Notes on High Dimension Probability

Subhrajyoty Roy

October 16, 2024

Abstract

This contains the lecture notes of the course on *High Dimensional Probability* by Roman Vershynin. The course is available for free at online https://www.math.uci.edu/~rvershyn/teaching/hdp/hdp.html.

Contents

1 Introduction to High Dimensional Ideas

Big data can come in one of two different ways.

1. # observations is big, this is usually easy, as classical statistical theory tells us how to deal with large number of samples. These are often better.

2. # dimensions is big. This is usually hard.

Empirical observation: it is exponentially harder to deal with larger # of dimensions rather than larger # of observations. To illustrate this, let's consider an example problem.

Example 1. *Let's say we want to numerically compute the integral*

$$\int_0^1 \cdots \int_0^1 f(x_1, \ldots, x_d) dx_1 \ldots dx_d$$

The usual way is to perform a numerical integration approach, based on Riemann sums. For $d = 1$, we can subdivide the interval $[0, 1]$ into grids of width (or resolution) ϵ, so there are $1/\epsilon$-grids. Then, we have the Riemann sum as

$$\int_{[0,1]} f(x)dx \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i), \ n = (1/\epsilon).$$

Note that, this will ensure that the bias is of order $O(\epsilon)$, assuming f is bounded. To see this,

$$\begin{aligned} \left| \int_{[0,1]} f(x)dx - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right| &\leq \frac{1}{n} \sum_{i=1}^{n} \left| \int_{[(i-1)/n, i/n]} f(x)dx - f(x_i) \right| \\ &= \frac{1}{n} \sum_{i=1}^{n} |f(t_i) - f(x_i)| \\ &= O(1/n) = O(\epsilon) \end{aligned}$$

where t_i is a tag obtained by using Integral Mean Value Theorem on f, for the interval $[(i-1)/n, i/n]$.

In general, for d-dimensional hypercube $[0, 1]^d$, we require $n = O(1/\epsilon^d)$ many points to achieve an error bound of $O(\epsilon)$.

Therefore, we need the number of points to be exponential in dimension to achieve the same level of accuracy. This is also called **the curse of dimensionality**.

However, there is a better way to solve this problem, by using Monte Carlo method, which uses probabily to achieve good result for this. Instead of choosing the points on the grid, choose uniformly at random. Pick N points:

$$S_N = \frac{1}{N} \sum_{i=1}^{N} f(x_i), \ x_i \sim \text{Uniform}([0, 1]^d)$$

Note that, $\mathbb{E}(S_N) = \mathbb{E}(f(X)) = \int_{[0,1]^d} f(x)dx$. The average L^2 error is given by

$$\begin{aligned} \mathbb{E}\left[\left(\frac{1}{N} \sum_{i=1}^{N} f(x_i) - \int_{[0,1]^d} f(x)dx \right)^2 \right] &= \text{Var}\left(\frac{1}{N} \sum_{i=1}^{N} f(x_i) \right) \\ &= \frac{\text{Var}(f(X))}{N} \leq C/N \end{aligned}$$

since $f(x)$ is bounded. Therefore, the RMSE $= O(\frac{1}{\sqrt{N}})$, independent of the dimension d.

Note

Thinking in terms of probability might help to overcome in high-dimensional inference problems.

1.1 Convexity

Usually in high-dimensional (HD) problems, convexity helps a lot.

Definition 1. *A set $T \subset \mathbb{R}^n$ is convex if $\forall x, y \in T$, the segment $[x, y] \in T$.*

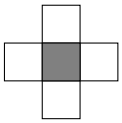
Let us consider a few examples as follows:


convex


convex


non convex


non convex (union of convex sets)


convex (intersection of convex sets)

This means union of convex sets may not be convex, but intersection of convex sets is convex. Thus, starting with a set T, we can consider the intersection of all such convex sets that are superset of T.

Definition 2. *The convex hull conv(T) of a set $T \subset \mathbb{R}^n$ is the smallest convex set that contains T.*

Theorem 1. *$\forall z \in$ conv(T), we have a decomposition, $z = \sum_{i=1}^{m} \lambda_i z_i$, where, $\lambda_i \geq 0$, $\sum_{i=1}^{m} \lambda_i = 1$, and each $z_i \in T$. Note that, the above combination or representation may not be unique or parsimonious.*

Theorem 2 (Caratheodory Theorem)**.** *$\forall z \in$ conv(T), $\exists$ a representation (convex combination) of $\leq (n+1)$ points in T, where, $T \subset \mathbb{R}^n$.*

Note that, the choice of basis in the representation obtained by the Caratheodory Theorem may depend on the choice of z. Also, the number $(n+1)$ is unimprovable, there is a dimension dependence here. However, it turns out that if we allow to get z back only approximately, we can use a similar trick like the Monte Carlo method to get a better representation, that may be dimension-independent. Additionally it says that we can choose the weights to be the same, and still get a good approximate solution.

Theorem 3 (Approximate Caratheodory Theorem)**.** *Let $T \subset \mathbb{R}^n$, and diam$(T) \leq 1$ (otherwise we can rescale). Then, $\forall z \in conv(T)$, $\forall k \in \mathbb{N}$, $\exists z_1, z_2, \ldots, z_k \in T$ (these points may be same) such that*

$$\left\| z - \frac{1}{k} \sum_{i=1}^{k} z_i \right\|_2 \leq \frac{1}{\sqrt{2k}}$$

This means, if we want error to be less than equal to ϵ, the choose, $k = \frac{1}{2\epsilon^2}$ (which is dimension free).

Proof. **This proof is also called Empirical Method of Maurey**.

Fix any $z \in \text{conv}(T)$, by Caratheory Theorem, we have, $z = \sum_{i=1}^{m} \lambda_i z_i$, $\lambda_i \geq 0$, $\sum_{i=1}^{m} \lambda_i = 1$, $z_i \in T$.

Now, consider a random variable (r.v.) Z that takes value z_i, with probability λ_i. Then, $\mathbb{E}(Z) = \sum_{i=1}^{m} \lambda_i z_i = x$. Let, $X_1, X_2, \ldots X_k \equiv Z$ be iid copies of Z. Therefore, the L^2 error is

$$\begin{aligned}
\mathbb{E}\left(\left\| z - \frac{1}{k} \sum_{i=1}^{k} X_i \right\|_2^2 \right) &= \mathbb{E} \left\| \frac{1}{k} \sum_{i=1}^{k} (X_i - z) \right\|_2^2 \\
&= \frac{1}{k^2} \sum_{i=1}^{k} \mathbb{E} \left\| X_i - \mathbb{E}(X_i) \right\|_2^2, \text{ since } z = \mathbb{E}(Z) = \mathbb{E}(X_i) \\
&= \frac{1}{k} \mathbb{E} \left\| Z - \mathbb{E}(Z) \right\|_2^2, \text{ since } X_i \equiv Z \\
&= \frac{1}{2k} \mathbb{E} \left\| Z - Z' \right\|_2^2, \text{ where } Z' \text{ is an independent copy of } Z \\
&\leq \frac{1}{2k}, \text{ as } \left\| Z - Z' \right\|_2^2 \leq \text{diam}(T) \leq 1
\end{aligned}$$

So since, the expectation $\leq \frac{1}{2k} \Rightarrow \exists$ a realization of X_i's such that L^2 error $\leq \frac{1}{2k}$. Therefore, $\left\| z - \frac{1}{k} \sum_{i=1}^{k} z_i \right\|_2^2 \leq \frac{1}{2k}$, where, z_i's are the realizations of X_i, and hence each $z_i \in T$. □

Exercise 1. *Let $x_1, x_2, \ldots x_n$ be an arbitrary set of unit vectors in $\mathbb{R}^n$. Prove that there exists $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ with $\epsilon_i \in \{-1, +1\}$ such that $\| \sum_{i=1}^{n} \epsilon_i x_i \|_2 \leq \sqrt{n}$.*

To solve this, we use probabilistic methods. Note that, if ϵ_is are random variables, then the expected squared L^2 norm is

$$\mathbb{E} \| \sum_{i=1}^{n} \epsilon_i x_i \|_2^2 = \sum_{i=1}^{n} \|x_i\|^2 \mathbb{E}(\epsilon_i^2) = n$$

since x_i is unit vector and $\mathbb{E}(\epsilon_i^2) = \text{Var}(\epsilon_i) + \mathbb{E}^2(\epsilon_i) = 1$. This means, there exists a realization that has squared L^2 norm less than equal to n, i.e., the L^2 norm less than equal to $\sqrt{n}$.

1.2 Applications of ACT

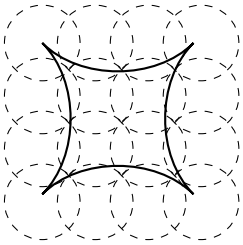
There are multiple applications of ACT. One useful idea is in portfolio building, where the ingredients are "stocks", and their linear combination is basically a mutual fund. The problem is to create a new mutual fund with a given combination of stocks by mixing available mutual funds, (i.e., the fund of funds). The solution is to use ACT to obtain a fast randomized solution approximately.

Note

Note that, ACT only provides the existence, and does not necessarily provide an algorithm to find that out. However, one simple algorithm could be to start with all the available mutual funds and then take any k of them randomly and average their returns.

Another prominent application of ACT is in covering numbers. This is what we will explore now.

Definition 3. *The **Covering Number** of a set $T \subset \mathbb{R}^n$ at scale $\epsilon > 0$ is the smallest # of Euclidean balls of radius ϵ needed to cover T. It is denoted by $N(T, \epsilon)$.*



$N(T, \epsilon) \leq 16$

Lemma 1. *Let B be the unit Euclidean ball, then we have $N(B, \frac{1}{2}) \geq 2^d$, i.e., the covering number is exponential in dimension.*

Proof. Assume B can be covered by N copies of $(\frac{1}{2}B)$ ball. Then clearly,

$$Vol(B) \leq N \cdot Vol\left(\frac{1}{2}B\right), \text{ since RHS might have some overlaps.}$$
$$\implies Vol(B) \leq N 2^{-d} Vol(B)$$
$$\implies N \geq 2^d$$

□

This is an unsurprising result. However, what is surprising is that the covering number for a polytope (i.e., a higher dimensional analogue of polygon), is independent of the dimension but dependent only on the number of vertices. Also, it is polynomial in the number of vertices.

Lemma 2. *Let P be a polytope in $\mathbb{R}^d$ with m vertices and $diam(P) \leq 1$. Then, $N(P, \epsilon) \leq m^{1/2\epsilon^2}$.*

Proof. If we consider, P to be nonconvex, then $conv(P)$ is a polytope with $\leq m$ vertices. This can only weaken the bound. Therefore, it is enough to show it for the cases when P is convex.

Let T be the set of vertices of P. Clearly, $P \subseteq conv(T)$. Note that ACT implies that $\forall z \in P$, is within distance $\frac{1}{\sqrt{2k}}$ from some point in the set

$$\mathcal{N} := \left\{ \frac{1}{k} \sum_{i=1}^{k} z_i : z_i \in T \right\}.$$

This means, $\forall x \in P$, is covered by a ball of radius $\frac{1}{\sqrt{2k}}$ and center $\in \mathcal{N}$. This means, the covering number

$$N(P, 1/\sqrt{2k}) \leq |\mathcal{N}| \leq m^k,$$

since to choose an element of $\mathcal{N}$, we pick any k vertices from m vertices which is $\binom{m}{k} = O(m^k)$. Choosing $\epsilon = 1/\sqrt{2k}$ now completes the proof. □

To make use of these results, usually, it helps by considering the intuition that, a small covering # ⇒ small volume, ⇒ easier to apply union-type bounds. As a result, since the covering number of a polytope is small, we expect the volume of the polytope also be to small compared to the Euclidean ball.

Theorem 4 (Carl-Pajor, 88)**.** *Let B be the unit Euclidean ball, and $P \subset B$ be any polytope with m vertices in $\mathbb{R}^n$. Then,*

$$\frac{Vol(P)}{Vol(B)} \leq \left(4 \frac{\sqrt{\log m}}{\sqrt{n}} \right)^n$$

What the Carl-Pajor theorem tells us is that the volume of the polytope is extremely small, unless the number of vertices m is exponential in n.

Proof. Let us consider ϵB-balls and cover P with these balls. By definition of covering number, we have

$$\begin{aligned} & Vol(P) \leq N(P, \epsilon) Vol(\epsilon B) \\ \implies & Vol(P) \leq N(P, \epsilon) \epsilon^n Vol(B) \\ \implies & Vol(P)/Vol(B) \leq \epsilon N(P, \epsilon) \leq \epsilon^n m^{2/\epsilon^2} \end{aligned}$$

Here, we use the previous Lemma, but with the understanding that since B is a unit Euclidean ball, its diameter is 2. Now, note that the above inequality is true for every $\epsilon > 0$. Therefore,

$$\frac{Vol(P)}{Vol(B)} \leq \inf_{\epsilon > 0} \epsilon^n \cdot m^{\frac{1}{2\epsilon^2}}.$$

Let $\ell(\epsilon) = \epsilon^n m^{\frac{1}{2\epsilon^2}} \Rightarrow \log \ell(\epsilon) = n \log \epsilon + \frac{1}{2\epsilon^2} \log m$. Now setting its derivative equal to 0, we get $\frac{\partial \log l}{\partial \epsilon} = 0 = \frac{n}{\epsilon} - \frac{1}{\epsilon^3} \log m \Rightarrow \epsilon = \sqrt{\frac{4 \log m}{n}}$. Putting this value of ϵ based into the expression, we get

$$\inf_{\epsilon > 0} \ell(\epsilon) = \exp \left[n \log \left(\frac{\sqrt{4 \log m}}{\sqrt{n}} \right) + \frac{2 \log m}{4 \log m} n \right]$$

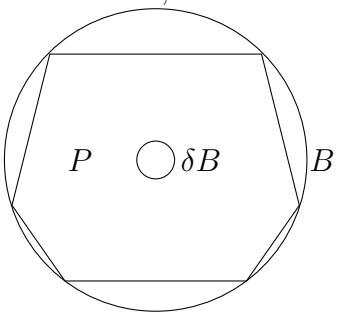
$$= \exp\left[\log\left(\frac{\sqrt{4\log m}}{\sqrt{n}}\right)^n e^{n/2}\right]$$
$$= \left(\frac{\sqrt{4e\log m}}{\sqrt{n}}\right)^n \leq \left(4\frac{\sqrt{\log m}}{\sqrt{n}}\right)^n, \text{ since } e \leq 4.$$

□

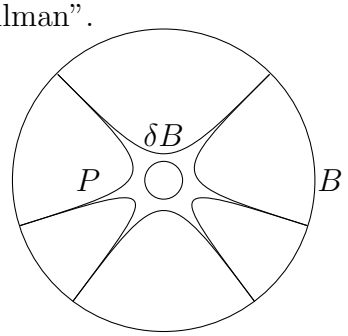
A few remarks we want to make here.

- Carl-Pajor proved a slightly better result, with $\log(m/n)$ rate, which is optimal.
- The optimal bound is attained at a random polytope. (Dafnis et al., 2003, 2009)
- Let $\delta = 4\sqrt{\frac{\log m}{n}}$, then note that, $Vol(\delta B) = \delta^n \cdot Vol(B)$, and correspondingly, $\Rightarrow \frac{Vol(\delta B)}{Vol(B)} = \delta^n \geq \frac{Vol(P)}{Vol(B)}$. As a result, we have $Vol(\delta B) \geq Vol(P)$. This means, a small ball around origin has at least the same volume as the polytope.

This means, the low-dimensional picture that we usually think is wrong.



For correct intuition in high dimension, we can consider the hyperbolic correction by "V. Milman".



The part δB is called the "core" of the polytope. Note that, the polytope is convex, but it does not look convex. This is because, for high dimensional inference, we cannot completely visualize all the information. So here, we drop the convexity (because we can work with it algebraically), and instead work with the volume-based ideas (because probabilities map to volume in terms of measures).

2 Concentration Inequalities

We know by law of large numbers that i.i.d. sum of a random variable $\boldsymbol{X}$ converges to $\mathbb{E}(X)$. The central limit theorem also tells that the rate of this convergence is $1/\sqrt{n}$. Since the central limit theorem and law of large numbers are very universal in nature, one might want to know to what extent these kinds of concentration holds. The fundamental aim of concentration inequalities is to demonstrate that a random variable (or i.i.d. sum of it) satisfies $X \approx \mathbb{E}(X)$ with high probability (exponentially close to 1).

For example, in case of normal distribution, $X \sim N(\mu, \sigma^2)$, we know that

$$\mathbb{P}(|X - \mu| > 3\sigma) = 0.9973.$$

For general distribution, we have

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}, \ \forall t > 0, \ \text{(because of Markov's inequality)}$$
$$\mathbb{P}(|X - \mathbb{E}(X)| > t) \leq \frac{\sigma^2}{t^2}, \ \forall t > 0, \ \text{because of Chebyshev's inequality}$$

However, for Gaussian distribution, the concentration around the mean is much faster. To see this, consider the following Lemma.

Theorem 5. *Suppose $g \sim N(0,1)$, then*

$$\mathbb{P}(g > t) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t},$$

i.e., the tail probability decays exponentially fast in t.

Proof.

$$\begin{aligned}
\mathbb{P}(g > t) &= \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-(t+y)^2/2} dy, \ \text{using substitution } x = y + t \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} e^{-ty} e^{-y^2/2} dy \\
&\leq \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} e^{-ty} dy, \ \text{since } e^{-y^2/2} \leq 1 \\
&= \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty} dy \\
&= \frac{1}{\sqrt{2\pi} t} e^{-t^2/2}
\end{aligned}$$

□

By symmetry, then we have for general $X \sim N(\mu, \sigma^2)$,

$$\mathbb{P}(|X - \mu| > t\sigma) \leq \frac{1}{t} \sqrt{\frac{2}{\pi}} e^{-t^2/2} \leq e^{-t^2/2}, \ \text{when } t \geq 1.$$

Therefore, the normal distribution has exponential tails. Also, the central limit theorem (CLT) tells that

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

so may be the tail probabilities are also close and exponential. But this does not work as the error from the CLT itself is of order $O(1/\sqrt{n})$, which is due to Berry-Esseen bound. So, the tail of $\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$ does exponentially decay in general like Gaussian, at least by this route.

Theorem 6 (Berry Essen Bound)**.** *Let X_i be i.i.d. random variables with mean* 0 *and variance* 1*, and finite third order moment. Then,*

$$\left|\mathbb{P}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i \geq t\right) - \mathbb{P}(Z \geq t)\right| \leq \frac{\mathbb{E}(|X_1|^3}{\sqrt{N}} = O(1/\sqrt{N}),$$

where $Z \sim N(0,1)$.

So we would like to sidestep this CLT and try to control the tail of this normalized sum directly. This introduces the study of concentration inequalities.

Exercise 2. *Let X be a standard normal random vector in $\mathbb{R}^n$ where $n \geq C_1$ (large constant). What are the values of $\mathbb{E}\|X\|_2^2$ and $\mathrm{Var}(\|X\|_2^2)$?*

To obtain this, let $X_1, X_2, \ldots X_n$ be the coordinates of the vector, note that

$$\mathbb{E}\|X\|_2^2 = \mathbb{E}(X_1^2 + X_2^2 + \cdots + X_n^2) = n\mathbb{E}(X_1^2) = n$$
$$\mathrm{Var}(\|X\|_2^2) = \mathrm{Var}(X_1^2 + \cdots + X_n^2) = n\mathrm{Var}(X_1^2) = n\left[\mathbb{E}(X_1^4) - \mathbb{E}^2(X_1^2)\right] = 2n$$

Exercise 3 (Continued)**.** *Show that, $|\|X\|_2^2 - n| \leq C\sqrt{n}$ for some large C with high probability (say* 0.99*).*

To show this, we make use of Chebyshev's inequality. Note that,

$$\mathbb{P}(|\|X\|_2^2 - n| > t) = \mathbb{P}(|\|X\|_2^2 - \mathbb{E}(\|X\|_2^2)| > t) \leq \frac{\mathrm{Var}(\|X\|_2^2)}{t^2} = \frac{2n}{t^2}$$

Choosing $t = C\sqrt{n}$ for large C yields, $\mathbb{P}(|\|X\|_2^2 - n| > C\sqrt{n}) \leq 2/C$, which can be made smaller than 0.01 *by choosing large C.*

Exercise 4 (Continued)**.** *Deduce that this means, $\sqrt{n}/2 \leq \|X\|_2 \leq 2\sqrt{n}$ with high probability say* 0.99*.*

Note that,

$$|\|X\|_2 - \sqrt{n}| = \frac{|\|X\|_2^2 - n|}{|\|X\|_2 + \sqrt{n}|} = \frac{\leq C\sqrt{n}}{\geq \sqrt{n}} \leq C$$

with high probability. This says that for a random normal vector X, $\|X\|_2 = \sqrt{n} + o(1)$.

Exercise 5. *Let X and Y be two independent standard normal random vectors in $\mathbb{R}^n$ for some $n \geq C$. Find $\mathbb{E}\langle X, Y\rangle^2$ and $\mathrm{Var}\langle X, Y\rangle^2$.*

Note that,

$$\begin{aligned}\mathbb{E}\langle X, Y\rangle^2 &= \mathbb{E}\left((\sum_{i=1}^n X_i Y_i)^2\right) \\ &= \mathbb{E}\left[\sum_{i=1}^n X_i^2 Y_i^2 + \sum_{i\neq j} X_i X_j Y_i Y_j\right] = n\end{aligned}$$

since $\mathbb{E}(X_i^2 Y_i^2) = \mathbb{E}(X_i^2)\mathbb{E}(Y_i^2)$, by independence, and the second term is zero.

Similarly,

$$\begin{aligned}\mathbb{E}\langle X, Y\rangle^4 &= \mathbb{E}\left[\left(\sum_{i=1}^n X_i Y_i\right)^4\right] \\ &= \mathbb{E}\left[\sum_{i=1}^n X_i^4 Y_i^4 + \sum_{i\neq j} X_i^2 Y_i^2 X_j^2 Y_j^2\right], \text{ other terms are zero} \\ &= n \times 3 \times 3 + n(n-1) \times 1 = n^2 + 8n\end{aligned}$$

Therefore, $\mathrm{Var}\langle X, Y\rangle^2 = \mathbb{E}\langle X, Y\rangle^4 - \mathbb{E}^2\langle X, Y\rangle^2 = n^2 + 8n - n^2 = 8n$.

Exercise 6 ((Continued))**.** *Show that if the angle between those vectors is denoted by θ, then $|\theta - \pi/2| = O(1/\sqrt{n})$ with large probability.*

We start by applying Chebyshev's inequality on $\langle X, Y\rangle^2$. Note that,

$$\mathbb{P}\left(|\langle X, Y\rangle^2 - n| \geq 3n\right) \leq \frac{\mathrm{Var}\langle X, Y\rangle^2}{9n^2} = \frac{8}{9n} \to 0$$

Therefore, with sufficiently large probability, $\langle X, Y\rangle^2 \leq 4n$. This means,

$$\cos^2(\theta) = \frac{\langle X, Y\rangle^2}{\|X\|_2^2 \|Y\|_2^2} = \frac{\leq 4n}{\geq (\sqrt{n} + o(1))^4} \leq \frac{C_2}{n}$$

i.e., $|\cos(\theta)| = O(1/\sqrt{n})$. Here we apply the result from previous exercise that $\|X\|_2^2 = (\sqrt{n} + o(1))$. Therefore, $|\theta - \pi/2| = O(1/\sqrt{n}) \to 0$ as $n \to \infty$.

What this exercise shows is that in high dimensions, almost any pair of unit random vectors are orthogonal to each other.

2.1 Inequalities

Definition 4. *A symmetric Bernoulli random variable is X which takes values $+1$ and (-1) with equal probabilities, i.e., $1/2$.*

Theorem 7 (Hoeffding's inequality)**.** *Let $X_1, X_2, \dots X_N$ be symmetric Bernoulli random variables, then*

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \geq t\right) \leq e^{-t^2/2}, \quad \forall t \geq 0,$$

i.e., the normalized sum exhibits Gaussian tail behaviour.

Proof. **This proof is also known as the MGF method.**

Let $\lambda > 0$ be a parameter. Then

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \geq t\right) &= \mathbb{P}\left(e^{\lambda \sum_{i=1}^{n} X_i} \geq e^{\lambda t\sqrt{n}}\right) \\
&\leq e^{-\lambda t\sqrt{n}}\mathbb{E}\left(e^{\lambda \sum_{i=1}^{n} X_i}\right), \text{ by Markov's inequality} \\
&= e^{-\lambda t\sqrt{n}}\prod_{i=1}^{n}\mathbb{E}\left(e^{\lambda X_i}\right), \text{ since } X_i\text{s are i.i.d.} \\
&= e^{-\lambda t\sqrt{n}}\left(\frac{e^{\lambda}+e^{-\lambda}}{2}\right)^n \\
&\leq e^{-\lambda t\sqrt{n}}e^{n\lambda^2/2}, \text{ since } \cosh(\lambda) = (e^{\lambda}+e^{-\lambda})/2 \leq e^{\lambda^2/2} \\
&= \exp\left(-\lambda t\sqrt{n} + n\lambda^2/2\right)
\end{aligned}$$

Now, we optimize this final bound over the choice of $\lambda > 0$ to complete the proof. □

The general version of Hoeffding's inequality can work with any bounded random variables. For example,

Theorem 8. *Let $X_1, X_2, \dots, X_n$ be i.i.d. r.v. such that $X_i \in [a_i, b_i]$ for all i. Then, $S_n = \sum_{i=1}^{n} X_i$ satisfies*

$$\mathbb{P}\left(S_n - \mathbb{E}(S_n) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Unfortunately, there is one problem with Hoeffding's inequality. It works only for bounded random variables, and does not take into account the variance component. Therefore, it will yield the same bound for a uniform random variable on $[a, b]$ and the random variable that takes probability $1/2$ on both the endpoints of $[a, b]$. But we expect the former to have a rapid decay.

Let us consider an empirical approximation. Let $X_1, X_2, \dots X_N \sim \text{Ber}(p)$, such that $p \to 0$ and $pN \to \mu$. Then,

$$\mathbb{P}\left(S_n = \sum_{i=1}^{n} X_i \geq t\right) \to \text{Poisson}(\mu)$$

Consider Poisson tails,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) = e^{-\mu}\sum_{k \geq t}\frac{\mu^k}{k!} \quad \left(\text{Stirling's bounds: } k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k\right)$$

$$\sim e^{-\mu}\mu^t\left(\frac{e}{t}\right)^t \quad \text{(only dominating term is } t\text{)}$$
$$= e^{-\mu}\left(\frac{\mu e}{t}\right)^t = O((C/t)^{-t})$$

Hence, we expect a tail that is like t^{-t}, or $e^{-t\log(t)}$, instead of the lighter Gaussian tail $e^{-t^2/2}$. This is illustrated through Chernoff's inequality.

Theorem 9 (Chernoff's inequality). *Let $X_i \sim Ber(p_i)$, and $S_N = \sum_{i=1}^N X_i$ has mean $\mathbb{E}(S_N) = \sum_{i=1}^N p_i = \mu$, and satisfies,*

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t, \ \forall t \geq \mu$$

Proof. Using the MGF method, we have $\mathbb{P}\left(S_n \geq t\right) \leq e^{-\lambda t}\prod_{i=1}^N \mathbb{E}\left(e^{\lambda X_i}\right)$. Now note that,

$$\mathbb{E}\left(e^{\lambda X_i}\right) = e^{\lambda}p_i + (1-p_i) \leq 1 + \left(e^{\lambda}-1\right)p_i \leq \exp\left(\left(e^{\lambda}-1\right)p_i\right) \quad \text{(as } 1+x \leq e^x\text{)}$$

Therefore, we have

$$\mathbb{P}(S_N \geq t) \leq e^{-\lambda t}\exp\left[(e^{\lambda}-1)\mu\right] = \exp\left[-\lambda t + (e^{\lambda}-1)\mu\right].$$

Similar to before, we now optimize over the choices of $\lambda > 0$. Differentiating with respect to λ yields, $(-t + e^{\lambda}\mu) = 0$, i.e., $\lambda = \log(t/\mu)$. Since $t \geq \mu$, this optimal value of $\lambda \geq 0$. Putting $\lambda = \log(t/\mu)$ back into the expression, we get

$$\exp\left(-t\log(t/\mu) + (t/\mu - 1)\mu\right) = \exp\left(-\mu + t - t\log(t) + t\log(\mu)\right) = e^{-\mu}\left(\frac{e\mu}{t}\right)^t,$$

as we wanted. □

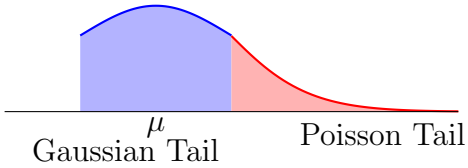
Although the Chernoff bound predicts a heavier tail compared to the Gaussian tail (i.e., $O(e^{-t\log(t)}$ instead of $O(e^{-t^2}))$, it makes sense only when $t >> \mu$. When $t \approx \mu$, i.e., for small deviations, we can get Gaussian approximations.

$$\mathbb{P}(S_n \geq (1+\delta)\mu) \leq e^{-\lambda\mu}\left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu}, \ \text{say } \delta \leq 1$$
$$= \exp\left[\mu\left(\delta - (1+\delta)\log(1+\delta)\right)\right]$$

By Taylor series, we have

$$\log(1+\delta) = \delta - \delta^2/2 + \delta^3/3 + \cdots \geq \delta - \delta^2/2$$
$$\delta - (1+\delta)\log(1+\delta) \leq \delta - (1+\delta)(\delta - \delta^2/2) = \delta - (\delta + \delta^2 - \delta^2/2 - \delta^3/3) \leq -\delta^2/6$$
$$\mathbb{P}\left(S_n \geq (1+\delta)\mu\right) \leq \exp\left(-\delta^2\mu/6\right), \ \forall \delta \leq 1,$$

Note that, this is like a Gaussian tail. Therefore, what Chernoff bound shows is that near the center, this sum of Bernoulli's behave like Gaussian (so CLT and other approximations work well) but in the tail region, it is fatter than the Gaussian.



2.2 Applications

2.2.1 Mean estimation

We start with an application of Hoeffding's inequality. Consider an i.i.d. sample $X_1, X_2, \ldots, X_n$ from (μ, σ^2), i.e., $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. We want to estimate these parameters. The classical estimator of the population mean is the sample mean, i.e., $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and it turns out that $\mathbb{E}(\hat{\mu}) = \mu$ and hence it is unbiased, and also $\text{Var}(\hat{\mu}) = \sigma^2/n$. One can show that this is UMVUE, hence this rate of variance is optimal.

The usual confidence intervals in this case looks like $(\widehat{\mu} - t\sigma/\sqrt{n}, \widehat{\mu} + t\sigma/\sqrt{n})$, for some t. However, this may not be very sharp or the best confidence interval. Because, we only get

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t\sigma/\sqrt{n}) \leq \frac{\sigma^2/n}{(t\sigma/\sqrt{n})^2} = \frac{1}{t^2},$$

by usage of Chebyshev's inequality. This is not a very sharp bound.

Can we get sharper exponentially close to 1 confidence for general distributions? Surprisingly YES! (Note that we only assume $\mathbb{E}|X|^2 < \infty$, not higher order moments).

This trick is called **Median of means** estimator. Assume $n = mK$. Partition the sample into K blocks of size m, denoted as $B_1, B_2, \ldots B_K$. Let, $\hat{\mu}_j = \frac{1}{m}\sum_{i \in B_j} X_i$. Define, $\tilde{\mu}$ to be the median of these blockwise means, i.e., median of $\hat{\mu}_1, \ldots \hat{\mu}_K$.

The confidence interval is given in the same way as before, $(\tilde{\mu} - t\sigma/\sqrt{n}, \tilde{\mu} + t\sigma/\sqrt{n})$, for some t. Note that, the error for each $\hat{\mu}_j$ is,

$$\mathbb{P}\left(\hat{\mu}_j \geq \mu + \frac{t\sigma}{\sqrt{n}}\right) \leq \frac{\sigma^2/m}{(t\sigma/\sqrt{n})^2} = \frac{n/m}{t^2} = \frac{K}{t^2}$$

We can choose K as per our convenience. Let us choose $K = t^2/4$. In this case,

$$\mathbb{P}\left(\hat{\mu}_j \geq \mu + \frac{t\sigma}{\sqrt{n}}\right) \leq \frac{1}{4}.$$

Now, by definition of median, we have

$$\mathbb{P}\left(\hat{\mu} > \mu + \frac{t\sigma}{\sqrt{n}}\right) \leq \mathbb{P}\left(\text{at least } \frac{K}{2} \text{ of } \hat{\mu}_j \text{ are } \geq \frac{t\sigma}{\sqrt{n}}\right)$$

Note that, $Y_j = \mathbf{1}(\hat{\mu}_j \geq t\sigma/\sqrt{n}) \sim \text{Ber}(p)$, where $p \leq 1/4$. Let, $S_K = \frac{1}{K}\sum_{i=1}^{K} Y_j$. Then, $\mathbb{E}(S_K) = Kp \leq K/4$ and $\text{Var}(S_K) = Kp(1-p) \leq 3K/16$. Hence, applying Hoeffding's inequality, we get

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu} > \mu + \frac{t\sigma}{\sqrt{n}}\right) &= \mathbb{P}(S_K \geq K/2) \\
&= \mathbb{P}(S_K - \mathbb{E}(S_K) \geq K/4) \\
&\leq e^{-CK}, \text{ by Hoeffding's inequality for some large } C.
\end{aligned}$$

Therefore, putting the value of K back, we get

$$\mathbb{P}\left(|\tilde{\mu} - \mu| > t\sigma/\sqrt{n}\right) \leq e^{-Ct^2/4}$$

Hence, the median of the means estimator provides a much narrower confidence interval for general distributions.

2.2.2 Random Graph Phase Transition

Definition 5. *An Erdös-Rényi random graph $G(n,p)$ is a random graph consisting of n nodes, and the edge between any two nodes exists with probably p independent of the other edges of the graph.*

The degree of a vertex $\deg(i) = d_i = \#\text{edges connected to } i$. Clearly, $d_i \sim \text{Binom}(n-1,p)$. Therefore, $\mathbb{E}(d_i) = (n-1)p := d$. Turns out there is a phase transition. If $d < c\log(n)$, the random graph ends up with a giant component in between with a large probability. On the other hand, when $d > C\log(n)$, then the random graph becomes connected and regular with high probability.

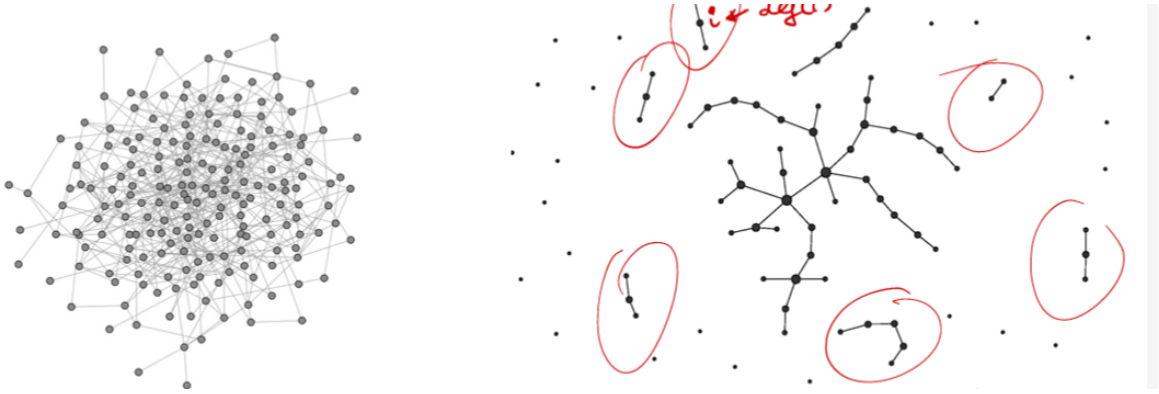


Figure 1: Phase transitions of Erdos-Renyi random graph

Theorem 10. *There exists absolute constant $C > 0$, such that if $d \geq C\log(n)$, then $G(n,p)$ is almost d-regular with high probability. That means,*

$$\mathbb{P}\left(\forall i, deg(i) : 0.9d \leq d_i \leq 1.1d\right) \geq 0.99$$

Proof. Fix any i, let E_i be the event that $|d_i - d| \leq \delta d$. Say $\delta \leq 1$. Then, by Chernoff inequality, we have

$$\mathbb{P}(E_i^c) \leq e^{-\delta^2 d/6} \leq e^{-\delta^2/6 \times C\log(n)} \leq \frac{1}{100n},$$

by choosing sufficiently large C. Now, union bound applies, i.e.,

$$\mathbb{P}\left(\bigcap_{i=1}^{N} E_i\right) \geq 1 - \sum_{i=1}^{n} P(E_i) \geq 1 - n \times \frac{1}{100n} = 0.99.$$

□

Theorem 11 (No regularity, but high degree clique)**.** *$\exists$ absolute constant $C > 0$ s.t. $d < C\log n$, the random graph $G(n,p)$ has a vertex with a large degree with high probability. These are called **hubs**. Mathematically,*

$$\mathbb{P}\left(\exists i, deg(i) = d_i \geq 10d\right) \approx 0.9.$$

Proof. Since $d_i \sim \text{Binom}(n-1, p)$ with $\mathbb{E}(d_i) = d$, by application of reverse Chernoff inequality, we have

$$\begin{aligned}\mathbb{P}(d_i \geq 10d) &\geq e^{-\mu}\left(\mu/t\right)^t, \quad \text{where } t = 10d \\ &= e^{-d}(1/10)^{10d} \geq e^{-20d} \\ &\geq e^{-20Clog(n)} \geq \frac{100}{n}\end{aligned}$$

by choosing sufficiently small C. Let, E_i be the event that $d_i \geq 10d$ for any fixed i.

Now if E_is are independent, then it is possible that

$$\mathbb{P}\left(\cup_{i=1}^n E_i\right) = 1 - \prod_{i=1}^n \mathbb{P}(E_i^c) = 1 - \left(1 - \frac{100}{n}\right)^n \geq 1 - e^{-100} \geq 0.9.$$

But unfortunately, the above proof won't work since E_is are not independent. Because if we have an edge between i and j, then both d_i and d_j increases by one. However, for a fixed i and j, if we ignore the edge between i and j, the other edges of i are independent of the other edges of j.

So, we make some changes to the random graph $G(n,p)$ and create a new graph G'. In this new graph, we create two partitions of vertices, each of size $n/2$ vertices. Then we remove all the edges within the left partition, and within the right partition, but the edges remain between left and right partitions. At this point, if you pick two vertex i and j in the left partition, their modified degrees $\tilde{d}_i$ and $\tilde{d}_j$ are independently distributed. However, $\tilde{d}_i \sim \text{Binom}(n/2-1, p)$. It still holds that

$$\frac{100}{n} \leq \mathbb{P}(E_i) = \mathbb{P}(d_i \geq 10d) = \mathbb{P}(d_i/2 \geq 5d) \leq \mathbb{P}(\tilde{d}_i \geq 5d),$$

Now, we have

$$\mathbb{P}\left(\cup_{i=1}^n E_i\right) \geq \mathbb{P}\left(\cup_{i \in \text{left half}} E_i\right) \geq 1 - \left(1 - \frac{100}{n}\right)^{n/2} \geq 1 - e^{-50} \geq 0.9,$$

which completes the proof. □

2.2.3 Discrepancy Theory

Let's say we throw N random points onto the 2d square $[0,1]^2$, which are i.i.d. uniform points in $[0,1]]^2$. Then, for all subset $I \subset [0,1]^2$, the expected # of points in I, $N_I \sim \text{Binom}(N, \lambda_I)$, where λ_I is the area of I. Clearly,

$$\begin{aligned}\mathbb{E}N_I &= N\lambda_I \\ \text{Var}(N_I) &= N\lambda_I(1-\lambda_I) < N\lambda_I \\ \text{sd}(N_I) &\leq \sqrt{N\lambda_I}\end{aligned}$$

Therefore, by Chebyshev's inequality this means, $N_I \approx N\lambda_I \pm C\sqrt{N\lambda_I}$ with high probability.

Note

However, once the points are given, it is possible to choose a subset I' such that $N_{I'} = 0$. Therefore, the above result holds only if the set is chosen predefined, and then the points are randomly distributed.

However, when we perform statistical methodologies, we choose the random sample first, and then analyze the data using different models. It is not that we fix the entire inference procedure beforehand, even before collecting the data. This identifies a core problem in all of statistical studies that we do currently.

This means we want some kind of result that holds regardless of the procedure used (or simultaneously for all nice classes of procedures). In the above example, it means to establish a result that is true for all nice sets I simultaneously. Usually we take this class of nice sets as the convex sets, or Euclidean balls or rectangles, etc.

Theorem 12. *A set of N i.i.d. uniform random points on $[0,1]^2$ satisfies the following with probability at least 0.99 (or any $1-\epsilon$ for a given $\epsilon > 0$): For all axis-aligned rectangle I, we have*

$$\lambda_I N - C\sqrt{\lambda_I N \log(N)} \le N_I \le \lambda_I N + C\sqrt{\lambda_I N \log(N)}$$

for an absolute constant C and sufficiently large N.

Note

1. We lose a $\sqrt{\log(N)}$ factor to establish the uniform bound.
2. Simple union bound may not be possible since there are infinite # of rectangles. The idea is similar to ϵ-nets and we consider ϵ-grid lines and the rectangles generated by them alone.

Proof. We use an ϵ-net argument, by considering ϵ-grid lines over $[0,1]^2$ and the rectangles generated by them. We call these **net rectangles**.

Step 1 (Concentration): Fix any net rectanlge I, with area λ_I. Then, as $N_I \sim \text{Binom}(N, \lambda_I)$, by Chernoff's inequality, we have

$$\begin{aligned} &\mathbb{P}\left(|N_I - \lambda_I N| \ge \delta \lambda_I N\right) \le 2e^{-\delta^2 \lambda_I N/6} \\ \implies &\mathbb{P}\left(|N_I - \lambda_I N| \ge \lambda_I N \times C\frac{\sqrt{\log N}}{\sqrt{\lambda_I N}}\right) \le 2\exp\left[-C^2 \log N/6\right] \le \frac{1}{N^{100}}, \end{aligned}$$

for sufficiently large C and N. Here we choose $\delta = C\frac{\sqrt{\log N}}{\sqrt{\lambda_I N}}$.

Step 2 (Union Bound): Let B_I be the event that rectangle I is bad, i.e., the number of points does not fall within the required bound, i.e., $|N_I - \lambda_I N| \ge C\sqrt{\lambda_I N \log(N)}$. Now,

$$\mathbb{P}(\exists \text{ a bad } I) = \mathbb{P}(\cup B_I) \le \sum_I \mathbb{P}(B_I) \le \left(\frac{1}{\epsilon}\right)^8 \frac{1}{N^{100}}.$$

By choosing $\epsilon = 1/N^{10}$, we have the probability that there is a bad net rectangle I to be less than $\frac{1}{N^{20}}$, which can be made smaller than 0.99 by choosing sufficiently large N.

Step 3 (Approximation): This result so far holds for uniformly with high probability, but only for net rectangles. We need to show now that the number of points in any rectangle can be approximated well by the number of points in the net rectangle. Let J

be another rectangle, then there is a net rectangle I (slightly smaller) such that

$$\lambda_J \leq \lambda_I \leq \lambda_J + \frac{4}{\epsilon} = \lambda_J + \frac{4}{N^{10}}.$$

Therefore,

$$\begin{aligned}
N_J \leq N_I &\leq \lambda_I N + C\sqrt{\lambda_I N \log(N)}, \quad \text{since } I \text{ is good} \\
&\leq \left(\lambda_J + \frac{4}{N^{10}}\right) N + C\left(\left(\lambda_J + \frac{4}{N^{10}}\right) N \log(N)\right)^{1/2} \\
&\leq \lambda_J N + C\sqrt{\lambda_J N \log(N)} + \left(\frac{4}{N^9} + C\frac{\sqrt{4\log(N)}}{N^9}\right), \text{ since } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \\
&\leq \lambda_J N + 2C\sqrt{\lambda_J N \log(N)}, \text{ since the second term is smaller than } \sqrt{\lambda_J N \log(N)}
\end{aligned}$$

This means, we only lose a factor of 2. □

2.3 Spaces of Random Variables

Definition 6. *A **normed space** is a vector space M over field F that is equipped with a norm $\|\cdot\|$. The norm $\|\cdot\| : M \to \mathbb{R}$ is a function such that*

1. $\|x\| \geq 0$ *for all* $x \in M$ *with equality if and only if* $x = 0$.
2. $\|\alpha x\| = |\alpha| \, \|x\|$ *for all* $\alpha \in F$ *and* $x \in M$.
3. $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in M$.

i.e., it has a measure of length.

Definition 7. *A **Hilbert** space is a vector space H over field F equipped with an inner product, $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$ such that*

1. $\langle x, x \rangle \geq 0$ *for all* $x \in H$ *with equality if and only if* $x = 0$.
2. $\langle x, y \rangle = \langle y, x \rangle$ *for all* $x, y \in H$.
3. $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$, *where* $a, b \in F$ *and* $x, y, z \in H$.

i.e., it has a measure of angle, which in turn, defines length.

Clearly, a Hilbert space is a normed space with $\|x\| = \sqrt{\langle x, x \rangle}$. We immediately have the Cacuhy-Schwartz inequality,

$$| \langle x, y \rangle | \leq \|x\| \, \|y\|$$

In case of L^2 space of random variables, this converts into $|\mathbb{E}(XY)| \leq (\mathbb{E}X^2)^{1/2}(\mathbb{E}Y^2)^{1/2}$. We also have a more general Hölder's inequality,

$$|\mathbb{E}(XY)| \leq \left(\mathbb{E}\,|X|^p\right)^{1/p} \left(\mathbb{E}\,|X|^q\right)^{1/q} = \|X\|_{L^p} \, \|X\|_{L^q},$$

where $(1/p + 1/q) = 1$ for some $p, q \geq 0$.

A more general is the **Jensen's** inequality, which says:

Theorem 13 (Jensen's inequality)**.** *Let X be a real-valued random variable and $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then,*

$$\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X)).$$

A fact that follows from the above is that the L^p norm $\|X\|_{L^p}$ is an increasing function of p for all $p \geq 1$. Before, we show the proof for this fact.

Proof. We will show that if $p \leq q$, then $\|X\|_{L^p} \leq \|X\|_{L^q}$. If $p < q$, choose a function $\phi(x) = x^{q/p}$, is convex as $q > p$ or $q/p > 1$.

By applying Jensen's inequality on this function with random variable $|X|^p$,

$$\begin{aligned} & \phi(\mathbb{E}\,|X|^p) \leq \mathbb{E}(\phi(|X|^p)) \\ \implies & (\mathbb{E}\,|X|^p)^{q/p} \leq \mathbb{E}(|X|^q) \\ \implies & \|X\|^{L^p} \leq \|X\|_{L^q} \end{aligned}$$

where the last line follows from taking p-th root from both sides, and since $p \geq 1$. □

As a result of this, we have

$$\lim_{p\to\infty} \|X\|_{L^p} = \|X\|_{L^\infty} =: \text{essential supremum of } |X|.$$

The essential supremum is the supremum of a random variable over a set of probability 1 discarding the set of measure zero. Also, we have,

$$L^1 \supset L^2 \supset \cdots \supset L^\infty,$$

where L^k is the space having finite k-th order moment and the L^∞ is the space of random variables that are bounded almost surely.

Note

Note that, $\cap_{p\geq 1} L^p \neq L^\infty$, since there are random variables (e.g. $N(0,1)$) that has finite moments for all order but still is not bounded.

This means, there must exists special kinds of norms that we can bound so that the resulting spaces remain in this gap. In general, we understand that L^p spaces look at the p-th moment by considering the function x^p. But we could look at finite expectation of different moments like e^x, or $\sin(x^2)$, or something like this.

Definition 8 (Orlicz Functions)**.** *A function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is called an **Orlicz function** if it is convex, increasing, and,*

$$\psi(x) \to \begin{cases} 0, & \text{as } x \to 0 \\ \infty, & \text{as } x \to \infty \end{cases}$$

e.g. $\psi(x) = x^p$ or $\psi(x) = e^x - 1$.

Definition 9 (Orlicz norm)**.** *For a given Orlicz function ψ, the Orlicz "norm" of a random variable X is defined as*

$$\|X\|_\psi = \inf\{k > 0 : \mathbb{E}[\psi(|X|/k)] \leq 1\}$$

To ensure that this definition is valid, we need to show the resulting quantity indeed defines a norm. Following is the justification for this.

1. Clearly, $\|X\|_\psi \geq 0$ for any random variable X, because it is infinimum of a set with positive quantities.

2. If $X \equiv 0$ almost surely, then by definition of Orlicz function, $\psi(|X|/k) = \psi(0) \equiv 0$ almost surely for any choice of $k > 0$. Therefore, the Orlicz norm is zero. Conversely, suppose that that Orlicz norm is 0 but X is not equal to 0 almost surely. Then, for any $k > 0$, we must have $\mathbb{E}(\psi(|X|/k)) \leq 1$. And since it is not true that X is zero almost surely, there exists $\epsilon > 0$ such that $|X| > \epsilon$ with probability at least δ for some $\delta > 0$. Also, since the Orlicz function is increasing, convex and $\psi(x) \to \infty$ as $x \to \infty$, there exists $M > 0$ such that $\psi(x) > 1/\delta$ for all $x \geq M$. Therefore, by choosing $k = \epsilon/M$, we have

$$\mathbb{E}(\psi(|X|/k)) = \int \psi(|X| M/\epsilon) dP(x) \geq \int_{|x|>\epsilon} \psi(|x| M/\epsilon) dP(x) \geq \psi(M)\delta \geq 1 > 0.$$

   This leads to a contradiction.

3. The linearity of the Orlicz norm is trivial.

4. For the triangle inequality, we need to know that for any two random variables X and Y, $\|X+Y\|_\psi \leq \|X\|_\psi + \|Y\|_\psi$. By definition of Orlicz norm, it is enough to show that for $k = \|X\|_\psi + \|Y\|_\psi$,

$$\mathbb{E}\psi(|X+Y|/k) \leq 1.$$

   To see this, note that

$$\begin{aligned}
& \mathbb{E}\psi(|X+Y|/k) \\
\leq\ & \mathbb{E}\psi(|X|/k + |Y|/k), \text{ since } \psi \text{ is increasing} \\
=\ & \mathbb{E}\psi\left(\frac{|X|}{\|X\|_\psi}\frac{\|X\|_\psi}{\|X\|_\psi+\|Y\|_\psi} + \frac{|Y|}{\|Y\|_\psi}\frac{\|Y\|_\psi}{\|X\|_\psi+\|Y\|_\psi}\right) = \quad \frac{\|X\|_\psi}{\|X\|_\psi+\|Y\|_\psi}\mathbb{E}\psi\left(\frac{|X|}{\|X\|_\psi}\right) + \frac{}{\|X\|} \\
\leq\ & \frac{\|X\|_\psi}{\|X\|_\psi+\|Y\|_\psi} + \frac{\|Y\|_\psi}{\|X\|_\psi+\|Y\|_\psi} = 1.
\end{aligned}$$

   The last inequality follows from the definition of $\|X\|_\psi$ and $\|Y\|_\psi$.

Now that we know that it is indeed a well-defined norm, we can define the Orlicz space as follows.

$$L_\psi = \left\{X : \|X\|_\psi < \infty\right\}$$

To illustrate further, consider the following examples.

1. If $\psi(x) = x^p$, then $\mathbb{E}[\psi(|X|/k)] = \mathbb{E}[(|X|/k)^p] \leq 1 \implies k \geq (\mathbb{E}[|X|^p])^{1/p}$ and $\|X\|_\psi = \|X\|_p$.

2. If $\psi(x) = e^x - 1$, then $\mathbb{E}\psi(|X|/k) = \mathbb{E}[e^{|X|/k} - 1] \leq 1 \implies \mathbb{E}[e^{|X|/k}] \leq 2$, which is something related to the MGF of a random variable.

Now turning our attention to the Hoeffding's inequality, we note that if X_is are i.i.d. symmetric Bernoulli random variables, then $\bar{X}_n = \sum_{i=1}^n X_i/n$ has Gaussian tails. The same thing holds if X_i are i.i.d. standard normal random variables. Also, you can verify that if X_i are i.i.d. uniform random variables on $[-1/2, 1/2]$, then also $\bar{X}_n$ has Gaussian tails.

Hence, the question is to find out the biggest class of distributions for which we have Gaussian tails.

Turns out, if $n = 1$, then $\bar{X}_n = X_1$ needs to satisfy,

$$\mathbb{P}(|X_1| \geq t) \leq 2e^{-ct^2/2}, \forall t > 0.$$

since the above needs to be true for all choices of n. Therefore, this is a necessary condition. Turns out this is also sufficient for the Hoeffding lemma to hold. The random variables which has tails satisfying the above inequality are called **sub-Gaussian random variables**.

2.3.1 Sub Gaussian Random Variables

Before we look at the general behaviour of them, let us look at the special case for $X_i \sim N(0, 1)$. In this case,

1. **Tails:** We have $\mathbb{P}(|X_i| \geq t) \leq 2e^{-t^2/2}$, $\forall t > 0$.

2. **Moments:** $\mathbb{E}(X_i^p) = 0$ if p is odd and is equal to $(p-1)!!$ if p is even. This means that the L^p norm of the Gaussian random variable satisfy

   $$\|X_i\|_p = (\mathbb{E}(|X_i|^p))^{1/p} \leq ((p-1)(p-3)\dots 5 \times 3 \times 1)^{1/p} \leq (p^{p/2})^{1/p} = \sqrt{p}$$

   for all $p \geq 1$.

3. **MGF:** The MGF of X_i is given by $\mathbb{E}(e^{tX_i}) = e^{t^2/2}$.

4. **MGF of square:** The MGF of X_i^2 is given by $\mathbb{E}(e^{tX_i^2}) = \dfrac{1}{\sqrt{1-2t}}$, if $t < 1/2$. Otherwise it is infinite. Therefore, if we take the Orlicz function $\psi_2(x) = e^{x^2} - 1$, then

   $$\begin{aligned} & \mathbb{E}\psi(|X_i|/k) \leq 1 \\ \implies & \mathbb{E}\left[e^{X_i^2/k^2}\right] \leq 2 \\ \implies & \frac{1}{\sqrt{1 - 2 \times 1/k^2}} \leq 2 \\ implies k \geq \sqrt{2}. \implies & \qquad \|X_i\|_{\psi_2} \leq \sqrt{2}. \end{aligned}$$

The following lemma establishes that this connection between the finiteness of all these quantities is not special for Gaussian random variables, but for all sub-Gaussian random variables.

Lemma 3 (sub-Gaussian Lemma)**.** *For all random variable X, the following the equivalent.*

1. *$\exists K_1$ such that $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/K_1^2}$, for all $t > 0$.*
2. *$\exists K_2$ such that $\|X\|_p \leq K_2\sqrt{p}$, for all $p \geq 1$.*
3. *$\exists K_3$ such that $\mathbb{E}\exp(X^2/K_3^2) \leq 2$.*
4. *If $\mathbb{E}(X) = 0$, then $\exists K_4$ such that $\mathbb{E}\exp(\lambda X) \leq e^{\lambda^2 K_4^2}$, for all $\lambda \in \mathbb{R}$.*

These constants $K_1, \ldots, K_4$, are all equivalent to the sub-Gaussian norm $\|X\|_{\psi_2}$.

Proof. (1) $\implies$ (2)**:** Without loss of generality, assume that $K_1 = 1$. Then,

$$\begin{aligned}
\mathbb{E}\,|X|^p &= \int_0^\infty \mathbb{P}(|X|^p > t)dt \\
&= \int_0^\infty \mathbb{P}(|X| > s)ps^{p-1}ds, \quad \text{by substitution of } t = s^p \\
&\leq \int_0^\infty 2e^{-s^2}ps^{p-1}ds \\
&= p(p/2-1)! \quad \text{as it is Gamma integral} \\
&\leq pp^{p/2-1} = p^{p/2}.
\end{aligned}$$

Therefore, $\|X\|_{L^p} \leq \sqrt{p}$.

(2) $\implies$ (3)**:** Without loss of generality, assume $K_2 = 1$. Then,

$$\begin{aligned}
\mathbb{E}(e^{X^2/100}) &= \mathbb{E}\left(\sum_{p=0}^\infty \frac{(X^2/100)^p}{p!}\right) \\
&= \sum_{p=0}^\infty \frac{1}{p!(100)^p}\mathbb{E}(X^{2p}) \\
&\leq \sum_{p=0}^\infty \frac{1}{p!(100)^p}(2p)^p \\
&\text{since by Stirling's approximation } p! \geq e^{-p}p^p \\
&\leq \sum_{p=0}^\infty \frac{(2p)^p}{(100)^p(p/e)^p} \\
&= \sum_{p=0}^\infty \left(\frac{2e}{100}\right)^p \leq 2.
\end{aligned}$$

(3) $\implies$ (4)**:** Without loss of generality, assume that $K_3 = 1$. We know that $e^x \leq x + e^{x^2}$ for all $x \in \mathbb{R}$. Therefore,

$$\mathbb{E}(e^{\lambda X}) \leq \mathbb{E}(\lambda X) + \mathbb{E}(e^{\lambda^2 X^2})$$

$$
\begin{aligned}
&= 0 + \left(\mathbb{E}(e^{X^2})\right)^{\lambda^2}, \quad \text{assume } |\lambda| \leq 1 \\
&\leq 2^{\lambda^2} = e^{\lambda^2 \log(2)}.
\end{aligned}
$$

(4) $\implies$ (1): Without loss of generality, assume that $K_4 = 1$. Then,

$$
\begin{aligned}
\mathbb{P}(|X| \geq t) &= \mathbb{P}(e^{\lambda|x|} \geq e^{\lambda t}) \\
&\leq e^{-\lambda t}\mathbb{E}(e^{\lambda|X|}), \quad \text{by Markov's inequality} \\
&\leq e^{-\lambda t} e^{\lambda^2} \\
&= e^{-\lambda t + \lambda^2}
\end{aligned}
$$

Optimizing over λ, we get $\lambda = t/2$, yielding the bound $e^{-t^2/4}$. □

Lemma 4. *If X_is are independent mean zero sub-Gaussian random variables, then $\|\sum_i X_i\|_{\psi_2} \leq C \sum_{i=1}^n \|X_i\|_{\psi_2}$ for some absolute constant C.*

Proof. We will make use of the sub-Gaussian Lemma. Note that,

$$
\begin{aligned}
\mathbb{E}(e^{\lambda \sum_i X_i}) &= \prod_{i=1}^n \mathbb{E}(e^{\lambda X_i}), \quad \text{by independence} \\
&\leq \prod_{i=1}^n e^{C\|X_i\|_{\psi_2}^2 \lambda_i^2} \\
&= e^{C\sum_i \|X_i\|_{\psi_2}^2 \lambda_i^2}.
\end{aligned}
$$

By equivalence of the sub-Gaussian properties, we get that $\sum_i X_i$ is also a sub-Gaussian random variable and that

$$
\left\| \sum_i X_i \right\|_{\psi_2} \leq C' \sum_{i=1}^n \|X_i\|_{\psi_2}
$$

for some constant C'. □

As a result, the Hoeffding's inequality follows for sub-Gaussian random variables as well, because the tail behaviour indicates Gaussianity.

Theorem 14 (Hoeffding's inequality for sub-Gaussian random variables)**.** *If X_i are independent mean zero sub-Gaussian random variables, then*

$$
\mathbb{P}\left(\left| \sum_i X_i \right| \geq t \right) \leq 2e^{-Ct^2/\sigma^2}
$$

where $\sigma^2 = \sum_{i=1}^n \|X_i\|_{\psi_2}^2$. This σ^2 is kind of a proxy for the variance.

However, it turns out that the sub-Gaussian class is not large enough to capture all kind of random variables that we usually deal with. For example, if X is sub-Gaussian, then X^2 is not sub-Gaussian.

$$
\mathbb{P}(X^2 \geq t) = \mathbb{P}(|X| \geq \sqrt{t}) \asymp e^{-c(\sqrt{t})^2} = e^{-ct} >> e^{-ct^2}, \ \forall t > 0.
$$

The tail in this case is heavier than the Gaussian case, but it still decays exponentially. We call this a **sub-exponential** random variable.

2.3.2 Sub-exponential Random Variables

A similar lemma as before can be stated for sub-exponential random variables.

Lemma 5 (sub-exponential Lemma)**.** *For all random variable X, the following the equivalent.*

1. *$\exists K_1$ such that $\mathbb{P}(|X| \geq t) \leq 2e^{-t/K_1}$, for all $t > 0$.*
2. *$\exists K_2$ such that $\|X\|_{L^p} \leq K_2 p$, for all $p \geq 1$.*
3. *$\exists K_3$ such that $\mathbb{E}\exp(|X|/K_3) \leq 2$.*
4. *$\exists K_4$ such that $\mathbb{E}\exp(\lambda X) \leq e^{\lambda^2 K_4^2}$, for all $|\lambda| \leq 1/K_4$.*

These constants $K_1, \ldots, K_4$, are all equivalent to the sub-exponential norm $\|X\|_{\psi_1}$.

A particular thing to note that is the last bound on the MGF does not hold for all λ, because the MGF may not converge for all λ. To understand this, note that

$$\mathbb{E}(e^{\lambda X}) = \int_0^\infty e^{\lambda x} p(x) dx \asymp \int_0^\infty e^{\lambda x} e^{-cx} dx$$

so the integral converges if and only if $\lambda < c$.

Clearly, Hoeffding's inequality does not hold for sub-exponential random variables in general. Since if $n = 1$, we cannot have $\mathbb{P}(|X| \geq t) \leq e^{-ct^2}$. However, once we accumulate multiple independent X_is, we can still obtain a Gaussian tail behaviour.

Theorem 15 (Bernstein's inequality)**.** *Let $X_i s$ are independent mean zero sub-exponential random variables. Then,*

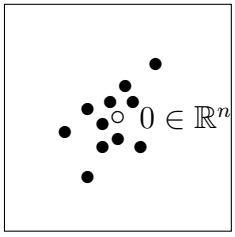
$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left[-C\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{K}\right\}\right]$$

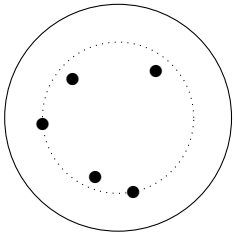
where $\sigma^2 = \sum_{i=1}^n \|X_i\|_{\psi_1}^2$ and $K = \max_{i=1}^n \|X_i\|_{\psi_1}$.

2.4 Applications: Part 2

2.4.1 Thin Shell Phenomenon

Consider a Gaussian random vector $g \sim N(0, I_n)$. We expect the samples from this $N(0, I_n)$ to look like the left one, but the actual picture looks like the right one in a high-dimensional space.





This means the samples concentrate on a thin ring close the to boundary of a the $\sqrt{n}$-ball, instead of concentrating near zero. The next theorem establishes this fact.

Theorem 16. *For $g \sim N(0, I_n)$, $\mathbb{P}(0.99\sqrt{n} \leq \|g\|_2 \leq 1.01\sqrt{n}) \geq 1 - 2e^{-cn}$, for some sufficiently small but absolute constant c.*

Proof. Note that,

$$\|g\|_2^2 - n = \sum_{i=1}^{n}(g_i^2 - 1) = \sum_{i=1}^{n}(g_i^2 - \mathbb{E}(g_i^2))$$

Since the above is a sum of i.i.d. terms, we can apply concentration inequalities. In particular, g_i is sub-Gaussian, so g_i^2 is only sub-exponential, hence we need to apply Bernstein's inequality. Using that, we get

$$\mathbb{P}\left(|\,\|g\|_2^2 - n| \geq 0.01n\right) \leq 2\exp\left[-c\min\left(\frac{n^2}{\sigma^2}, \frac{n}{K}\right)\right]$$

where

$$\sigma^2 = \sum_{i=1}^{n}\left\|g_i^2 - 1\right\|_{\psi_1} \leq nC$$
$$K = \max_i \left\|g_i^2 - 1\right\|_{\psi_1} = \left\|g_1^2 - 1\right\|_{\psi_1} (\text{ as i.i.d.}) \leq C$$

Therefore, we have

$$\mathbb{P}\left(0.99n \leq \|g\|_2^2 - n \leq 1.01n\right) \geq 1 - 2e^{-c'n}$$

for some c'. Now, note that

$$\frac{\|g\|_2 - \sqrt{n}}{\sqrt{n}} = \frac{\|g\|_2^2 - n}{\sqrt{n}(\|g\|_2 + \sqrt{n})} = \frac{\leq 0.01n}{\geq \sqrt{n} \times \sqrt{n}} \leq 0.01n,$$

Therefore, the result follows. □

2.4.2 Dimension Reduction

Let $x_1, x_2, \ldots x_N \in \mathbb{R}^d$ be a set of d-dimensional vectors. We want to reduce or translate them into $y_1, y_2, \ldots y_N \in \mathbb{R}^n$ where $n << d$ so that the pairwise distances are approximately preserved. Turns out, to do this effectively, we can choose n to be as small as $O(\log N)$. This means, for N datapoints, we can effectively capture them in a $\log(N)$ dimensional vector space without losing pairwise distances. This was proved by Johnson and Linderstrass in 1984.

Theorem 17 (Johnson Linderstrass, 1984)**.** *For all $x_1, \ldots x_N \in \mathbb{R}^d$, there exists a linear map $T : \mathbb{R}^d \to \mathbb{R}^n$ such that with $n \leq C\log(N)$ for some absolute constant C, we have*

$$0.99\,\|x_i - x_j\|_2 \leq \|T(x_i) - T(x_j)\|_2 \leq 1.01\,\|x_i - x_j\|_2\,,$$

for all $i, j = 1, 2, \ldots N$.

Proof. Choose a random map T corresponding to a random matrix $1/\sqrt{n}G$ where G is a random Gaussian matrix, i.e., $G_{ij} \sim N(0,1)$.

Fix any $z \in \mathbb{R}^d$ such that $\|z\|_2 = 1$. Then, $Gz \sim N(0, I_n)$. Therefore, by Thin-shell phenomenon, we have

$$\mathbb{P}\left(0.99\sqrt{n} \leq \|Gz\|_2 \leq 1.01\sqrt{n}\right) \geq 1 - 2e^{-cn}.$$

Now for any i and j such that $x_i \neq x_j$, using the above, we have

$$\mathbb{P}\left(0.99\sqrt{n} \leq \left\|G\frac{(x_i - x_j)}{\|x_i - x_j\|_2}\right\|_2 \leq 1.01\sqrt{n}\right) \geq 1 - 2e^{-cn}$$
$$\implies \mathbb{P}\left(0.99\,\|x_i - x_j\|_2 \leq \|T(x_i) - T(x_j)\|_2 \leq 1.01\,\|x_i - x_j\|_2\right) \geq 1 - 2e^{-cn}$$

Now, an application of union bound reveals that, the probability that the above event happens for all pair of points x_i and x_j, is at least $1 - 2N^2e^{-cn}$. By choosing $n = 4\log(N)/c$, we get that the lower bound is at least $1 - 2/N^2$, which is positive for all $N > 2$.

This means, the probability that a random map will satisfy the criterion is strictly positive, hence there exists such a projection that preserves the pairwise distances approximately. □

3 Combinatorial Optimization

Combinatorial optimizations refer to the class of optimization problems that usually requires integer or natural number solutions, also sometimes restricted to have an upper bound or lower bound limit, or takes values in a finite set. Let us look at a few example problems.

Example 2. *Given a list of numbers $\{a_i\}_{i=1}^n$, find the maximum possible value of $\sum_{i=1}^n a_i x_i$ subject to the restriction that $x_i \in \{+1, -1\}$. This is a linear objective function, and there exists a direct solution by taking $x_i = sign(a_i)$.*

Example 3. *Given a list of numbers $\{a_{ij}\}_{i,j=1}^n$, find the maximum possible value of $\sum_{i=1}^n \sum_{j=1}^n a_{ij}x_iy_j$ subject to the restriction that $x_i \in \{+1, -1\}, y_j \in \{+1, -1\}$. This is a quadratic objective function. Turns out this is an NP-hard problem. Even if we restrict $x_i = y_i$, the problem still remains NP-hard.*

Therefore, we should look at some kind of relaxation in order to be able to solve this problem efficiently.

1. **Spectral Relaxation:** Currently, we need to optimize over the boolean cube $\{-1, +1\}^n$. Instead, we can optimize this over the $\sqrt{n}$-ball $B(0, \sqrt{n})$. This allows us to algebraic manipulation to use theories of vector space. In particular, note that

$$\max_{x \in \{-1,+1\}^n} \sum_{i,j} a_{ij}x_ix_j \leq \max_{\|x\| \leq \sqrt{n}} \sum_{i,j} a_{ij}x_ix_j = (\sqrt{n})^2 \max_{\|x\| \leq 1} x^\intercal Ax = n\lambda_1(A)$$

and the maximizing x is the eigenvector of A corresponding to the largest eigenvalue. Here, we optimize an upper bound instead of the original objective function. However, using the upper bound allows us to apply Spectral theory and find a solution analytically. To get a proper solution to the boolean cube, we can then take $x_i = \text{sign}(e_{1i}(A))$, i.e., the signs of the coordinates of the first eigenvetor $e_1(A)$.

2. **Semi-definite programming:** The other relaxation is to reparametrize the system bringing in new variables $z_{ij} = x_i x_j$ which leads to the linear objective function $\sum_{i,j} a_{ij} z_{ij}$, which is easy to solve. However, the matrix Z comprising of the elements z_{ij} need to satisfy a few broad conditions:

   (a) Z is symmetric, since $z_{ij} = x_i x_j = x_j x_i = z_{ji}$.

   (b) $Z_{ii} = x_i^2 = 1$, i.e., the diagonals are equal to 1.

   (c) Z is positive definite.

   Therefore, one can consider a relaxed version of the problem with new parameter z_{ij} where Z satisfy the above conditions, and the objective function becomes linear. Moreover, the constraints lead to a convex set hence the entire system can be efficiently solved by using convex programming.

Now, we present a few small applications where this kind of combinatorial optimization problem may appear.

Ising Model: Ising model is a model to explain electromagnetism. According to this model, every atom has a spin or polarity. When electricity is passed through the object, the polarity of the atoms of that object lines up together, and the overall magnetism property starts building up. Let x_i be the state of the i-th atom (i.e., $+1$ or -1, depending on the polarity of the atom), and let a_{ij} be the interaction strength between atom i and atom j (which is reciprocal of the squared distance between atom i and atom j). Then, the overall energy of the system is given by the Hamiltonian,

$$H(x) = \sum_{i,j} a_{ij} x_i x_j,$$

which is also called free energy. Given a material, the objective is to find the maximum energy or magnetism possible.

Clustering of Points: Let $v_1, v_2, \ldots v_n \in \mathbb{R}^d$ be points, and let $a_{ij} = \|v_i - v_j\|_2$. We want to cluster these points. Let, x_i be a quantity that takes values $+1$ or -1 depending on which cluster the point v_i is in. Now, the core objective of the clustering should be,

1. If $x_i x_j = 1$, then a_{ij} should be small since both v_i and v_j belongs to the same cluster.

2. If $x_i x_j = -1$, then a_{ij} would be large as they belong to the different clusters.

Therefore, the objective is to minimize the objective function

$$\min_{x_i \in \{-1,+1\}} \sum_{i,j} a_{ij} x_i x_j = \sum \text{distance in cluster 1} + \sum \text{distance in cluster 2} - \sum \text{distance between clusters}$$

Frieze-Jerrum (1997) generalizes the same idea for multiple clusters and solves it.

Max-cut of a graph: Given a graph $G = (V, E)$, the max cut problem aims to find two disjoint partitions V_1 and V_2 of the vertex set V such that the # of edges between V_1 and V_2 are largest possible. Therefore, the maximum cut objective function is

$$\textbf{Max cut of } G = \max_{V = V_1 \cup V_2} |E(V_1, V_2)|.$$

To solve this, consider the adjacency matrix A such that $a_{ij} = \mathbf{1}((i,j) \in E)$. Let, $x_i = 1$ if the vertex i is in V_1, otherwise it is (-1). Therefore,

$$(1 - x_i x_j) = \begin{cases} 2 & \text{if } (i,j) \in E(V_1, V_2) \\ 0 & \text{otherwise} \end{cases}$$

So, $E(V_1, V_2)$ is the number of edges between crossing over two partitions and is given by $\sum_{i,j} a_{ij}(1 - x_i x_j)/4$.

Now that we have seen some applications where this combinatorial problem is applicable, and also seen how we can create a relaxation of this, let us try to solve the relaxed problem. We will show that in certain cases, the solution to the relaxed version is not very far from the largest value of the quadratic forms.

Definition 10 (Gram Matrix)**.** *The gram matrix of the vectors $v_1, v_2, \dots v_n \in \mathbb{R}^d$ is a $n \times n$-matrix with entries $\langle v_i, v_j \rangle$.*

Lemma 6. *For any $n \times n$ positive semi-definite matrix, there exists a choice of vectors $v_1, v_2, \dots v_n$ such that it can be written as a Gram matrix of this set of vectors.*

Proof. The proof follows easily from the Cholesky decomposition of the p.d. matrix. □

Using this lemma along with the semi-definite relaxation, we can write the matrix Z as a Gram matrix, so that $z_{ij} = \langle v_i, v_j \rangle$. Also, we need $z_{ii} = 1 = \|v_i\|_2^2$, i.e., the vectors $v_1, \dots v_n$ are unit vectors. Therefore, the objective becomes

$$\max_{z_{ij}} \sum_{i,j} a_{ij} z_{ij} = \max_{v_i \in \mathbb{R}^n, \|v_i\|=1} \sum_{i,j} a_{ij} \langle v_i, v_j \rangle$$

Compare this to the original combinatorial problem

$$\max_{x_i \in \{-1,+1\}} \sum_{i,j} a_{ij} x_i x_j.$$

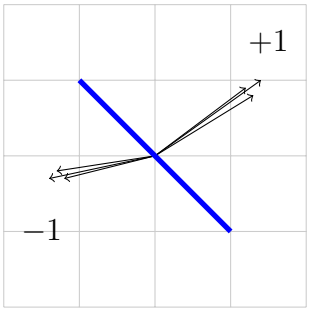
One may think of these x_is as unit vectors in $\mathbb{R}^1$ instead of $v_i \in \mathbb{R}^n$.

Note

It says that the semi-definite relaxation of the optimization problem replaces that one-dimensional vectors with n-dimensional vectors, (apparently that would have made things more difficult, but this is not the case here).

Now, given a solution to the SDP (semi-definite problem), we can get back the original x_i and x_j but probably clustering the original vectors v_is together. Ideally, we should have a clustering of the vectors as follows. Then, we can draw a line to separate them, that accordingly we can pick x_i and x_j. If the separation vector is v, then we would be looking at $x_i := \text{sign}(\langle v, v_i \rangle)$.

Turns out, even a randomized cut (the algorithm is called **Randomized Rounding**), also works equally well in most cases. This means, we simply that $g \sim N(0, I_n)$, a



randomly distributed Gaussian vector, and use the mapping $v \to \text{sign}(\langle v, g\rangle)$ to get the solution.

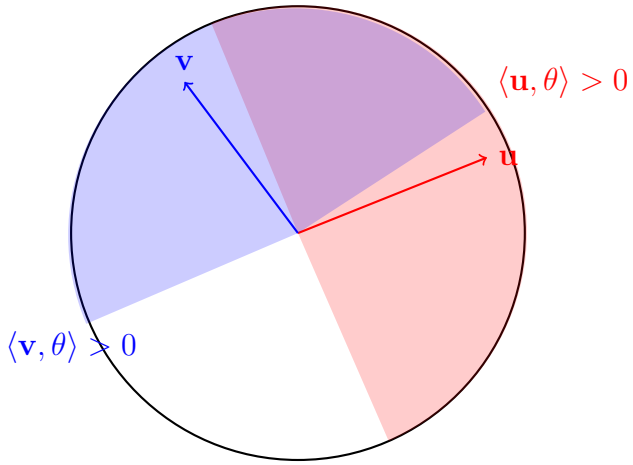
Lemma 7 (Grothendieck's identity)**.** *For any vectors $u, v \in \mathbb{R}^n$, such that $\|u\| = \|v\| = 1$, and a random vector $g \sim N(0, I_n)$, the following holds:*

$$\mathbb{E} sign(\langle u, g\rangle) sign(\langle v, g\rangle) = \frac{2}{\pi} \sin^{-1}(\langle u, v\rangle)$$

Proof. Note that, if $\theta = g/\|g\|$, then θ is a random vector uniformly distributed on the unit n-dimensional hypersphere. Let P be the plane on which both the vectors u and v reside. Since $S \cap P$ is the unit circle, θ restricted to P becomes uniformly distributed on a unit circle.

Hence, it is enough to restrict the problem to the plane P since none of the inner product changes because of that. Therefore, essentially it is enough to prove the above for only $n = 2$ case.

Let us now assume $u, v \in \mathbb{R}^2$ and θ is a uniformly distributed vector on the unit circle. Let's try to find out the region on the choice of θ for which we have $\text{sign}(\langle u, \theta\rangle) = 1$, i.e., it is the region where $\langle u, \theta\rangle > 0$. This means, we need to look at the perpendicular line $u^\perp$, and the region on the side of u is this particular region where $\langle u, \theta\rangle > 0$.



By using this idea, we note that,

$$\text{sign}(\langle u, \theta\rangle)\text{sign}(\langle v, \theta\rangle) = \begin{cases} 1 & \text{where } \theta \in [-v^\perp, u^\perp] \cup [v^\perp, -u^\perp] \\ (-1) & \text{where } \theta \in [u^\perp, v^\perp] \cup [-u^\perp, -v^\perp]. \end{cases}$$

Let, α be the angle between u and v. Then,

$$\begin{aligned}
\mathbb{E}\text{sign}(\langle u, \theta\rangle)\text{sign}(\langle v, \theta\rangle) &= \frac{1}{2\pi}\text{arclength}\left(2 \times (2(\pi/2 - \alpha) + \alpha) - (2\pi - 2 \times (2(\pi/2 - \alpha) + \alpha))\right) \\
&= \frac{1}{2\pi}\text{arclength}\left(2\pi - 2\alpha - (2\pi - (2\pi - 2\alpha))\right) \\
&= \frac{1}{2\pi}\text{arclength}(2\pi - 4\alpha) \\
&= 1 - \frac{4}{2\pi}\cos^{-1}(\langle u, v\rangle) \\
&= \frac{2}{\pi}\sin^{-1}(\langle u, v\rangle)
\end{aligned}$$

□

Theorem 18 (Goeman and Williamson's algorithm)**.** *The expected cut based on the randomized rounding partition of the solution to the SDP problem is at least* 0.878 *times of the solution to the max-cut problem.*

Proof. The expected cut for a graph G with partitions given by x_is, we have

$$\begin{aligned}
\text{Expected cut} &= \mathbb{E}\left(\frac{1}{4}\sum_{i,j} a_{ij}(1 - x_i x_j)\right) \\
&= \frac{1}{4}\sum_{i,j} a_{ij}(1 - \mathbb{E}x_i x_j) \\
&= \frac{1}{4}\sum_{i,j} a_{ij}(1 - \mathbb{E}\text{sign}(\langle v_i, g\rangle)\text{sign}(\langle v_j, g\rangle)) \\
&= \frac{1}{4}\sum_{i,j} a_{ij}(1 - 2/\pi\sin^{-1}(\langle v_i, v_j\rangle)), \quad \text{by Grothendieck's identity} \\
&= \frac{1}{4}\sum_{i,j} a_{ij}\frac{2}{\pi}\cos^{-1}(\langle v_i, v_j\rangle) \\
&\geq \frac{1}{4}\sum_{i,j} a_{ij} \times 0.878(1 - \langle v_i, v_j\rangle), \quad \text{by linearization of arc cosine function} \\
&\qquad \text{since } 2/\pi\cos^{-1}(\theta) \geq 0.878(1 - \theta) \\
&= 0.878 \times \left(\frac{1}{4}\sum_{i,j} a_{ij}(1 - \langle v_i, v_j\rangle)\right) \\
&= 0.878 \times \text{SDP}(G) \geq 0.878 \times \text{Max cut}(G)
\end{aligned}$$

□

While the above shows that the randomized rounding algorithm does not lose much in terms of tightness, one crude step we use here is that the maximum value of the SDP is at least as large as the maximum value of the original max-cut problem. The below result shows the reverse inequality up to a scaling factor to show that this relaxation is not too bad.

Theorem 19 (Grothendieck's inequality (1953))**.** *For any a_{ij} and any dimension d, we have*

$$\max_{u_i,v_j\in\mathbb{R}^d,\|u_i\|=\|v_j\|=1} \sum_{i,j} a_{ij} \langle u_i, v_j\rangle \leq 1.782 \times \max_{x_i,y_j\in\{+1,-1\}} \sum_{i,j} a_{ij}x_iy_j$$

Proof. The proof uses two techniques: First, a randomized rounding process. Second, a kernel trick to reach the intended nonlinearity to apply Grothendieck's lemma.

Suppose there exists two function ϕ, ψ such that

$$\langle \phi(u), \psi(v)\rangle = \sin\left(\frac{\beta\pi}{2}\langle u, v\rangle\right), \quad \text{where } \beta = \frac{2}{\pi}\ln(1+\sqrt{2}).$$

Then by Grothendieck's identity, we have

$$\mathbb{E}\text{sign}(\langle\phi(u_i), g\rangle)\text{sign}(\langle\psi(v_j), g\rangle) = \frac{2}{\pi}\sin^{-1}\left(\langle\phi(u_i), \psi(v_j)\rangle\right) = \beta\langle u_i, v_j\rangle.$$

Therefore,

$$\begin{aligned}
\sum_{i,j} a_{ij}\langle u_i, v_j\rangle &= \frac{1}{\beta}\sum_{i,j} a_{ij}\mathbb{E}\text{sign}(\langle\phi(u_i), g\rangle)\text{sign}(\langle\psi(v_j), g\rangle) \\
&= \frac{1}{\beta}\mathbb{E}\left(\sum_{i,j} a_{ij}x_i^*y_j^*\right), \text{ where } x_i^*, y_j^* \in \{-1, +1\} \\
&\leq \frac{1}{\beta}\max_{x_i,y_j\in\{-1,+1\}}\sum_{i,j} a_{ij}x_iy_j, \quad \text{and } 1/\beta = 1.782
\end{aligned}$$

Taking maximum on the left-hand side now completes the solution. □

Now, the proof relies on the existence of β, ϕ and the ψ functions. So, we will show the existence of these feature maps using tensor calculus. We shall do this in several steps.

1. Let $u \in \mathbb{R}^n$. Then define tensor product $u \otimes u = [u_iu_j]_{i,j=1}^n = uu^\intercal \in \mathbb{R}^{n^2}$. Note that,

$$\langle u\otimes u, v\otimes v\rangle = \sum_{i,j} u_iu_jv_iv_j = (\sum_i u_iv_i)(\sum_j u_jv_j) = \langle u, v\rangle^2.$$

2. Note that, similarly define $u \otimes u \otimes u$ to be the 3d-tensor with entries $u_iu_ju_k$ for all possible values of (i, j, k). Similar to above, $\langle u\otimes u\otimes u, v\otimes v\otimes v\rangle = \langle u, v\rangle^3$. In general,

$$\left\langle u^{\otimes k}, v^{\otimes k}\right\rangle = \langle u, v\rangle^k, \text{ for any } k \geq 1.$$

3. For any two vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ (m may be different from n), define the direct sum of them as

$$u \oplus v = [u_1, \ldots, u_n, v_1, \ldots, v_m] \in \mathbb{R}^{m+n}.$$

This helps us build,

$$\langle u\oplus v, x\oplus y\rangle = \langle u, x\rangle + \langle v, y\rangle, \text{ provided the dimensions match}$$

4. Therefore, by combining the above two procedures (tensor product and direct sum), we can build any polynomial of the inner products. In case we want a negative coefficient, it is possible to do so by carefully choosing ϕ and ψ to be different in one coordinate on the polynomial. For example, if we want to build $2\langle u, v\rangle + 3\langle u, v\rangle^2 - 5\langle u, v\rangle^3$, we take

$$\phi(u) = \sqrt{2}u \oplus \sqrt{3}u^{\otimes 2} \oplus \sqrt{5}u^{\otimes 3} \in \mathbb{R}^{n+n^2+n^3}$$
$$\psi(v) = \sqrt{2}v \oplus \sqrt{3}v^{\otimes 2} \oplus (-\sqrt{5})v^{\otimes 3} \in \mathbb{R}^{n+n^2+n^3}$$

5. For all real analytic function $f(x) = \sum_{k=0}^{\infty} c_k x^k$, we can now obtain feature maps ϕ and $\psi : \mathbb{R}^d \to L^2$ such that $\langle \phi(u), \psi(v)\rangle = f(\langle u, v\rangle)$. Moreover, $\|\phi(u)\|_2^2 = \|psi(v)\|_2^2 = \sum_{k=0}^{\infty} |c_k|$. For the sine function,

$$f(x) = \sin(cx) = cx - \frac{(cx)^3}{3!} + \frac{(cx)^5}{5!} \dots$$
$$\implies \|\phi\| = \sum_{k=0}^{\infty} = c + \frac{c^3}{3!} + \frac{c^5}{5!} + \cdots = (e^c - e^{-c})/2 := 1, \quad \text{we want it to map to unit vector}$$
$$\implies c = \ln(1 + \sqrt{2}) := \frac{\beta\pi}{2}$$
$$\implies \beta = \frac{2}{\pi}\ln(1 + \sqrt{2})$$

3.1 Support Vector Machines (SVM) and Kernel Trick

In machine learning, we use the Support Vector Machine (SVM) to perform binary classification which uses this "kernel trick" described above. Let's say, you have a dataset (x_i, y_i) with $x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$. We want to find a linear separation using $w \in \mathbb{R}^d$ such that

$$\langle w, x_i\rangle \begin{cases} > 1 & \text{if } y_i = 1 \\ < (-1) & \text{if } y_i = (-1) \end{cases}$$

i.e., $\langle w, x_i\rangle y_i > 1$ for all $i = 1, 2, \dots n$.

However, having a linear separation is overly optimistic, in practice we may not have that. Instead, we can find a feature map $\phi : \mathbb{R}^d \to H$ (some Hilbert space) such that on H, there is a linear separation. Therefore, we can find w^* such that $\langle w^*, \phi(x_i)\rangle y_i > 1$ for all i. If we have n datapoints, then if we have H to be n-dimensional, we must have a linear separation. Thus, we can consider the points x_js as basis, so that $w = \sum_j \alpha_j x_j$. With the feature map transformation, we then would have $w = \sum_j \alpha_j \phi(x_j)$. Therefore, the final objective becomes

$$\text{Find } \alpha_j \text{ such that } \sum_{j=1}^{n} y_j \alpha_j \langle \phi(x_j), \phi(x_i)\rangle > 1, \ \forall i.$$

Therefore, to solve this problem, one simply requires the knowledge of the inner product in the feature space, but not necessarily the feature map H. Hence, we can pick any positive definition function K (called the kernel) and assume that

$$\langle \phi(x_1), \phi(x_2)\rangle := K(x_1, x_2).$$

This algorithm is called **Hard-margin SVM**.

Note that, hard-margin SVM perfectly classifies the binary classes, and allows no outliers. This means, even if one or two points misclassify, SVM can produce a feature map which is extremely complex. In order to allow small amount of outliers, we can combine the objectives along with a hinge-loss function.

Definition 11 (Hinge Loss)**.** *The hinge loss function is defined as* $(1-t)_+ = \max\{1-t, 0\}$.

The **Soft-margin SVM** considers the objective function to minimize

$$l(w) = \sum_{i=1}^{n} (1 - \langle w, x_i \rangle y_i)_+ + \lambda \|w\|_2^2$$

where λ is a penalization parameter.

Theorem 20 (Mercer)**.** *For any continuous, symmetric and positive semi-definite kernel function* K*, there exists a map* $\phi : \mathbb{R}^d \to H$ *such that* $K(x, y) = \langle \phi(x), \phi(y) \rangle$.

4 Spectral Analysis

4.1 Principal Component Analysis

To visualize high-dimensional data, we use PCA (Principal Component Analysis). Let $X \in \mathbb{R}^d$, $\mathbb{E}(X) = 0$ and $\mathbb{E}(XX^\intercal) = \Sigma$. Then we can perform eigen-decomposition of Σ to obtain the eigenvalues $\lambda_i(\Sigma)$ and the eigenvectors $v_i(\Sigma)$, which are called the "principal components" of X.

However, we do not know Σ, we can only estimate it using a sample covariance matrix in the light of data. Given a finite sample $X_1, \dots X_n \in \mathbb{R}^d$, we have the estimated sample covariance matrix

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\intercal,$$

assume there are mean-centered. We hope that, the PCA applied on Σ_n will be close to PCA applied on Σ. There are two questions:

1. How many samples do we require to establish this approximation?
2. What is the error rate?

The answer to the first question turns out to be surprisingly only $O(d)$. That means, you need only linear many samples as the dimension to get a full picture. For the second question, we need to study some basic tools of random matrix theory.

Let's start by considering the answer to the first question. The first goal will be to show that $\Sigma_n \approx \Sigma$ in operator norm.

Let S^{d-1} be the d-dimensional unit sphere in $\mathbb{R}^d$. Then,

$$\|\Sigma_n - \Sigma\|_2^2 = \max_{v \in S^{d-1}} |v^\intercal (\Sigma_n - \Sigma) v| = \max_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 - \mathbb{E} \langle X, v \rangle^2 \right|.$$

Let us call this quantity as $Z(v) = \frac{1}{n}\sum_{i=1}^{n} \langle X_i, v\rangle^2 - \mathbb{E}\langle X, v\rangle^2$. Note that, $\{Z(v) : v \in S^{d-1}\}$ is a random process indexed by $v \in S^{d-1}$. For a fixed v, we can show a bound by using concentration inequality. Since the maximum is taken over an uncountable set, we cannot use union bound to proceed. We need to reduce it to an ϵ-net type idea.

Lemma 8. *For all $\epsilon > 0$, the sphere S^{d-1} has an ϵ-net $\{x_1, x_2, \dots x_N\}$, such that $N \leq (2/\epsilon + 1)^d$.*

Proof. Consider the minimal ϵ-net. The $\epsilon/2$-ball centered at the x_is must be disjoint. If not, there exists $i \neq j$ such that $\|x_i - x_j\| < \epsilon$, because of triangle inequality. This means, the ϵ-net is not minimal since we can remove x_j.

Note that, all balls lie in the $(1 + \epsilon/2)$-ball. This means,

$$\begin{aligned} &\text{Vol}(B(1+\epsilon/2)) \geq N\text{Vol}(B(\epsilon/2)) \\ \implies &N \leq \frac{(1+\epsilon/2)^d}{(\epsilon/2)^d} = (2/\epsilon + 1)^d \end{aligned}$$

□

The next step in the plan is to approximate the operator norm (i.e., the maximum over S^{d-1}) with the maximum over an ϵ-net.

Lemma 9. *For a real symmetric matrix A, let $\mathcal{N} \subseteq S^{d-1}$ be an ϵ-net of S^{d-1}, then $\|A\| \leq 1/(1-\epsilon)\max_{x\in\mathcal{N}} \|Ax\|$.*

Proof. The definition of the operator norm says that there exists $u \in S^{d-1}$ such that $\|Au\| = \|A\|$, since we can take u to the eigenvector corresponding to the first eigenvalue of A (which will have real coordinates as A is real symmetric). By definition of the ϵ-net, there exists $x \in \mathcal{N}$ such that $\|x - u\| \leq \epsilon$, hence, by Cauchy-Schwartz inequality,

$$\|Ax - Au\| = \|A(x-u)\| \leq \|A\|\,\|x-u\| \leq \|A\|\,\epsilon.$$

Now, it follows that by triangle inequality,

$$\|Ax\| = \|Au - (Au - Ax)\| \geq \|Au\| - \|Ax - Au\| \geq (1-\epsilon)\,\|A\|\,.$$

Taking maximum over all $x \in \mathcal{N}$ now completes the proof. □

As a corollary, if we choose $\epsilon = 1/4$, then our previous results imply that $|\mathcal{N}| \leq 9^d$. Also, we get that

$$\|\Sigma_n - \Sigma\| \leq 2\max_{v\in\mathcal{N}} \left| \frac{1}{n}\sum_{i=1}^{n} \langle X_i, v\rangle^2 - \mathbb{E}\langle X, v\rangle^2 \right|.$$

Now, if we assume that X_is are independent and subGaussian, then $\langle X_i, v\rangle$ is simply a linear combination of the coordinates of X_i, hence is also subGaussian. As a result, $\langle X_i, v\rangle^2$ are sub-exponential. Let $K = \left\|\langle X_i, v\rangle^2\right\|_{\psi_1} = \|X_i^\intercal X_i\|_{\psi_1}$ For a fixed $v \in \mathcal{N} \subseteq S^{d-1}$, using Berstein's inequality,

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} \langle X_i, v\rangle^2 - \mathbb{E}\langle X, v\rangle^2 \right| > \delta \right) = \mathbb{P}\left(\left| \sum_{i=1}^{n} \langle X_i, v\rangle^2 - \mathbb{E}\langle X, v\rangle^2 \right| > \delta n \right)$$

$$\leq 2\exp\left[-c\min\left(\frac{\delta^2 n^2}{K^2 n}, \frac{\delta n}{K}\right)\right]$$
$$= 2\exp\left[-c\min\left(\frac{\delta^2}{K^2}, \frac{\delta}{K}\right) n\right]$$

Theorem 21. *If X_i are independent subGaussian random variables and $n \geq Cd$ for some absolute constant C, then we have $\|\Sigma_n - \Sigma\| \leq 0.1\,\|\Sigma\|$, with probability at least $1 - 2e^{-cn}$.*

Proof. Without loss of generality, assume that $\|\Sigma\| = 1$, otherwise, we can rescale. For a fixed $v \in S^{d-1}$, since $\|\Sigma\| = 1$, we have

$$\mathbb{E}\langle X, v\rangle^2 \leq \|\Sigma\| = 1 \implies \|X_i\|_{\psi_1} \leq C,$$

for some absolute constant C. Therefore, by using Berstein's inequality as shown above,

$$\mathbb{P}\left(|Z(v)| > \delta = 0.01\right) \leq 2e^{-cn}, \quad \text{for some c.}$$

By using union bound, we now have

$$\mathbb{P}\left(\max_{v\in\mathcal{N}} |Z(v)| > 0.01\right) \leq |\mathcal{N}|\, 2e^{-cn} = 9^d \times 2e^{-cn} \leq 2e^{-c'n/2},$$

if $d < C"n$. This means, the complement event holds with probability $1 - 2^{-c'n/2}$, and when it holds, we have

$$\|\Sigma_n - \Sigma\| \leq 2\max v \in \mathcal{N}|Z(v)| \leq 2 \times 0.01 \leq 0.1,$$

which completes the proof. □

Now that we have the "closeness" of Σ_n and Σ, we want to show that the eigenvalues and the eigenvectors also remain approximately close.

Theorem 22 (Weyl's inequality)**.** *For all $d \times d$ symmetric matrices A and B, $\max_i |\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|$.*

Since we have $\|\Sigma_n - \Sigma\|$ to be small, this shows that the eigenvalues of the sample covariance matrix will closely resemble the eigenvalues of the true covariance matrix.

In the case of eigenvectors, there are some difficulties:

1. Two eigenvectors may be the same up to the same sign.
2. For a given eigenvalue with more than one geometric multiplicity, there are multiple eigenvectors that can only have a rotational invariance.
3. Suppose two eigenvalues are close to each other, but the perturbation switches their order. Then, the corresponding eigenvectors also switch. This means, these swapped eigenvectors will be orthogonal to each other, not at all close.

Therefore, we need to have something called a "spectral gap", i.e., all the eigenvalues are sufficiently apart.

Theorem 23 (Davis Kahan inequality)**.** *Let A, B be two $d \times d$ symmetric matrices. Let P_A be the projection operator onto the span of $\{v_1(A), \ldots, v_k(A)\}$ for some $k = 1, 2, \ldots, d$. Similarly defined P_B to be the projection operator onto the span of eigenvectors of B. Then,*

$$\|P_A - P_B\| \leq \frac{\|A - B\|}{\lambda_k(A) - \lambda_{k+1}(A)}.$$

This entire analysis now shows why the method of principal component analysis works. A few important remarks are mentioned below.

Note

1. Linear sample size $n = O(d)$ is the optimal choice in general.
2. For structured data, it may be possible to do principal component analysis with a smaller sample size. For example, if $X = (X_1, \ldots, X_1)$, then X is essentially one-dimensional and then $n = O(1)$.
3. If the data is inherently k-dimensional, then $n = O(k)$ is usually enough. This general result is proved by Koltchinskii and Lounici [2017].

4.2 Bulk Spectram Analysis

Figure 2: Spectral distribution of real-life data

When we analyse real-life data, we often have the above kind of picture if we plot the spectrum of the covariance matrix under consideration. Bulk of the spectrum will be distributed across a range, which consists of the noises, and then there are a few spectral values that stands out as outliers, these are only the meaningful components of the signal. Therefore, to identify the signal, we need to understand how the spectrum of pure noise would look like, as then we can remove that and get the proper signal out of the data.

The spectrum behaviour of the pure noise data is given by "Marchenko Pastur Law (1967)". We shall talk about this in a short while.

4.2.1 Wigner's Law

We begin by considering a simpler model instead. Let W be a symmetric $n \times n$ random matrix such that all of its entries on and above the diagonal are i.i.d. standard normal random variables. Then Wigner's law demonstrates the asymptotic behaviour of the spectral density of $W/\sqrt{n}$.

Theorem 24 (Wigner's Law (1958))**.** *For the above random matrix W, for any $-2 \leq a \leq b \leq 2$, the number of eigenvalues of $W/\sqrt{n}$ lying between $[a,b]$, say denoted by $N_{W/\sqrt{n}}([a,b])$ satisfy*

$$\lim_{n\to\infty} \frac{N_{W/\sqrt{n}}([a,b])}{n} = \int_a^b \rho_{sc}(x)dx,$$

where $\rho_{sc}(x) = \sqrt{4-x^2}/2\pi$ where $x \in [-2,2]$. This is often called the "semicircle" density.

Before we aim to prove this, we need a few tools at our disposal.

1. If we have a partition matrix
$$M = \begin{bmatrix} a & b^\intercal \\ b & D \end{bmatrix},$$
then if D is invertible, then the formula for Schur's complement formula says
$$M_{11}^{-1} = (1,1)\text{-th entry of } M^{-1} = \frac{1}{a - b^\intercal D^{-1} b}.$$

2. If $g \sim N(0, I_n)$, then $\mathbb{E}(g^\intercal M g) = \text{Trace}\,(M)$.

3. For a symmetric matrix M, we have
$$\lambda_i(M^{-1}) = \frac{1}{\lambda_i(M)}, \text{ and, } \lambda_i(M - zI) = \lambda_i(M) - z, \ z \in \mathbb{R}.$$

4. One function that contains information about eigenvalues is $f(z) = \det(A - zI)$, for which the roots of this function are the eigenvalues.

5. Consider another function where the eigenvalues are the poles of the function. Hence,
$$f(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i - z} = \frac{1}{n}\sum_{i=1}^n \lambda_i((A-zI)^{-1}) = \frac{1}{n}\text{Trace}\left((A-zI)^{-1}\right).$$
This last matrix $(A - zI)^{-1}$ is often called "resolvent function" of the matrix A.

6. On a similar note, one can define **Stieljes transform**. Stieljes transform of a distribution of a random variable X is
$$S(z) = \mathbb{E}\left[\frac{1}{X - z}\right], z \in \mathbb{C}.$$
It is possible to show that $S(z)$ can uniquely define the distribution of X (much like the characteristic function). In fact, its inversion formula is given by
$$f(x) = \lim_{\epsilon \to 0+} \frac{S(x + i\epsilon) - S(x - i\epsilon)}{2\pi i}.$$

The following is an informal proof of Wigner's law based on these ideas.

Proof. **Step 1:** Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $W/\sqrt{n}$. Define the Stieljes transform of the empirical spectral density,

$$S_n(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i - z} = \frac{1}{n}\text{Trace}\left(\left(\frac{1}{\sqrt{n}}W - zI\right)^{-1}\right) = \frac{1}{n}\sum_{i=1}^{n} M_{ii}^{-1},$$

where M is equal to the matrix $(W/\sqrt{n} - zI)$. Note that, M can be partitioned as

$$M = \begin{bmatrix} \frac{W_{11}}{\sqrt{n}} - z & \frac{1}{\sqrt{n}}g^\intercal \\ \frac{1}{\sqrt{n}}g & \bar{M} \end{bmatrix},$$

where $g \sim N(0, I_{n-1})$. Therefore, by the formula of Schur's complement, we have

$$\begin{aligned} M_{11}^{-1} &= \frac{1}{\frac{W_{11}}{\sqrt{n}} - z - \frac{1}{n}g^\intercal \bar{M}^{-1} g} \\ &\approx \frac{1}{0 - z - \frac{1}{n}\mathbb{E}(g^\intercal \bar{M}^{-1} g)} \\ &= \frac{1}{-z - \frac{1}{n}\text{Trace}\left((\bar{M})^{-1}\right)} \\ &\approx \frac{1}{-z - \frac{n}{n-1}S_{n-1}(z)}, \quad \text{assuming } \bar{M} \approx M, \text{ but just of dimension } (n-1) \end{aligned}$$

As $n \to \infty$, we have in the limit,

$$S_\infty(z) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1} M_{ii}^{-1} = -\frac{1}{z + S_\infty(z)},$$

i.e., $S_\infty(z) + (z + S_\infty(z))^{-1} = 0$. Solving for $S_\infty(z)$ yields

$$S_\infty(z) = \frac{-z + \sqrt{z^2 - 4}}{2}.$$

Step 2: Due to the Stieljes inversion, it is now enough to show that the semicircle distribution $\rho_{sc}(x)$ has the same Stieljes transformation.

To see this, we note that

$$\begin{aligned} S_{sc}(z) &= \frac{1}{2\pi}\int_{-2}^{2}\frac{\sqrt{4 - x^2}}{x - z}dx \\ &= \frac{1}{2\pi}\int_{\pi}^{0}\frac{-4\sin^2(\theta)}{2\cos(\theta) - z}d\theta, \quad \text{letting } x = \cos(\theta) \\ &= \frac{1}{\pi}\int_{0}^{\pi}\frac{2\sin^2(\theta)}{2\cos(\theta) - z}d\theta \\ &= -\frac{1}{4i\pi}\int_{\|\alpha\|=1}\frac{(\alpha^2 - 1)^2}{\alpha^2(\alpha^2 + 1 - \alpha z)}d\alpha, \quad \text{letting } \alpha = e^{i\theta} \end{aligned}$$

Now in the last integral, we need to apply Cauchy's Residue formula. Note that, the given function has 3 poles, at $\alpha_0 = 0$ of order 2, at $\alpha_1 = (z + \sqrt{z^2-4})/2$ and at $\alpha_2 = (z - \sqrt{z^2-4})/2$.

The pole at $\alpha_1 = (z+\sqrt{z^2-4})/2$ lies outside the unit circle, hence to use the Residue theorem, we calculate the residues at α_0 and α_2.

$$\text{Res}(\alpha_0) = \frac{4\alpha_0(\alpha_0^2-1)(\alpha_0^2+1-z\alpha_0) - (\alpha_0^2-1)^2(2\alpha_0 - z)}{(\alpha_0^2+1-z\alpha_0)^2} = z$$
$$\text{Res}(\alpha_2) = \frac{(\alpha_2^2-1)^2}{\alpha_2^2(\alpha_2 - (z+\sqrt{z^2-4})/2)} = -\sqrt{z^2-4}$$

Therefore, by applying the Residue theorem for integrals,

$$S_{sc}(z) = -\frac{1}{4\pi i}\int_{\|\alpha\|=1} \frac{(\alpha^2-1)^2}{\alpha^2(\alpha^2+1-\alpha z)} = -\frac{2\pi i}{4\pi i}\left[\text{Res}(\alpha_0) + \text{Res}(\alpha_2)\right] = \frac{-z+\sqrt{z^2-4}}{2},$$

which matches with the Stieljes transform $S_\infty(z)$ derived earlier.
Step 3: Now we apply the Steiljes inverse to complete the proof. □

We have a similar theorem for the singular values which may arise from a matrix with different dimensions.

Theorem 25 (Marchenko-Pastur, 1967)**.** *Let $X_1, X_2, \ldots, X_n \sim N(0, I_d)$ and $\Sigma_n = n^{-1}\sum_{i=1}^n X_i X_i^\intercal$, the sample covariance matrix. Suppose that $d/n \to r \in (0,1)$ as $n \to \infty$. Then, the spectral density of Σ_n converges to the Marchenko Pastur law, given by*

$$\rho_{MP}(x) = \frac{1}{2\pi r x}\sqrt{(x-a)(b-x)},\ x \in [a,b],$$

where $a = (1-\sqrt{r})^2$ and $b = (1+\sqrt{r})^2$.

To prove this result, we need to know the following two results.

1. **Stieljes transformation of MP law:** Using similar calculation as above, we can calculate the Stieljes transformation of the Marchenko Pastur law as

$$S_{MP}(z) = \frac{1-z-r+\sqrt{(1-z-r)^2-4zr}}{2}.$$

2. **Sherman-Morrison Formula:** Given two vectors $x, y \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$, we have

$$(M+xy^\intercal)^{-1} = M^{-1} - \frac{M^{-1}xy^\intercal M^{-1}}{1+y^\intercal M^{-1}x}.$$

Let $q = y^\intercal M^{-1}x$. Then, we get that

$$y^\intercal(M+xy^\intercal)^{-1}x = q - \frac{q^2}{1+q} = 1 - \frac{1}{1+q} = 1 - \frac{1}{y^\intercal M^{-1}x}.$$

The following is again an informal proof of Marchenko-Pastur theorem using these two basic ideas.

Proof. Our steps will be similar to the Wigner's theorem. We shall derive the limiting Stieljes transform of the spectral density of Σ_n and show that it matches with the Stieljes transform of MP law. Let $S_n(z)$ be the Stieljes transform of the spectral density of Σ_n, i.e.,

$$S_n(z) = \frac{1}{d}\sum_{k=1}^{d} \lambda_k(\Sigma_n - zI)^{-1} = \frac{n}{d}\text{Trace}\left(\left(\sum_{i=1}^{n} X_i X_i^\intercal - nzI\right)^{-1}\right) \to \frac{1}{r}\text{Trace}\left(A^{-1}\right),$$

where A is the matrix it is replacing, i.e., $A = \sum_{i=1}^{n} X_i X_i^\intercal - nzI$. Let $B = \sum_{i=1}^{n-1} X_i X_i^\intercal - nzI = A - X_n X_n^\intercal$. Now, it follows that

$$\begin{aligned}
& X_n^\intercal A^{-1} X_n = X_n^\intercal (B + X_n X_n^\intercal)^{-1} X_n = 1 - \frac{1}{1 + X_n^\intercal B^{-1} X_n} \\
\implies & X_n^\intercal A^{-1} X_n \approx 1 - \frac{1}{1 + \mathbb{E}(X_n^\intercal B^{-1} X_n)} \\
\implies & X_n^\intercal A^{-1} X_n \approx 1 - \frac{1}{1 + \text{Trace}\,(B^{-1})}, \quad \text{since } X_n \text{ are } B \text{ independent} \\
\implies & X_n^\intercal A^{-1} X_n \approx 1 - \frac{1}{1 + rS_{n-1}(nz/(n-1))} \\
\implies & \sum_{k=1}^{n} X_k^\intercal A^{-1} X_k \approx n\left[1 - \frac{1}{1 + rS_{n-1}(nz/(n-1))}\right] \\
\implies & \text{Trace}\left(\sum_{k=1}^{n} X_k X_k^\intercal A^{-1}\right) \approx n\left[1 - \frac{1}{1 + rS_{n-1}(nz/(n-1))}\right] \\
\implies & \text{Trace}\left((A + nzI)A^{-1}\right) \approx n\left[1 - \frac{1}{1 + rS_{n-1}(nz/(n-1))}\right] \\
\implies & d + nzrS_n(z) = d(1 + zS_n(z)) \approx n\left[1 - \frac{1}{1 + rS_{n-1}(nz/(n-1))}\right]
\end{aligned}$$

Taking limit on both sides and using $S_\infty(z)$ to denote the limiting Stieljes transform, we obtain the relation

$$1 + zS_\infty(z) = \frac{1}{r}\left[1 - \frac{1}{1 + rS_\infty(z)}\right] = \frac{S_\infty(z)}{1 + rS_\infty(z)}.$$

Solving this yields,

$$S_\infty(z) = \frac{1 - z - r + \sqrt{(1 - z - r)^2 - 4zr}}{2}.$$

□

References

Koltchinskii, Vladimir, and Karim Lounici. "Concentration inequalities and moment bounds for sample covariance operators." Bernoulli (2017): 110-133.