\{\max_{x\in\mathcal{N}, y\in\mathcal{M}}\langle Ax, y\rangle \geq u\right\} \leq 9^{n+m} \cdot 2\exp\left(-3(n + m) - t^2\right) \leq 2\exp(-t^2).$$

Finally, combining all three steps together we conclude that

$$\mathbb{P}\left\{\|A\| \geq 2u\right\} \leq 2\exp(-t^2).$$

Recalling our choice of u, which matches the bound used in the statement of the theorem we complete the proof. □

4.2 Stochastic Block Model

The stochastic block model is a generalization of a completely random graph. It is based on the fact that observed graphs tend to have clusters, think about friendship networks, some types of people are more likely to be friends than others. In the following example taken from Vershynin (2018), we will work on deriving guarantees for clustering groups of people into their respective communities. In particular we start with 2 even communities to make the math simple.

This model divides n vertices into two sets ("communities") of sizes $n/2$ each. Construct a random graph G by connecting every pair of vertices independently with probability p if they

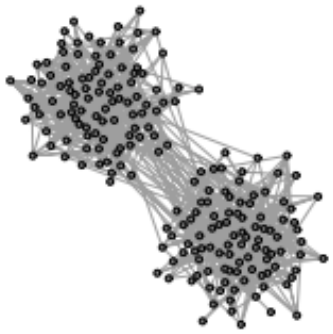

Figure 3: A random graph generated according to the stochastic block model $G(n, p, q)$ with $n = 200$, $p = 1/20$ and $q = 1/200$.

belong to the same community and q if they belong to different communities. This distribution on graphs is called the *stochastic block model* and is denoted $G(n, p, q)$.

In the special case where $p = q$ we obtain the Erdős-Rényi model $G(n, p)$. But we assume that $p > q$ here. In this case, edges are more likely to occur within than across communities. This gives the network a community structure; see Figure 3.

We can identify a graph G by its adjacency matrix A. For a random graph $G \sim G(n, p, q)$, the adjacency matrix A is a random matrix, and we will examine A using the tools we developed on concentration for matrices. Note that to simplify the discussion, we allow nodes to form self-directed edges with probability p, this would not be the case in practical observed graphs.

We split A into deterministic and random parts,

$$A = D + R,$$

where D is the expectation of A. We may think about D as an informative part (the "signal") and R as "noise".

To see why D is informative, let us compute its eigenstructure. The entries A_{ij} have a Bernoulli distribution; they are either $\mathrm{Ber}(p)$ or $\mathrm{Ber}(q)$ depending on the community membership of vertices i and j. Thus the entries of D are either p or q, depending on the membership. For illustration, if we group the vertices that belong to the same community together, then for $n = 4$ the matrix D will look like:

$$D = \mathbb{E}A = \begin{bmatrix} p & p & q & q \\ p & p & q & q \\ q & q & p & p \\ q & q & p & p \end{bmatrix}.$$

For arbitrary n this matrix D has rank 2, and the non-zero eigenvalues λ_i and the corresponding eigenvectors u_i

$$\lambda_1 = \left(\frac{p+q}{2}\right) n, \quad u_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

$$\lambda_2 = \left(\frac{p-q}{2}\right) n, \quad u_2 = \begin{bmatrix} 1 \\ \vdots \\ - \\ \vdots \\ -1 \end{bmatrix}. \tag{4.17}$$

The important object here is the second eigenvector u_2. It contains all information about the community structure. If we knew u_2, we would identify the communities precisely based on the sizes of coefficients of u_2. But we do not know $D = \mathbb{E}A$, and so we do not have access to u_2. Instead, we know $A = D + R$, a noisy version of D. The level of the signal D is

$$\|D\| = \lambda_1 \asymp n$$

while the level of the noise $R = A - D$ can be estimated using

$$\|R\| \leq C\sqrt{n} \quad \text{with probability at least } 1 - 4e^{-n},$$

as every entry of the adjacency matrix is independent and bounded between $[-1, 1]$, since A only has $0, 1$ entries and $D's$ entries are between $[0, 1]$.

Thus, for large n, the noise R is much smaller than the signal D. In other words, A is close to D, and thus we should be able to use A instead of D to extract the community information. This can be justified using the classical perturbation theory for matrices.

The following theorem is very helpful in determining how well we can estimate the second eigenvector.

Theorem 15. *(Davis-Kahan) Let S and T be symmetric matrices with the same dimensions. Fix i and assume that the i-th largest eigenvalue of S is well separated from the rest of the spectrum:*

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0.$$

Then the angle between the eigenvectors of S and T corresponding to the i-th largest eigenvalues (as a number between 0 and $\pi/2$) satisfies

$$\sin \angle(v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}.$$

The conclusion of the Davis-Kahan theorem implies that the unit eigenvectors $v_i(S)$ and $v_i(T)$ are close to each other up to a sign, namely

$$\exists \theta \in \{-1, 1\} : \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{2^{3/2}\|S - T\|}{\delta}.$$

Our intended method is to use the sign of the estimated second eigenvector to cluster each observation into their respective communities, we use Davis-Kahan to bound the amount of mistakes we will make with high-probability.

Let us apply the Davis-Kahan Theorem for $S = D$ and $T = A = D + R$, and for the second largest eigenvalue. We need to check that λ_2 is well separated from the rest of the spectrum of D, that is from 0 and λ_1. The distance is

$$\delta = \min(\lambda_2, \lambda_1 - \lambda_2) = \min\left(\frac{p-q}{2}, q\right) n =: \mu n.$$

Recalling the bound from Theorem 14 on $R = T - S$ and applying it here, we can bound the distance between the unit eigenvectors of D and A. It follows that there exists a sign $\theta \in \{-1, 1\}$ such that

$$\|v_2(D) - \theta v_2(A)\|_2 \leq \frac{C\sqrt{n}}{\mu n} = \frac{C}{\mu\sqrt{n}}$$

with probability at least $1 - 4e^{-n}$. We already computed the eigenvectors $u_i(D)$ of D, but there they had norm $\sqrt{n}$ (every entry is either 1 or -1). So, multiplying both sides by $\sqrt{n}$, we obtain in this normalization that

$$\|u_2(D) - \theta u_2(A)\|_2 \leq \frac{C}{\mu}.$$

It follows that the *signs* of most coefficients of $\theta v_2(A)$ and $v_2(D)$ must agree. Indeed, we know that

$$\sum_{j=1}^{n} |u_2(D)_j - \theta u_2(A)_j|^2 \leq \frac{C}{\mu^2}. \tag{14}$$

Since the coefficients $u_2(D)_j$ are all ± 1, every coefficient j on which the signs of $\theta v_2(A)_j$ and $v_2(D)_j$ disagree contributes at least 1 to the sum in Equation 14. Thus the number of disagreeing signs must be bounded by

$$\frac{C}{\mu^2}.$$

Summarizing, we can use the vector $v_2(A)$ to accurately estimate the vector $v_2 = v_2(D)$, whose signs identify the two communities; this guarantee is quite strong as it means that with high-probability we will only ever make a finite number of mistakes even if the graph were to grow. This method for community detection is usually called *spectral clustering* as we are working with the spectral decomposition of the adjacency matrix. Finally we note that the having then $\theta \in \{-1, 1\}$ factor doesn't affect the accuracy of the clustering process as it is simply a change in label; you may have seen this in Gaussian mixture models as well.

4.3 Bernsetin bounds for matrices

In this subsection we are trying to show a concentration bound on the sum of independent matrices instead on a single matrix, with the end goal of providing bounds on covariance estimation under general assumptions. Recall that we were able to use the fact that

$$E[\exp(t(X + Y))] = E[\exp(tX)\exp(tY)],$$

and use Chernoff's method to obtain tail bounds if X and Y are real valued random variables. To generalize this approach we need to define what a *matrix exponential* means and furthermore is it true that for matrices A and B:

$$\exp(A)\exp(B) = \exp(A + B)?$$

(it's not!). First let us defined a function of a symmetric matrix:

Definition 10. *For a function $f : \mathbb{R} \to \mathbb{R}$ and an $n \times n$ symmetric matrix*

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^\top,$$

define

$$f(X) := \sum_{i=1}^{n} f(\lambda_i) u_i u_i^\top.$$

Meaning that we apply the function to the eigenvalues while keeping the eigenvectors constant; remember that symmetric matrices always have an eigen-decomposition with real eigenvalues. For a convergent power series expansion of f about x_0:

$$f(x) = \sum_{k=1}^{\infty} a_k (x - x_0)^k.$$

It is the case that series of matrix terms converges, and

$$f(X) = \sum_{k=1}^{\infty} a_k (X - x_0 I)^k.$$

As an example, for each $n \times n$ symmetric matrix X we have

$$e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots$$

A useful fact that we will use about matrix exponential is that they have positive eigenvalues. For example for a symmetric matrix A with positive eigenvalues it is true that:

$$\lambda_{\max}(A) \leq \mathrm{Tr}[A] \leq n \times \max \lambda_{\max}(A)$$

as the trace is the sum of the eigenvalues of a matrix. Also recall that we can generalize inequality on matrices via a partial ordering:

Proposition 12. *(Positive semidefinite order). We say*

$$0 \preceq X,$$

if X is a symmetric positive semidefinite matrix. And we say

$$X \preceq Y$$

if $X - Y \preceq 0$.

A useful fact we will use in the sequel is:

Proposition 13. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be two functions. If $f(x) \leq g(x)$ for all $|x| \leq K$, then $f(X) \preceq g(X)$ for $|A| \leq K$.*

Proof. If $|A| \leq K$ then all of its eigenvalues $|\lambda_i(A)| \leq K$, therefore:

$$\begin{aligned} g(X) - f(X) &= \sum_{i=1}^{n} g(\lambda_i) u_i u_i^\top - \sum_{i=1}^{n} f(\lambda_i) u_i u_i^\top \\ &= \sum_{i=1}^{n} \{g(\lambda_i) - f(\lambda_i)\} u_i u_i^\top \succeq 0 \end{aligned}$$

as $g(\lambda_i) - f(\lambda_i) > 0$ for all $|\lambda_i| \leq K$. □

Let us consider two generalization of the property $\exp(x + y) = \exp(x) \exp(y)$ for matrices.

Theorem 16. *(Golden-Thompson inequality). For any $n \times n$ symmetric matrices A and B, we have*

$$tr(e^{A+B}) \leq tr(e^A e^B).$$

Unfortunately, Golden-Thompson inequality does not hold for three or more matrices: in general, the inequality $\mathrm{tr}(e^{A+B+C}) \leq \mathrm{tr}(e^A e^B e^C)$ may fail.

Theorem 17. *(Lieb's inequality). Let H be an $n \times n$ symmetric matrix. Define the function on matrices*

$$f(X) := tr\exp(H + \log X).$$

Then f is concave on the space on positive definite $n \times n$ symmetric matrices.

A proof of matrix Bernstein's inequality can be based on either Golden-Thompson or Lieb's inequalities. We use Lieb's inequality, which we will now restate for random matrices. If X is a random matrix, then Lieb's and Jensen's inequalities imply that

$$\mathbb{E}f(X) \preceq f(\mathbb{E}X),$$

for the function defined in Lieb's inequality as a matrix function is applied to its eigenvalues (recall that $\preceq$ denotes the partial ordering on symmetric matrices). Applying this with $X = e^Z$, we obtain the following.

Lemma 8. *(Lieb's inequality for random matrices). Let H be a fixed $n \times n$ symmetric matrix and Z be a random $n \times n$ symmetric matrix. Then*

$$\mathbb{E}tr\exp(H + Z) \leq tr\exp(H + \log \mathbb{E}\exp(Z)).$$

Proof. This follows from Jensen's inequality as well as the fact that $X = \exp(Z)$ has positive eigenvalues. □

Theorem 18. *(Matrix Bernstein's inequality). Let $X_1, \ldots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\| \geq t\right\} \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Here $\sigma^2 = \left\|\sum_{i=1}^N \mathbb{E}X_i^2\right\|$ is the norm of the matrix variance of the sum.

In particular, we can express this bound as the mixture of sub-gaussian and sub-exponential tails:

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\| \geq t\right\} \leq 2n \exp\left[-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right].$$

Proof. **Step 1: Reduction to MGF.** To bound the norm of the sum

$$S := \sum_{i=1}^N X_i,$$

we need to control the largest and smallest eigenvalues of S. We can do this separately. To put this formally, consider the largest eigenvalue

$$\lambda_{\max}(S) := \max_i \lambda_i(S)$$

and note that

$$\|S\| = \max_i |\lambda_i(S)| = \max\big(\lambda_{\max}(S), \lambda_{\max}(-S)\big).$$

We only bound $\lambda_{\max}(S)$, the proof for $\lambda_{\max}(-S)$ is similar. To bound $\lambda_{\max}(S)$, we proceed with the Chernoff approach as we did in the scalar case for sub-Gaussian and sub-exponential concentration. Fix $\lambda \geq 0$ and use Markov's inequality to obtain

$$\mathbb{P}\left\{\lambda_{\max}(S) \geq t\right\} = \mathbb{P}\left\{e^{\lambda \cdot \lambda_{\max}(S)} \geq e^{\lambda t}\right\} \leq e^{-\lambda t}\mathbb{E}e^{\lambda \cdot \lambda_{\max}(S)}.$$

By definition of a function on a matrix the eigenvalues of e^S are $e^{\lambda \cdot \lambda_i(S)}$, we have

$$E := \mathbb{E}e^{\lambda \cdot \lambda_{\max}(S)} = \mathbb{E}\lambda_{\max}(e^S).$$

Since the eigenvalues of e^S are all positive, the maximal eigenvalue of e^S is bounded by the sum of all eigenvalues, the trace of e^S, which leads to

$$E \leq \mathbb{E}\operatorname{tr} e^S.$$

Step 2: Application of Lieb's inequality. To prepare for an application of Lieb's inequality (Lemma 8), let us separate the last term from the sum S:

$$E \leq \mathbb{E}\operatorname{tr}\exp\left[\sum_{i=1}^{N-1}\lambda X_i + \lambda X_N\right] = \mathbb{E}\left\{\mathbb{E}\operatorname{tr}\exp\left[\sum_{i=1}^{N-1}\lambda X_i + \lambda X_N\Big| X_1, \ldots, X_{n-1}\right]\right\},$$

by the law of total expectation. Conditional on $(X_i)_{i=1}^{N-1}$, we apply Lemma 8 for the fixed matrix $H := \sum_{i=1}^{N-1}\lambda X_i$ and the random matrix $Z := \lambda X_N$ and obtain

$$E \leq \mathbb{E}\operatorname{tr}\exp\left[\sum_{i=1}^{N-1}\lambda X_i + \log \mathbb{E}e^{\lambda X_N}\right].$$

We continue in a similar way: separate the next term λX_{N-1} from the sum $\sum_{i=1}^{N-1}\lambda X_i$ and apply Lemma 8 again for $Z = \lambda X_{N-1}$. Repeating N times, we obtain

$$E \leq \operatorname{tr}\exp\left[\sum_{i=1}^{N}\log \mathbb{E}e^{\lambda X_i}\right].$$

Step 3: MGF of the individual terms. It remains to bound the matrix-valued moment generating function $\mathbb{E}e^{\lambda X_i}$ for each term X_i. This is a standard task, and the argument will be similar to the scalar case.

Lemma 9. *(Moment generating function). Let X be an $n \times n$ symmetric mean zero random matrix such that $\|X\| \leq K$ almost surely. Then*

$$\mathbb{E}\exp(\lambda X) \preceq \exp\left(g(\lambda)\mathbb{E}X^2\right) \quad \textit{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3},$$

provided that $|\lambda| < 3/K$.

Proof. First, note that we can bound the (scalar) exponential function by the first few terms of its Taylor's expansion as follows:

$$e^z \leq 1 + z + \frac{1}{1 - |z|/3}\cdot\frac{z^2}{2}, \quad \text{if } |z| < 3.$$

(To get this inequality, write $e^z = 1 + z + z^2 \cdot \sum_{p=2}^{\infty} z^{p-2}/p!$ and use the bound $p! \geq 2 \cdot 3^{p-2}$.) Next, apply this inequality for $z = \lambda x$. If $|x| \leq K$ and $|\lambda| < 3/K$ then we obtain

$$e^{\lambda x} \leq 1 + \lambda x + g(\lambda) x^2,$$

where $g(\lambda)$ is the function in the statement of the lemma.

Finally, we can transfer this inequality from scalars to matrices using Lemma [X]. We obtain that if $\|X\| \leq K$ and $|\lambda| < 3/K$, then

$$e^{\lambda X} \preceq I + \lambda X + g(\lambda) X^2.$$

Take expectation of both sides and use the assumption that $\mathbb{E}X = 0$ to obtain

$$\mathbb{E} e^{\lambda X} \preceq I + g(\lambda) \mathbb{E} X^2.$$

To bound the right hand side, we may use the inequality $1 + z \leq e^z$ which holds for all scalars z. Thus the inequality $I + Z \preceq e^Z$ holds for all matrices Z, and in particular for $Z = g(\lambda)\mathbb{E}X^2$. This yields the conclusion of the lemma.

Step 4: Completion of the proof. Let us return to bounding E. Using the bound of the MGFs, we obtain

$$E \leq \operatorname{tr} \exp \left[\sum_{i=1}^{N} \log \mathbb{E} e^{\lambda X_i} \right] \leq \operatorname{tr} \exp \left[g(\lambda) Z \right], \quad \text{where} \quad Z := \sum_{i=1}^{N} \mathbb{E} X_i^2.$$

Since the trace of $\exp[g(\lambda)Z]$ is a sum of n positive eigenvalues, it is bounded by n times the maximum eigenvalue, so

$$E \leq n \cdot \lambda_{\max} \left(\exp \left[g(\lambda) Z \right] \right) = n \cdot \exp \left[g(\lambda) \lambda_{\max}(Z) \right]$$

$$= n \cdot \exp \left[g(\lambda) \|Z\| \right] \quad (\text{since } Z \succeq 0)$$

$$= n \cdot \exp \left[g(\lambda) \sigma^2 \right] \quad (\text{by definition of } \sigma) \, .$$

Plugging this bound for $E = \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)}$

$$\mathbb{P} \left\{ \lambda_{\max}(S) \geq t \right\} \leq n \cdot \exp \left[-\lambda t + g(\lambda) \sigma^2 \right].$$

We can choose the following value: $\lambda = t/(\sigma^2 + Kt/3)$ to make the expression simper instead of direct minimization (which you can do and then realize it's not worth doing like I did). Substituting it into the bound above and simplifying yields

$$\mathbb{P} \left\{ \lambda_{\max}(S) \geq t \right\} \leq n \cdot \exp \left(-\frac{t^2/2}{\sigma^2 + Kt/3} \right).$$

Repeating the argument for $-S$ and combining the two bounds gives the desired conclusion.

□

You can integrated the tail probability to obtain the following bound on the expectation of the operator norm as well:

Lemma 10. *(Matrix Bernstein's inequality: expectation).: Let $X_1, \dots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \le K$ almost surely for all i. Then*

$$\mathbb{E}\left\|\sum_{i=1}^N X_i\right\| \lesssim \left\|\sum_{i=1}^N \mathbb{E}\, X_i^2\right\|^{1/2} \sqrt{1+\log n} + K(1+\log n).$$

We wil use this to bound the estimation error of covariance matrices in the subsection to follow.

4.4 Application: covariance estimation for general distributions

In this section we will derive bounds for estimating covariance matrices of random vectors which have bounded L^2 norm. We estimate the second moment matrix $\Sigma = \mathbb{E}XX^T$ by its sample version

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^m X_i X_i^T.$$

If X has zero mean, then Σ is the covariance matrix of X and Σ_m is the sample covariance matrix of X.

Theorem 19. *(General covariance estimation). Let X be a random vector in $\mathbb{R}^n$, $n \ge 2$. Assume that for some $K \ge 1$,*

$$\|X\|_2 \le K(\mathbb{E}\|X\|_2)^{1/2} \quad \textit{almost surely.} \tag{15}$$

Then, for every positive integer m, we have

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le C\left(\sqrt{\frac{K^2 n \log n}{m}} + \frac{K^2 n \log n}{m}\right)\|\Sigma\|.$$

Proof. Before we start proving the bound, let us pause to note that $\mathbb{E}\|X\|_2^2 = \operatorname{tr}(\Sigma)$. So the assumption (15) becomes

$$\|X\|_2^2 \le K^2 \operatorname{tr}(\Sigma) \quad \text{almost surely.} \tag{16}$$

Lemma 10 for the sum of i.i.d. mean zero random matrices $X_i X_i^T - \Sigma$ to show

$$\mathbb{E}\|\Sigma_m - \Sigma\| = \frac{1}{m}\mathbb{E}\left\|\sum_{i=1}^m (X_i X_i^T - \Sigma)\right\| \lesssim \frac{1}{m}\left(\sigma\sqrt{\log n} + M\log n\right) \tag{17}$$

where

$$\sigma^2 = \left\|\sum_{i=1}^m \mathbb{E}(X_i X_i^\top - \Sigma)^2\right\| = m\left\|\mathbb{E}(XX^\top - \Sigma)^2\right\|$$

and M is any number chosen so that

$$\|XX^\top - \Sigma\| \le M \quad \text{almost surely.}$$

To complete the proof, it remains to bound σ^2 and M. Let us start with σ^2. Expanding the square, we find that

$$\mathbb{E}(XX^\top - \Sigma)^2 = \mathbb{E}(XX^\top)^2 - \Sigma^2 \preceq \mathbb{E}(XX^\top)^2.$$

Further, the assumption (16) gives

$$(XX^\top)^2 = \|X\|^2 XX^\top \preceq K^2 \operatorname{tr}(\Sigma) XX^\top.$$

Taking expectation and recalling that $\mathbb{E}XX^\top = \Sigma$, we obtain

$$\mathbb{E}(XX^\top)^2 \preceq K^2 \operatorname{tr}(\Sigma)\Sigma.$$

Substituting this bound into the expression for σ, we obtain the folllowing

$$\sigma^2 \leq K^2 m \operatorname{tr}(\Sigma)\|\Sigma\|.$$

Now we bound M,

$$\begin{aligned} \|XX^\top - \Sigma\| &\leq \|X\|_2^2 + \|\Sigma\| \quad \text{(by triangle inequality)} \\ &\leq K^2 \operatorname{tr}(\Sigma) + \|\Sigma\| \quad \text{(by assumption (16))} \\ \leq 2K^2 \operatorname{tr}(\Sigma) =: M \quad &\text{(since } \|\Sigma\| \leq \operatorname{tr}(\Sigma) \text{ and } K \geq 1). \end{aligned}$$

Substituting our bounds for σ and M into (17), we get

$$\mathbb{E}\|\Sigma_m - \Sigma\| \leq \frac{1}{m}\left(\sqrt{K^2 m \operatorname{tr}(\Sigma)\|\Sigma\|} \cdot \sqrt{\log n + 2K^2 \operatorname{tr}(\Sigma)} \cdot \log n\right).$$

To complete the proof, use the inequality $\operatorname{tr}(\Sigma) \leq n\|\Sigma\|$ and simplify the bound. □

To use this bound for unbounded vectors, you can remove the boundedness assumption by a truncation argument and then bounding the bias this induces on the estimate, where you can trade off the bias and the concentration gain for the optimal tresholding (or just try to find another bound that works for your problem!).

References

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford.

He, Y., B. Meng, Z. Zeng, and G. Xu (2021). On the phase transition of wilks' phenomenon. *Biometrika 108*(3), 741–748.

Rigollet, P. and J.-C. Hütter (2023). High-dimensional statistics. *arXiv preprint arXiv:2310.19244*.

Saumard, A. and J. A. Wellner (2014). Log-concavity and strong log-concavity: a review. *Statistics surveys 8*, 45–114.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, Cambridge.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge.