Notes on High Dimension Probability

Subhrajyoty Roy

September 6, 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 **1**
1.1 Convexity 2
1.2 Applications of ACT 4

2 Concentration Inequalities **8**
2.1 Inequalities 11
2.2 Applications 13
2.2.1 Mean estimation 13
2.2.2 Random Graph Phase Transition 14
2.2.3 Discrepancy Theory 16
2.3 Spaces of Random Variables 18
2.3.1 Sub Gaussian Random Variables 21

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

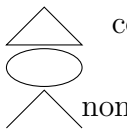
Thinking in terms of probability might help to overcome in high-dimensional inference problems.

1.1 Convexity

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

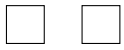
Definition 1. *A set $T \subset \mathbb{R}^n$ is convex if $\forall x, y \in T$, the segment $[x, y] \in T$.*

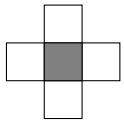
Let us consider a few examples as follows:


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

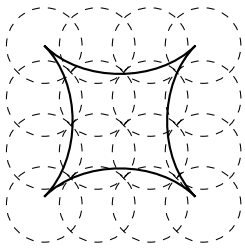
There are multiple applications of ACT. One useful idea is in portfolio building, where the ingredients are "stocks", and their linear combination is basically a mutual fund. The problem is to create a new mutual fund with a given combination of stocks by mixing available mutual funds, (i.e., the fund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

$$\begin{aligned} Vol(B) &\leq N \cdot Vol\left(\frac{1}{2}B\right), \text{ since RHS might have some overlaps.} \\ \implies Vol(B) &\leq N2^{-d}Vol(B) \\ \implies N &\geq 2^d \end{aligned}$$

□

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. $\square$

To make use of these results, usually, it helps by considering the intuition that, a small covering # $\Rightarrow$ small volume, $\Rightarrow$ 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

$$\begin{aligned} \inf_{\epsilon > 0} \ell(\epsilon) &= \exp\left[n \log\left(\frac{\sqrt{4 \log m}}{\sqrt{n}} \right) + \frac{2 \log m}{4 \log m} n \right] \\ &= \exp\left[\log\left(\frac{\sqrt{4 \log m}}{\sqrt{n}} \right)^n e^{n/2} \right] \end{aligned}$$

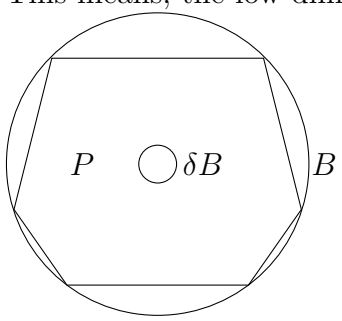
$$= \left(\frac{\sqrt{4e\log m}}{\sqrt{n}}\right)^n \leq \left(4\frac{\sqrt{\log m}}{\sqrt{n}}\right)^n, \text{ since } e \leq 4.$$

$\square$

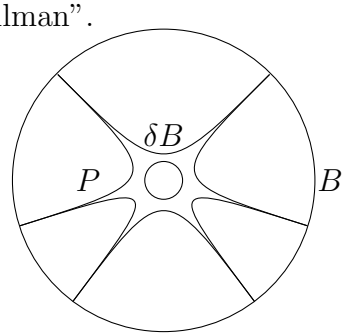
A few remarks we want to make here.

- 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* $\operatorname{Var}\langle X, Y\rangle^2$*.*

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

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,

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

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 $X_1, X_2, \ldots X_N$ 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. $\square$

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

Theorem 8. *Let $X_1, X_2, \ldots, X_n$ 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 $X_1, X_2, \ldots X_N \sim \text{Ber}(p)$, 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,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) &= e^{-\mu} \sum_{k \geq t} \frac{\mu^k}{k!} \quad \left(\text{Stirling's bounds: } k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k\right) \\
&\sim e^{-\mu} \mu^t \left(\frac{e}{t}\right)^t \quad (\text{only dominating term is } t) \\
&= e^{-\mu}\left(\frac{\mu e}{t}\right)^t = O((C/t)^{-t})
\end{aligned}$$

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,

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

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. $\square$

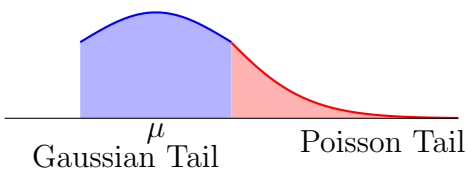
Although the Chernoff bound predicts a heavier tail compared to the Gaussian tail (i.e., $O(e^{-t\log(t)}$ instead of $O(e^{-t^2})$), it makes sense only when $t >> \mu$. When $t \approx \mu$, i.e., for small deviations, we can get Gaussian approximations.

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

By Taylor series, we have

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

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 $\mathrm{Var}(X_i) = \sigma^2$. 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 $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$. 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 $B_1, B_2, \dots B_K$. Let, $\hat{\mu}_j = \dfrac{1}{m}\sum_{i \in B_j} X_i$. Define, $\tilde{\mu}$ to be the median of these blockwise means, i.e., median of $\hat{\mu}_1, \dots \hat{\mu}_K$.

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 $\deg(i) = d_i = \#$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.

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,*

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

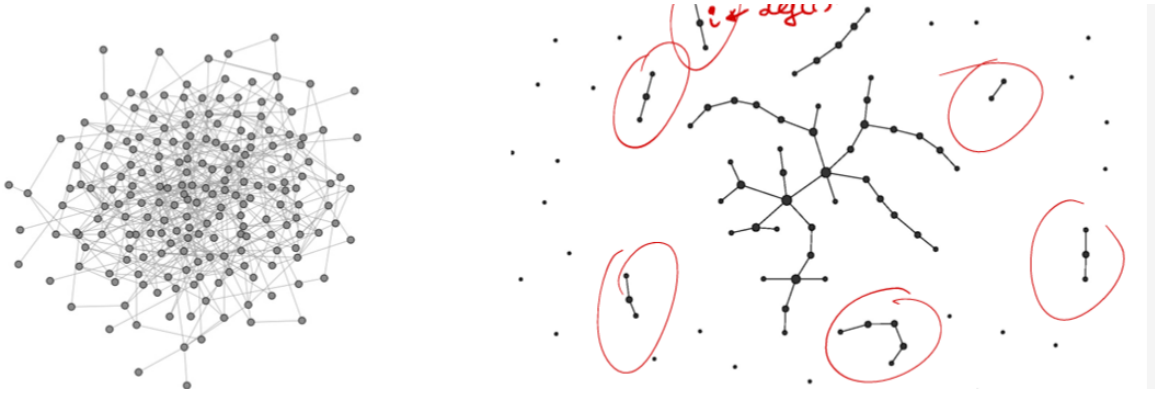


Figure 1: Phase transitions of Erdos-Renyi random graph

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*

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

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

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

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., $|N_I - \lambda_I N| \geq C\sqrt{\lambda_I N \log(N)}$. Now,

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

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,

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

$$\leq \lambda_J N + C\sqrt{\lambda_J N \log(N)} + \left(\frac{4}{N^9} + C\frac{\sqrt{4\log(N)}}{N^9}\right), \text{ since } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$$

$$\leq \lambda_J N + 2C\sqrt{\lambda_J N \log(N)}, \text{ since the second term is smaller than } \sqrt{\lambda_J N \log(N)}$$

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,

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

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

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

   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

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

   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

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

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. □

CHECK FROM HERE

Proof: (1) $\Rightarrow$ (2), WLOG, assume $k_1 = 1$,

Then $\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(|X|^p > t)dt$ (change of integral lemma)

$= \int_0^\infty \mathbb{P}(|X| > s)ps^{p-1}ds$ $(t = s^p \Rightarrow dt = ps^{p-1}ds)$

$\leq \int_0^\infty 2e^{-s^2}ps^{p-1}ds = p(\frac{p}{2} - 1)!$ (Gamma integral)

$\leq pp^{p/2-1} = p^{p/2}$

$\Rightarrow \|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq \sqrt{p}$

(2) $\Rightarrow$ (3) WLOG, assume $k_2 = 1$,

$\mathbb{E}(e^{X^2/100}) = \mathbb{E}(\sum_{p=0}^\infty \frac{(X^2/100)^p}{p!})$ (Taylor series)

$= \sum_{p=0}^\infty \frac{1}{p!} \cdot \frac{1}{(100)^p} \cdot \mathbb{E}(X^{2p}) \leq \sum_{p=0}^\infty \frac{1}{(100)^p p!} \cdot (\sqrt{2p})^{2p}$

By Stirling's approx., $p! \geq e^{-p}p^p$

$\Rightarrow \mathbb{E}(e^{X^2/100}) \leq \sum_{p=0}^\infty \frac{1}{(100)^p p!} \cdot (\sqrt{2p})^{2p} = \sum_{p=0}^\infty (\frac{2e}{100})^p \leq 2.$

$\frac{1}{50}$

(3) $\Rightarrow$ (4) WLOG, $k_3 = 1$, use $e^x \leq x + e^{x^2}$, $\forall x \in \mathbb{R}$

Therefore, $\mathbb{E}e^{\lambda X} \leq \mathbb{E}(\lambda X) + \mathbb{E}(e^{\lambda^2 X^2})$ $\downarrow$ $= 0$ $\rightarrow \leq$ $(\mathbb{E}e^{X^2})^{\lambda^2}$, if $|\lambda| < 1$ $\leq 2^{\lambda^2} = e^{\lambda^2 \log 2}$

(4) $\Rightarrow$ (1) WLOG, $k_4 = 1$

$\mathbb{P}(|X| > t) = \mathbb{P}(e^{\lambda X} > e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}(e^{\lambda X}) \leq e^{-\lambda t} \cdot e^{\lambda^2/2} = \exp(\lambda^2/2 - \lambda t)$ Markov) minimizing over λ yields $\leq e^{-t^2/4}$. $t/2)$

• Since, $\|X\|_{\psi_2} \leftarrow$ Sub-Gaussian norm we have for orlicz function $(e^{x^2} - 1)$. So, in the Sub-Gaussian lemma, we can take all k_i's as some constant times $\|X\|_{\psi_2}$.

Homework: (Classical form of Chernoff bound) Let, $S \sim Bin(N, \mu/N)$. Then $\mathbb{P}(S > t) < e^{-t}$, $\forall t > e\mu$

Proof: We know from our version, $\mathbb{P}(S > t) \leq e^{-\mu}(\frac{e\mu}{t})^t$, $\forall t > \mu$.

Since, $t > e\mu \Rightarrow \log t > 2 + \log \mu$ $\Rightarrow t \log t > 2t + t \log \mu > 2t + t \log \mu - \mu$ $\Rightarrow -t > -\mu + t \log \mu + t - t \log t$ $\Rightarrow e^{-t} > e^{-\mu} \cdot \mu^t \cdot e^t \cdot t^{-t} \geq \mathbb{P}(S > t)$.

Proposition: If X_i's are indep, mean zero, sub-Gaussian, then $\|\sum_{i=1}^N X_i\|_{\psi_2} \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2$

Proof: $\mathbb{E}e^{\lambda \sum X_i} = \prod_{i=1}^N \mathbb{E}e^{\lambda X_i} \leq \prod_{i=1}^N e^{ck_i^2\lambda^2}$ $\uparrow$ $k_i^2 = \|X_i\|_{\psi_2}^2$ (independence) $\leq e^{c(\sum_{i=1}^N k_i^2)\lambda^2}$

by equivalence of subgaussian properties, $\sum X_i$ is Sub-Gaussian, & $\|\sum X_i\|_{\psi_2} \leq C \sum k_i$

Renormalizing this & using equivalence to tail behaviour, we get

Thm: (Sub-Gaussian Hoeffding lemma) If X_i are indep, mean zero, sub-gaussian, then

$\mathbb{P}(|\sum_{i=1}^N X_i| > t) \leq 2\exp(-ct^2/\sigma^2)$, where, $\sigma^2 = \sum_{i=1}^N \|X_i\|_{\psi_2}^2$ is kind of "variance proxy"!

Turns out, sub-gaussian class is not large enough. If $X \sim$ sub-gaussian, then, $X^2 \not\sim$ sub-gaussian.

$\mathbb{P}(X^2 > t) = \mathbb{P}(|X| > \sqrt{t}) \sim e^{-(\sqrt{t})^2} = e^{-t} \gg e^{-\sqrt{t}}$, $\forall t > 0$ he ier than

Sub-exponential and Thin Shell Phenomenon

Sub-exponential lemma: $\forall$ r.v. X, the following are equivalent:

(1) $\exists k_1$, $\mathbb{P}(|X| > t) \leq 2e^{-t/k_1}$, $\forall t > 0$

(2) $\exists k_2$ s.t. $\|X\|_p \leq k_2 p$, $\forall p \geq 1$

(3) $\exists k_3$ s.t. $\mathbb{E}\exp(|X|/k_3) \leq 2$ $\longrightarrow$ yields us, $k_3 = \|X\|_{\psi_1}$ (sub-exponential norm)

(4) $\exists k_4$ s.t. $\mathbb{E}e^{\lambda X} \leq e^{\lambda^2 k_4^2}$, $\forall |\lambda| \leq \frac{1}{k_4}$

(The last one follows as $\int$ in the MGF, $\mathbb{E}e^{\lambda X} = \int e^{\lambda x}\mathbb{P}(dx)$ $\sim e^{-cx} \Rightarrow$ This is similar to the Chernoff case, near the tails, it follows Gaussian behaviour), (so if $\lambda > c$, the integral even if the tail is heavy). won't converge)

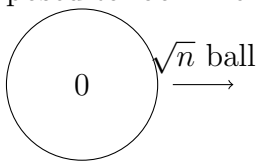
Does Hoeffding's bound hold for sub-exponential X_i? $\longrightarrow$ No! If $N = 1$, we cannot have, Surprisingly, once you accumulate multiple indep X_i's $\mathbb{P}(|X| > t) \leq e^{-ct^2}$ they produce Gaussian tail:

Bernstein's inequality: If X_i's are indep, mean 0, sub-exponential, then,

$\mathbb{P}(|\sum_{i=1}^n X_i| > t) \leq 2\exp[-c\cdot\min(\frac{t^2}{\sigma^2}, \frac{t}{K})]$, (Proof is same MGF where, $\sigma^2 = \sum_{i=1}^n \|X_i\|_{\psi_1}^2$, $K = \max_i \|X_i\|_{\psi_1}$ method) $\uparrow$ notice ψ_1

Thin Shell Phenomenon:

Consider a Gaussian random vector $g \sim N(0, I_n)$ So, the samples from $N(0, I_n)$ is supposed to look like $\boxed{0 \in \mathbb{R}^n}$ but this $\longrightarrow$ is not correct, instead we have...



centrate Thm: For $g \sim N(0, I_n)$ $\mathbb{P}(\sqrt{n} - 0.99 \leq \|g\|_2 \leq 1.01\sqrt{n}) \geq 1 - 2e^{-cn}$ The points in a thin shell near $\sqrt{n}$-ball

Proof: Consider, $\|g\|_2^2 - n = \sum_{i=1}^n (g_i^2 - 1)$ $g_i =$ coordinate of g $= \sum_{i=1}^n (g_i^2 - \mathbb{E}g_i^2) \longrightarrow$ sum of iid terms

g_i^2 is sub-exponential

So, we apply Bernstein's ineq., $\mathbb{P}(|\|g\|_2^2 - n| \geq 0.01n) \leq 2e^{-c\min(\frac{n^2}{\sigma^2}, \frac{n}{K})}$

$\sigma^2 = \sum_{i=1}^n \|g_i^2 - 1\|_{\psi_1} \leq nC$, and, $K = \max_i \|g_i^2 - 1\|_{\psi_1} = \|g_1^2 - 1\|_{\psi_1} \leq C$ Since iid

$\uparrow$

Therefore, $\mathbb{P}(0.99n \leq \|g\|_2^2 \leq 1.01n) \geq 1 - 2e^{-c\min(\frac{n^2}{Cn}, \frac{n}{C})} \geq 1 - 2e^{-cn}$ $\underbrace{\text{take square root now}}$ & get another constant

Dimension Reduction

Let $x_1, x_2, \ldots, x_N \in \mathbb{R}^n$ be given map to $\mathbb{R}^m$, $m \ll n$ where we preserve the pairwise distances (approximately) Is it possible with $m = O(\log N)$?

Thm (Johnson - Lindenstrauss lemma, 1984) $\forall x_1, \ldots, x_N \in \mathbb{R}^d$, $\exists$ a linear map T: $T : \mathbb{R}^d \to \mathbb{R}^m$, such that, $m < C\log N$, and, $0.99\|x_i - x_j\|_2 \leq \|T(x_i) - T(x_j)\|_2 \leq 1.01\|x_i - x_j\|_2, \forall i, j = 1, 2, \ldots, N$

Proof: Choose a random map, say, matrix corresponding to T is, G_i, with $G_{ij} \overset{iid}{\sim} N(0,1)$

$m\left[G\right]_d = \left[I\right]_m$ Fix any $z \in \mathbb{R}^d$, s.t. $\|z\|_2 = 1$.

Then, $Gz \sim N(0, I_m)$, as, $(Gz)_i = \sum_{j=1}^d G_{ij} z_j$ $N(0, \sum_{j=1}^d z_j^2) = N(0,1)$

By thin shell theorem: $\mathbb{P}(0.99\sqrt{m} \leq \|Gz\|_2 \leq 1.01\sqrt{m}) \geq 1 - 2e^{-cm}$ (1)

Now for any i, j, using (1), we have, $\mathbb{P}(0.99\sqrt{m} \leq \|G(\frac{x_i - x_j}{\|x_i - x_j\|_2})\|_2 \leq 1.01\sqrt{m}) \geq 1 - 2e^{-cm}$

$\Rightarrow \mathbb{P}(0.99\sqrt{m}\|x_i - x_j\|_2 \leq \|T(x_i) - T(x_j)\|_2 \leq 1.01\sqrt{m}\|x_i - x_j\|_2) \geq 1 - 2e^{-cm}$

$\Rightarrow$ By union $\forall i, j$ pairs, bound, where, $T = \frac{1}{\sqrt{m}} G$.

$\mathbb{P}(\forall i, j, 0.99\|x_i - x_j\|_2 \leq \|T(x_i) - T(x_j)\|_2 \leq 1.01\|x_i - x_j\|_2) \geq 1 - 2N^2 e^{-cm}$ $1 - 2\exp(-cm + 2\log N)$

Choose $cm = 4\log N$, so to get, $\geq 1 - 2\exp(-2\log N) = 1 - \frac{2}{N^2} > 0$, if $N \geq 2$

$\Rightarrow m = \frac{4}{c}\log N = O(\log N)$.

$\Rightarrow \exists$ a map for which this holds deterministically.

Obs: $\Rightarrow$ The std $\sigma(X)$ captures only "2nd moment" & average spread, while Sub-Gaussian norm captures all moments, MGF, etc. This exercise shows Sub-Gauss. norm dominates $\sigma(X)$ but not other way around.

$\Rightarrow$ Let, X be sub-gaussian, & assume $\mathbb{E}X = 0$. WLOG. $\rightarrow$ absolute Then, $\sigma(X) = (\mathbb{E}X^2)^{1/2} = \|X\|_2 \leq \sqrt{2}K_2\|X\|_{\psi_2} \leq C\|X\|_{\psi_2}$

$\Rightarrow$ Let, $X \sim N(\mu, \sigma^2)$, then, $\sigma(X) = \sigma$, and, let $\|X\|_{\psi_2} = K$ (*)

Then, $\mathbb{E}(e^{X^2/K^2}) \leq 2 \Rightarrow \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{x^2/K^2} \cdot e^{-x^2/2\sigma^2} dx < 2 \Rightarrow \frac{\sigma^2}{K^2} < \frac{1}{2\sigma^2}$ $K^2 > 2\sigma^2$

$\Rightarrow K = \sqrt{2}\sigma$ works.

$\Rightarrow \|X\|_{\psi_2} \leq \sqrt{2}\sigma = 4$ $\boxed{\|X\|_{\psi_2} \leq C\sigma(X)}$ $\rightarrow$ in this case reverse $\uparrow$ (*) inequality holds

$\Rightarrow$ But consider, $X \sim Poisson(\lambda)$ $\|X\|_{\psi_2} = \infty$, as it is not sub-gaussian So, we do not have (*). But, $Var(X) = \lambda \Rightarrow \sigma(X) = \sqrt{\lambda} < \infty$,