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!).

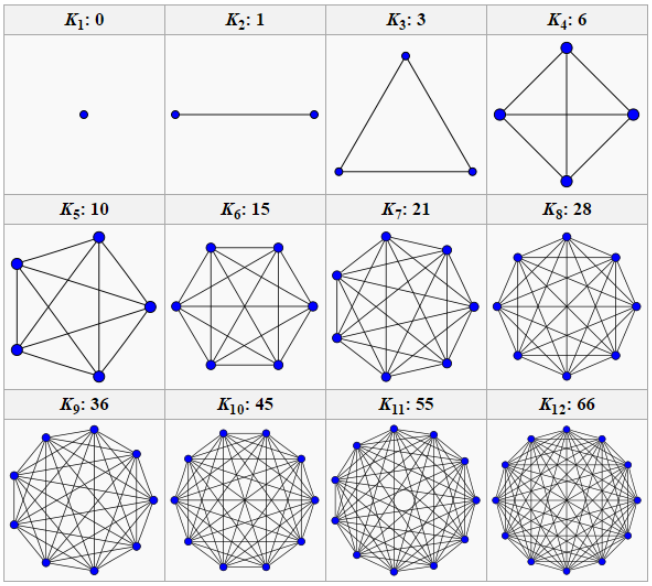

Figure 1: Illustration of what different clique sizes looks like, clique sizes are used as a descriptive statistic for graphs. The title of each sub-plot: $K_i : j$ indicate the number of nodes (i) and the number of total edges (j). You can think of them as groups of friends or perhaps more realistically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

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

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,

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

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

Note that 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 the dependence structure of this term 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 let us try to reduce the dimensionality of this 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 yields

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

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, but for the bound in probability we will need the the last step in the proof of Theorem 3 that

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

with probability $1-\delta$, □

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

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

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

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

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

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

where $K-K = \{x-y : x,y\in K\}$. It is easy to see that if K is symmetric 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 $v^{(1)},\dots,v^{(N)} \in \mathbb{R}^d$ and let $X\in\mathbb{R}^d$ be a random vector such that $[v^{(i)}]^\top X, i=1,\dots,N$, are sub-Gaussian random variables with variance proxy σ^2. Then*

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

and

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

Moreover, for any $t>0$,

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

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 $\mathcal{V} = \{e_1, -e_1, \ldots, e_d, -e_d\}$, 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 $\mathbb{X}\mathcal{V} = \{\mathbb{X}_1, -\mathbb{X}_1, \ldots, \mathbb{X}_d, -\mathbb{X}_d\}$ 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 proof of Theorem 6, we 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*

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

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, effectively 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:

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

for some penalty τ. In practice τ 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)*** *We say that 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*
$$\frac{|\mathbb{X}_i^T \mathbb{X}_j|}{n} \leq \frac{1}{32k}.$$

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*

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

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

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

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

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

Applying Hölder's inequality

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

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

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

In particular, it implies that

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

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 $|S| \leq k$,

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

Combining this result with Equation 4, we find

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

This concludes the proof of the bound on the MSE. To prove the upper bound 3, we use Lemma 2.17 once again to get

$$|\hat{\theta}^{\mathcal{L}} - \theta^*|_2^2 \leq 2\,\text{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) \leq 64k\tau^2.$$

□

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.

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*

$$|\theta|_* = \sum_{j=1}^{d} \lambda_j \theta_j^*, \tag{2.24}$$

or equivalently as

$$|\theta|_* = \max_{\phi \in S_d} \sum_{j=1}^{d} \lambda_j |\theta_{\phi(j)}|. \tag{2.25}$$

The Slope estimator is then given by

$$\hat{\theta}^S \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\theta\|_2^2 + 2\tau |\theta|_* \right\} \tag{2.26}$$

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

Under construction.

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 (of course it turns out that it does both). 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.

4 Covariance and Matrix Estimation

5 Graphical Models (maybe)

6 Generalized Linear Models (maybe)

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.