High Dimension Probability

2021-04-24

Contents

Chapter 1

Preliminaries on random variables

Lemma 1.1. *Let X be a non-negative r.v. then*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\left\{X > t\right\} dt$$

Proof. Note that for Lebesgue measure, there exists:

$$x = \mu[0, x) = \mu(0, x) = \mu[0, x] = \int_0^\infty {}_{\{t<x\}} dt$$

Then for some certain $\omega \in \Omega$,

$$X(\omega) = \int_0^\infty {}_{t<X(\omega)} dt$$

Then

$$\mathbb{E}X = \int_\Omega X d\mathbb{P} = \int_\Omega \int_0^\infty {}_{\{t<X\}} dt d\mathbb{P}$$

According to Fubini's Theorem,

$$\mathbb{E}X = \int_0^\infty dt \int_\Omega {}_{\{t<X\}} d\mathbb{P} = \int_0^\infty \mathbb{P}\left\{X > t\right\} dt$$

□

Exercise 1.1 (Generalization of integral identity)**.** Show that for any *r.v.* X

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\{X > t\} dt - \int_{-\infty}^0 \mathbb{P}\{X < t\} dt$$

Solution. Note $X = X^+ - X^-$, then

$$\begin{aligned}
\mathbb{E}\, X &= \mathbb{E}\, X^+ - \mathbb{E}\, X^- \\
&= \int_0^\infty \mathbb{P}\{X^+ > t\}dt - \int_0^\infty \mathbb{P}\{X^- > t\}dt \\
&= \int_0^\infty \mathbb{P}\{X > t\}dt - \int_0^\infty \mathbb{P}\{-X > t\}dt \\
&= \int_0^\infty \mathbb{P}\{X > t\}dt - \int_0^\infty \mathbb{P}\{X < -t\}dt \\
&= \int_0^\infty \mathbb{P}\{X > t\}dt - \int_{-\infty}^0 \mathbb{P}\{X < t\}dt
\end{aligned}$$

Exercise 1.2 (p-moments via tails)**.** Let X be a *r.v.* and $p \in (0, \infty)$, show that

$$\mathbb{E}\,|X|^p = \int_0^\infty pt^{p-1}\,\mathbb{P}\{|X| > t\}dt$$

Solution. By the integral identity

$$\begin{aligned}
\mathbb{E}\,|X|^p &= \int_0^\infty \mathbb{P}\{|X|^p > t\}dt \\
&= \int_0^\infty \mathbb{P}\{|X| > t^{\frac{1}{p}}\}dt \\
&= \int_0^\infty \mathbb{P}\{|X| > t^{\frac{1}{p}}\}\frac{pd(t^{\frac{1}{p}})}{t^{\frac{1}{p}-1}} \\
&= \int_0^\infty \mathbb{P}\{|X| > t\}\frac{pdt}{t^{1-p}} \\
&= \int_0^\infty pt^{p-1}\,\mathbb{P}\{|X| > t\}dt
\end{aligned}$$

Exercise 1.3 (Chebyshev's inequality)**.** Let X be a *r.v.* with mean μ and variance σ^2, then for any $t > 0$, we have

$$\mathbb{P}\{|X - \mu| \geq t\} = \mathbb{P}\{|X - \mu|^2 \geq t^2\} = \frac{\sigma^2}{t^2}$$

Theorem 1.1 (DeMoivre-Laplace Theorem)**.** *Let $(X_i)_{i\in\mathbb{N}^*}$ are i.i.d. Bernoulli variables with mean $\mu = p$ and variance $\sigma^2 = p(1-p)$, then*

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathfrak{Z}$$

Theorem 1.2 (Lindeberg-Levy Theorem)**.** *Let $(X_i)_{i\in\mathbb{N}^*}$ be i.i.d. with mean μ and variance σ^2, both finite, then*

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathfrak{Z}$$

Proof. In view of theorem **??**, the claim is

$$\varphi_{Z_n}(t) \to \varphi_{\mathfrak{Z}}(t) = e^{-t^2/2}$$

Let φ denote the $ch.f.$ of $\frac{X_n-\mu}{\sigma}$, then Taylor's theorem yields

$$\begin{aligned}\varphi(t) &= \varphi(0) + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2(1+h(t)) \\ &= 1 - \frac{1}{2}t^2(1+h(t))\end{aligned}$$

for some h $s.t.$ $\lim_{t\to\infty}|h(t)| = 0$. As (X_n) are independent, note $Z_n = \sum \frac{X_n-\mu}{\sigma}/\sqrt{n}$:

$$\varphi_{Z_n}(t) = \varphi^n(\frac{t}{\sqrt{n}}) = [1 - \frac{r^2/2}{n}(1+h(\frac{r}{\sqrt{n}}))]^n$$

note $\frac{r}{\sqrt{n}} \to 0$ as $n \to \infty$, thus

$$[1 - \frac{r^2/2}{n}(1+h(\frac{r}{\sqrt{n}}))]^n \to (1 - \frac{r^2/2}{n})^n \to e^{-r^2/2}$$

and claim follows.

□

Exercise 1.4. Let $(X_i)_{i\in\mathbb{N}^*}$ be $i.i.d.$ $r.v.'s$ with mean μ and finite variance, then

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}X_i - \mu\right| = O(\frac{1}{\sqrt{N}})$$

Solution. Suppose the variance is $\sigma^2 < \infty$, then

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}X_i - \mu\right| &= \lim_{N\to\infty}\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)\right| \\ &= \mathbb{E}\lim_{N\to\infty}\left|\frac{S_n - n\mu}{N}\right| \\ &= \frac{\sigma}{\sqrt{N}}\mathbb{E}\lim_{N\to\infty}\left|\frac{S_n - n\mu}{\sigma\sqrt{N}}\right| \\ &= \frac{\sigma}{\sqrt{N}}\mathbb{E}\,|\mathfrak{Z}| = \frac{\sigma\sqrt{2}}{\sqrt{N\pi}} = O(\frac{1}{\sqrt{N}})\end{aligned}$$

Chapter 2

Concentration of sums of independent random variables

2.1 Why concentration inequalities?

Concentration inequalities quantify how a random variable X deviates around its mean, the simplest is Chebyshev's inequality 1.3. It can be applied to general $r.v.$ but often too weak.

Example 2.1. Suppose $(X_i)_{i\in\mathbb{N}^*}$ is $i.i.d.$ and distributed as $\mathrm{Ber}(\frac{1}{2})$, let $S_n = \sum_{i=1}^n X_i$, then $\mathbb{E}\, S_n = \frac{n}{2}$ and $\mathrm{Var}\, S_n = \frac{n}{4}$, consider $\mathbb{P}\{S_n > \frac{3}{4}n\}$.

Let S_n denote the number of heads, then Chebyshev's applies to

$$\mathbb{P}\{S_n \geq \frac{3}{4}n\} \leq \frac{4}{n}$$

To see whether it's appropriate, by CLT, we have

$$Z_n = \frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \xrightarrow{d} \mathfrak{Z}$$

so we have

$$\mathbb{P}\left\{S_n \geq \frac{3}{4}n\right\} \approx \mathbb{P}\left\{\mathfrak{Z} \geq \sqrt{\frac{n}{4}}\right\}$$

Proposition 2.1 (tails of normal distribution)**.** *Let $g \sim \mathcal{N}(0,1)$, then for $t > 0$, we have:*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(g \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Proof. For the upper bound:

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

the lower bound follows from

$$\int_t^\infty (1 - 3x^{-4}) e^{-\frac{x^2}{2}} dx = \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-\frac{t^2}{2}}$$

This completes the proof.

□

Thus $\mathbb{P}\{S_n \geq \frac{3}{4}n\} \leq \frac{1}{\sqrt{2\pi}} e^{-\frac{n}{8}}$, which is much better in view of figure 2.1

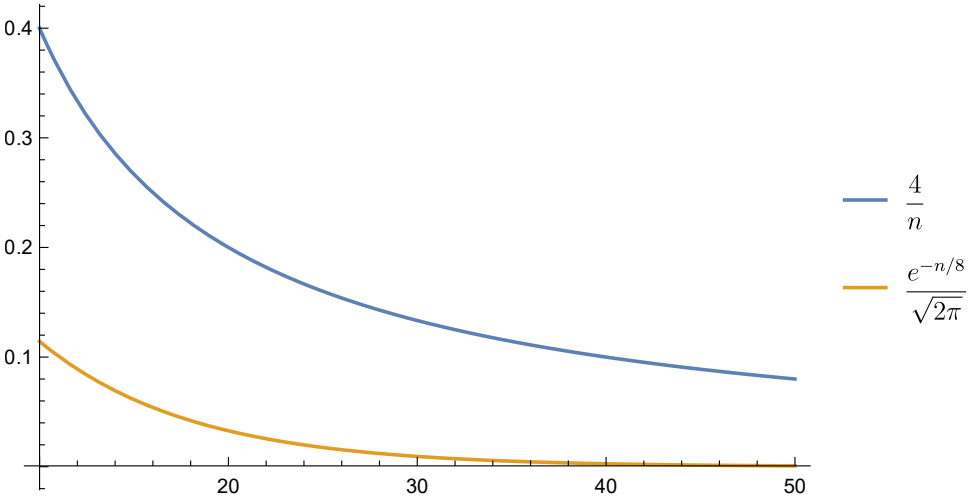

Figure 2.1: linerly and exponential decay rate

However, the way we give the exponential decay rate is not rigorous, in fact, we can't do so since the error term of approximation given by CLT decays too slow:

Theorem 2.1 (Berry-Esseen CLT)**.** *In the setting of theorem 1.2, we have*

$$|\mathbb{P}\{Z_n \geq t\} - \mathbb{P}\{\mathfrak{Z} \geq t\}| \leq \frac{\mathbb{E}|X - \mu|^3}{\sigma^3}$$

In order to resolve this issue, we develop alternative, direct approaches to concentration, which bypass the central limit theorem.

Proposition 2.2 (Truncated normal distribution)**.** *Show that for all $t \geq 1$, we have*

$$\mathbb{E}\,\mathfrak{Z}^2\mathbf{1}_{\{\mathfrak{Z}>t\}} = t \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} + \mathbb{P}\{\mathfrak{Z} > t\} \leq \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Proof. The first inequality follows from directly integrity. The second follows from:

$$\begin{aligned}
\mathbb{P}\{\mathfrak{Z} > t\} &= \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{-x^2}{2}} dx \\
&= \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{-x^2}{2}} dx \\
&\leq \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{-x^2}{2}} \frac{x}{t} dx \\
&= \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}
\end{aligned}$$

□

2.2 Hoeffding's inequality

Definition 2.1. Symmetric Bernoulli distribution takes value -1 and 1 with probability $\frac{1}{2}$ each.

Theorem 2.2 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_N$ be independent symmetric Bernoulli r.v. and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ then for any $t \geq 0$ we have:*

$$\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Proof. WLOG, suppose that $\|a\|_2 = 1$.

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} &= \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^N a_i X_i\right) \geq \exp(\lambda t)\right\} \\
&\leq e^{-\lambda t} \mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right)
\end{aligned}$$

Now consider $\mathbb{E}\exp\left(\lambda \sum_{i=1}^N a_i X_i\right)$, from the independency, we find

$$\mathbb{E}\exp\left(\lambda \sum_{i=1}^N a_i X_i\right) = \prod_{i=1}^N \mathbb{E}\exp(\lambda a_i X_i)$$

Since the property of symmetric Bernoulli distribution, notice that for some fixed i,

$$\mathbb{E}\exp(\lambda a_i X_i) = \frac{\exp(\lambda a_i) + \exp(-\lambda a_i)}{2} = \cosh(\lambda a_i)$$

Exercise 2.1. Show that

$$\cosh(x) \leq \exp(\frac{x^2}{2}) \text{ for all } x \in \mathbb{R}$$

It follows that

$$\mathbb{E}\exp(\lambda a_i X_i) \leq \exp\left(\frac{\lambda^2 a_i^2}{2}\right)$$

and hence

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} \leq e^{-\lambda t}\prod_{i=1}^N \exp(\lambda^2 a_i^2) &= \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^N a_i^2\right) \\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2}\right)\end{aligned}$$

Now pick $\lambda = t$, $f(\lambda) = -\lambda t + \frac{\lambda^2}{2}$ get the minimum, then

$$\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} \leq \exp\left(\frac{t^2}{2}\right)$$

□

Now return to 2.1, take the tail *r.v.* to $X_i' = 2X_i - 1$, then Hoeffding's inequality applies:

$$\mathbb{P}\left\{S_n > \frac{3}{4}n\right\} = \mathbb{P}\left\{S_n' > \frac{1}{2}n\right\} \leq \exp\left(\frac{-n}{8}\right)$$

Note we can part $\mathbb{P}\{|S| \geq t\}$ as $\mathbb{P}\{S \geq t\} + \mathbb{P}\{-S \geq t\}$ and thus

Theorem 2.3 (Hoeffding's inequality, two sided)**.** *Let $X_1, \ldots, X_N$ be independent symmetric Bernoulli r.v. and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ then for any $t \geq 0$ we have:*

$$\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Proposition 2.3 (Chernoff bounds)**.** *Let X_i by independent, then for any $\lambda > 0$,*

$$\mathbb{P}\{X_i \geq t\} \leq e^{-\lambda t} M_{X_i}(t)$$

and if $S_n = \sum_{i=1}^n X_i$, we have

$$\mathbb{P}\{S_n \geq t\} \leq e^{-\lambda t}\prod_{i=1}^n M_{X_i}(t)$$

Similarly, we have

$$\mathbb{P}\{S_n \leq t\} \leq e^{\lambda t} M_{X_i}(-t)$$

Proof. Note $X_i \geq t \iff e^{\lambda X_i} \geq e^{\lambda t}$, then claim follows by Markov's inequality.

□

The following extension of Hoeffding's inequality is valid for general bounded random variables:

Theorem 2.4 (Hoeffding's inequality for general bounded random variables)**.** *Let $X_1, \dots, X_N$ be independent with bounded $X_i \in [m_i, M_i]$, then for any $t > 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{N}(X_i - \mathbb{E}\,X_i) \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

Proof.

Lemma 2.1 (Hoeffding's lemma)**.** *Let X be r.v. bounded by $X \in [m, M]$ a.s. with mean μ, then*

$$M_{X-\mu}(\lambda) \leq \exp\left(\frac{\lambda^2(M-m)^2}{8}\right)$$

Now we apply Hoeffding's lemma. The Chernoff bound tell us:

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right\} &\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mu_i)\right)\right]\exp(-\lambda t) \\
&= \exp(-\lambda t)\prod_{i=1}^{n}\mathbb{E}\exp\left[\lambda(X_i - \mu_i)\right] \\
&\leq \exp(-\lambda t)\prod_{i=1}^{n}\exp\left(\frac{\lambda^2(M_i - m_i)^2}{8}\right) \\
&= \exp\left[\frac{\lambda^2}{8}\sum_{i=1}^{n}(M_i - m_i)^2 - \lambda t\right]
\end{aligned}$$

To minimize over $\lambda \geq 0$, we have

$$\lambda = \frac{4t}{\sum_{i=1}^{n}(M_i - m_i)^2}$$

then substituting it complete the proof.

□

Exercise 2.2 (Boosting randomized algorithms)**.** Suppose $(X_i)_{i\in\mathbb{N}^*}$ are *i.i.d.* with $\mathrm{Ber}(\frac{1}{2} - \delta)$, where $\delta > 0$. Show that

$$\mathbb{P}\left\{S_n \geq \frac{1}{2}n\right\} \leq \epsilon$$

whenever $n \geq \frac{1}{2\delta^2}\ln\left(\frac{1}{\epsilon}\right)$.

Solution. By the Hoeffding's inequality 2.4:

$$\mathbb{P}\left\{S_n \geq \frac{1}{2}n\right\} = \mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \frac{1}{2} + \delta) \geq n\delta\right\} \leq \exp(-\frac{2n^2\delta^2}{n})$$

to ensure $\exp\left(2n\delta^2\right) \leq \epsilon$, we must have $n \geq \frac{1}{2\delta^2}\ln\left(\frac{1}{\epsilon}\right)$.

Exercise 2.3 (Robust estimation of the mean)**.** Suppose we want to estimate μ from sample $X_1, \dots, X_n$ with ϵ-accurate, *i.e.*, as small $\mathbb{P}\left\{|\hat{\mu} - \mu| > \epsilon\right\}$ as possible.

1. Show that a sample of size $n = O(\frac{\sigma^2}{\epsilon^2})$ is sufficient to compute an ϵ-accurate estimate with probability at least $\frac{3}{4}$, where $\sigma^2 = \operatorname{Var} X$.
2. Show that a sample of size $n = O(\log(\delta^{-1})\frac{\sigma^2}{\epsilon^2})$ is sufficient to compute an ϵ-accurate estimate with probability at least $1 - \delta$.

Solution. 1. Let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$, then we see that

$$\mathbb{P}\left\{|\hat{\mu} - \mu| > \epsilon\right\} \leq \frac{\mathbb{E}(\hat{\mu} - \mu)^2}{\epsilon^2} = \frac{\frac{1}{n^2}\mathbb{E}\sum_{i=1}^n (X_i - \mu)^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

2. Suppose $(\hat{\mu}_i)_{i=1}^m$ are drawn independent as in part **1**, let $Y_i = \mathbf{1}_{\hat{\mu}_i - \mu > \epsilon}$, then Y_i are Bernoulli with $p_i \leq \frac{1}{4}$. Denote

$$\hat{\mu} = \text{Median of } (\hat{\mu}_i)_{i=1}^m$$

then by general Hoeffding's inequality 2.4:

$$\begin{aligned}\mathbb{P}\left\{\hat{\mu} - \mu > \epsilon\right\} &= \mathbb{P}\left\{\sum_{i=1}^m Y_i \geq \frac{m}{2}\right\} \\ &\leq \mathbb{P}\left\{\sum_{i=1}^m (Y_i - p_i) \geq \frac{m}{4}\right\} \\ &\leq \exp\left(-\frac{m}{8}\right)\end{aligned}$$

Similarly, we have $\mathbb{P}\left\{\hat{\mu} - \mu < -\epsilon\right\} \leq \exp\left(-\frac{m}{8}\right)$ and hence

$$\mathbb{P}\left\{|\hat{\mu} - \mu| \geq \epsilon\right\} \leq 2\exp\left(-\frac{m}{8}\right)$$

To ensure that smaller than δ, we should have $m \geq 8\ln(2\delta^{-1})$ and claim follows by observe that we use $m \times n$ samples.

Exercise 2.4 (Small ball probabilities)**.** Let $X_1, \ldots, X_n$ be non-negative independent $r.v.'s$ with continuous distribution. Assume that the densities of X_i are bounded by 1.

1. Show that for all $t > 0$, the MGF of X_i satisfies

$$\mathbb{E}\exp(-tX_i) \leq \frac{1}{t}$$

2. For any $\epsilon > 0$, show that

$$\mathbb{P}\left\{\sum_{i=1}^n X_i \leq \epsilon n\right\} \leq (e\epsilon)^n$$

Solution. 1. Note

$$\mathbb{E}\left[\exp(-tX_i)\right] = \int_0^\infty f_i(x)e^{-tx}dx \leq \int_0^\infty e^{-tx} = \frac{1}{t}$$

2. For any ϵ, note

$$\sum_{i=1}^{n} X_i \leq \epsilon n \iff \exp\left[-\lambda \sum_{i=1}^{n} \frac{X_i}{\epsilon}\right] \geq \exp(-\lambda n)$$

then by Markov's inequality:

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \leq \epsilon n\right\} \leq e^{\lambda n} \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(-\frac{\lambda}{\epsilon} X_i\right)\right] \leq e^{\lambda n} \left(\frac{\epsilon}{\lambda}\right)^n$$

minimize over $\lambda > 0$ and we find $\lambda = 1$ give the desired results.

2.3 Chernoff's inequality

The general Hoeffding's inequality are not so sharp in case $(X_i)_{i\in\mathbb{N}^*}$ are Bernoulli with parameter p_i small enough to apply Poisson Limit Theorem: suppose $\mathbb{E}\, S_n = \sum p_i = \mu$, then by Hoeffding's inequality 2.4:

$$\mathbb{P}\left\{S_n - \mu \geq t\right\} \leq \exp(-\frac{2t^2}{n})$$

which is not so sharp as Chernoff's inequality giving below, since it's insensitive to the magnitude of p_i (and thus, μ). Figure 2.2 shows their difference.

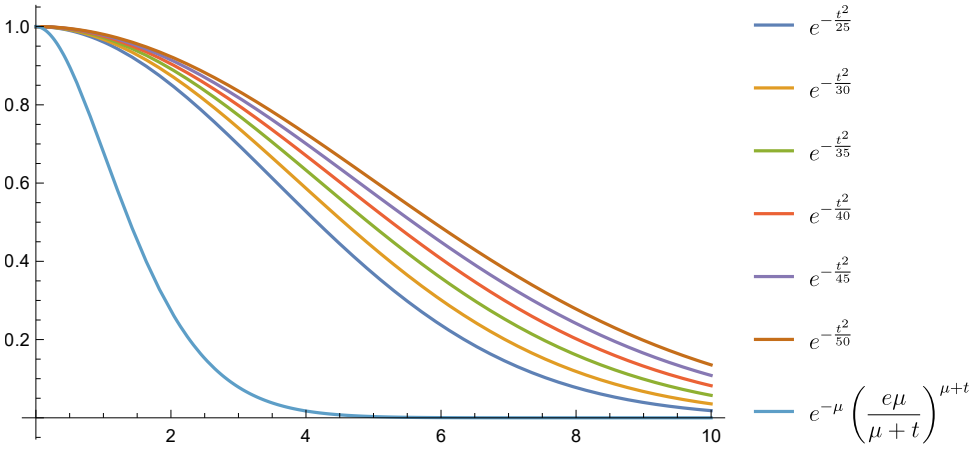

Figure 2.2: behavior when $\mu = 1$

Theorem 2.5 (chernoff inequality)**.** *Let X_i be independent Bernoulli r.v. with parameter p_i. Consider their sum $S_n = \sum_{i=1}^{n} X_i$ and denote $\mu = \mathbb{E}S_n$ then for $t > \mu$ we have:*

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

Proof. By Chernoff bound and independency of X_i:

$$\mathbb{P}\left\{S_n \geq t\right\} \leq e^{-\lambda t} M_{S_n}(t) = e^{-\lambda t} \prod_{i=1}^{n} M_{X_i}(\lambda)$$

Note the numeric inequality $1 + x \leq e^x$:

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

Then

$$\prod_{i=1}^N \mathbb{E}\exp(\lambda X_i) \leq \exp\left[(e^\lambda - 1)\sum_{i=1}^N p_i\right] = \exp[(e^\lambda - 1)\mu]$$

Finally, pick $\lambda = \ln(t/\mu)$, we complete the proof.

□

Proposition 2.4 (lower tail)**.** *For any $t < \mu$ we have*

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

Proof. Note:

$$\begin{aligned}\mathbb{P}\left\{S_N \leq t\right\} = \mathbb{P}\left\{-S_N \geq -t\right\} &\leq \frac{\mathbb{E}\exp(-\lambda S_N)}{\exp(-\lambda t)} \\ &= e^{\lambda t}\cdot\prod_{i=1}^N \mathbb{E}e^{-\lambda X_i} \leq e^{\lambda t}\exp\left[(e^{-\lambda} - 1)\mu\right]\end{aligned}$$

Then replace $-\lambda = \ln(t/\mu)$.

□

Proposition 2.5 (Poisson tail)**.** *Let $X \sim Pois(\lambda)$, then for any $t > \mu$, we have*

$$\mathbb{P}\left\{X \geq t\right\} \leq e^{-\lambda}\left(\frac{e\lambda}{t}\right)^t$$

Proof. According to the Poisson limit theorem, if $X \sim \text{Pois}(\lambda)$, we can construct a series of r.v. where $X_i \sim \text{Ber}(p_i)$ s.t. $X = \sum_{i=1}^\infty X_i$ and note that $X_i \geq 0$ for every i so we can use MCT.

$$\begin{aligned}\mathbb{P}\left\{X \geq t\right\} &\leq \exp(-\lambda t)\cdot\prod_{i=1}^\infty \mathbb{E}\exp(\lambda X_i) \\ &\leq \exp\left(\lambda(e^t - 1)\right)\end{aligned}$$

It follows some similar step and we done.

□

Proposition 2.6 (Chernoff's inequality: small deviations)**.** *In the setting of theorem 2.5, for $\delta \in (0, 1]$, we have*

$$\mathbb{P}\left\{|S_n - \mu| \geq \delta\mu\right\} \leq 2\exp\left(-c\mu\delta^2\right)$$

where c is an absolute constant.

Proof. We only show the positive side, by chernoff inequality

$$\mathbb{P}\left\{S_n - \mu \geq \delta\mu\right\} \leq e^{-\lambda(\delta+1)\mu}\,\mathbb{E}\exp\left(\lambda S_n\right) \leq e^{\delta\mu}\left(\frac{1}{\delta+1}\right)^{(\delta+1)\mu}$$

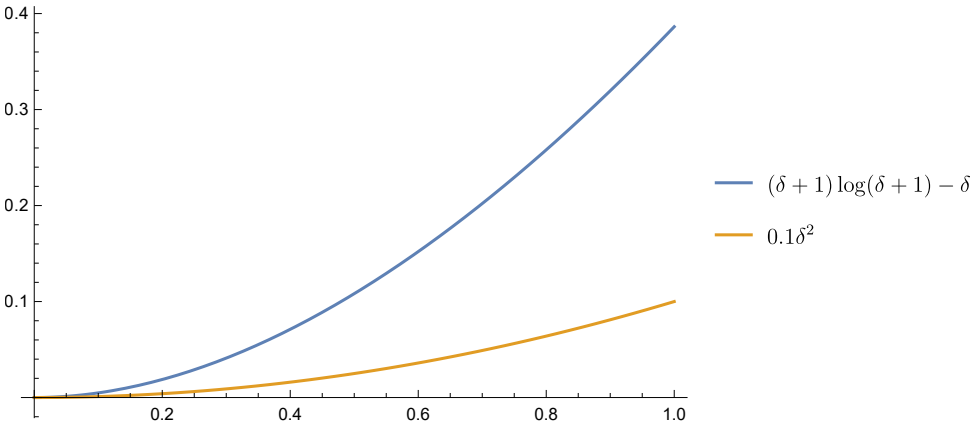

Figure 2.3: small deviations

Taking log each side, then we show that

$$-\delta + (\delta+1)\ln(\delta+1) \geq c\delta^2$$

holds for some c. It's possible by noting

□

As consequence, by Poisson Limit Theorem, we have when t is small, Poisson is almost sub-gaussian(see later):

Corollary 2.1. *Let $X \sim Pois(\lambda)$, for $t \in (0, \lambda]$, we have*

$$\mathbb{P}\{|X-\mu| \geq t\} \leq 2\exp\left(-\frac{ct^2}{\lambda}\right)$$

2.4 Application: degrees of random graphs

Definition 2.2. Erd¨os-R´enyi model $G(n,p)$ is a random graph with n vertices which connecting each pair with probability p. Where p can be constant or function depending n.

Clearly, the expected degree of each vertices in $G(n,p)$ is clearly $(n-1)p =: d$. The relatively **dense graphs**, which is defined by $d \gg \log n$, are almost regular with high probability:

::: {.proposition dense-regular name="Dense graphs are almost regular"}

For a random graph $G \sim G(n,p)$ with expected degree $d \geq C\log n$. Then, it's regular with a high probability.

:::

Proof. For a fixed vertex i, its degree d_i is a sum of $n-1$ independent Bernoulli variable, by Chernoff's inequality 2.6, we have

$$\mathbb{P}\{|d_i - d| \geq \delta d\} \leq 2e^{-cd}$$

The probability that G is not almost regular is

$$\mathbb{P}\{\exists i : |d_i - d| \geq \delta\} \leq n2e^{-cd} \leq 2n^{1-cC}$$

can be arbitrary small.

□

2.5 Sub-gaussian distributions

Now suppose which kind of $r.v.$ have similar results with Hoeffding's inequality, suppose $n = 1$, we expect

$$\mathbb{P}\{|X_i| > t\} \leq 2e^{-ct^2}$$

By intuition, we expect Gaussian distribution is sub-gaussian. In fact, we have

$$\mathbb{P}\{|\mathfrak{Z}| > t\} \leq 2e^{-\frac{t^2}{2}}$$

that is sub-gaussian tails.

Proof. It's sufficient to show that $\mathbb{P}\{\mathfrak{Z} > t\} \leq e^{-\frac{t^2}{2}}$. Note

$$\mathbb{P}\{\mathfrak{Z} > t\} = \int_t^\infty \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx, \quad e^{-\frac{t^2}{2}} = \int_t^\infty x e^{-\frac{x^2}{2}} dx$$

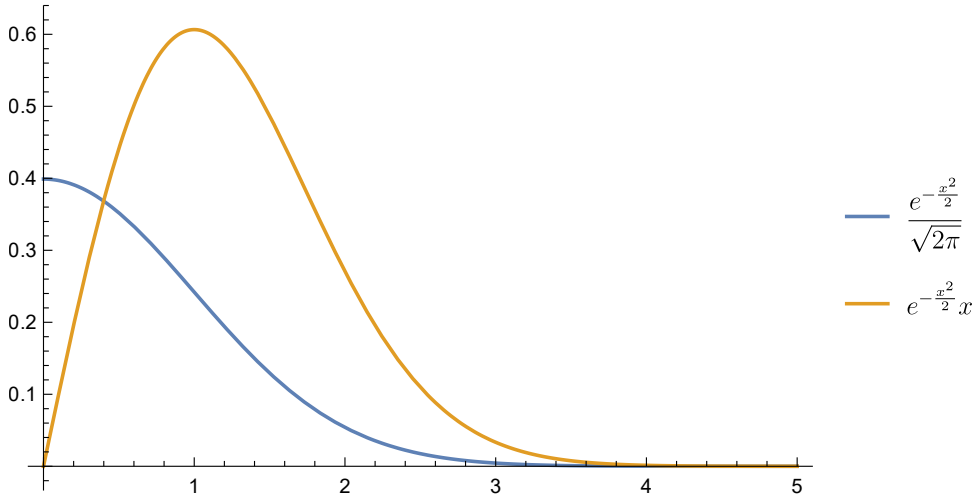

Figure 2.4: density of normal distribution

Clearly, we have $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \leq x e^{-\frac{x^2}{2}}$ when $x \geq 1$. Then we claim that

$$\int_0^1 \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx \leq \int_0^1 x e^{-\frac{x^2}{2}} dx$$

This do hold and complete the proof.

□

Exercise 2.5 (Moments of normal distribution)**.** Show that for $p \geq 1$:

$$\|\mathfrak{Z}\| = (\mathbb{E}\,|\mathfrak{Z}|^p)^{\frac{1}{p}} = \sqrt{2}\left[\frac{\Gamma(\frac{1+p}{2})}{\Gamma(\frac{1}{2})}\right]^{\frac{1}{p}} = O(\sqrt{p})$$

Solution. Note

$$\begin{aligned}\|\mathfrak{Z}\|_p &= \left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} |x|^p\, e^{-\frac{x^2}{2}}\, dx\right)^{\frac{1}{p}} = \left(\frac{2}{\sqrt{2\pi}}\int_{\mathbb{R}} x^p e^{-\frac{x^2}{2}}\, dx\right)^{\frac{1}{p}} \\ &= \left(\frac{2}{\sqrt{2\pi}}\int_{\mathbb{R}} \sqrt{2\omega}^p e^{-\omega}\frac{1}{\sqrt{2\omega}} d\omega\right)^{\frac{1}{p}} \text{ change variable } \frac{x^2}{2} = \omega \\ &= \left(\frac{2}{\sqrt{2\pi}}\sqrt{2}^{p-1}\Gamma(\frac{p+1}{2})\right)^{\frac{1}{p}} = \sqrt{2}\left[\frac{\Gamma(\frac{1+p}{2})}{\Gamma(\frac{1}{2})}\right]^{\frac{1}{p}}\end{aligned}$$

Hence $\|\mathfrak{Z}\|^p = O[(\frac{p}{2}!)^{\frac{1}{p}}] = O(\frac{p}{2}^{\frac{p}{2}\cdot\frac{1}{p}}) = O(\sqrt{p})$.

Finally, we have $M_{\mathfrak{Z}}(\lambda) = e^{\frac{\lambda^2}{2}}$.

2.5.1 Sub-gaussian properties

Now let X be a general random variable. The following proposition states that the properties we just considered are equivalent.

Proposition 2.7. *Let X be a r.v., TFAE:*

1. *X has the tail satisfy*
$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-t^2/K_1^2) \text{ for all } t > 0$$
2. *The moment of X satisfy*
$$\|X\|_{L^p} = (\mathbb{E}\,|X|^p)^{1/p} \leq K_2\sqrt{p} \text{ for all } p \geq 1$$
3. *The MGF of X^2 :*
$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(K_3^2\lambda^2) \text{ for all } \lambda \text{ s.t. } |\lambda| \leq \frac{1}{K_3}$$
4. *The MGF of X^2 is bounded at some point:*
$$\mathbb{E}\exp(X^2/K_4^2) \leq 2$$

   Moreover, if $\mathbb{E}\,X = 0$, then above properties equivalent to:

5. *The MGF of X satisfy*
$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2)$$

Where each K_i differ from each other by at most an absolute constant factor.

Proof. 1 $\implies$ 2: WLOG, suppose $K_1 = 1$,

$$\begin{aligned}
\mathbb{E}\,|X|^p &= \int_0^\infty \mathbb{P}\left\{|X|^p > t\right\} dt \\
&= \int_0^\infty \mathbb{P}\left\{|X| > u\right\} pu^{p-1} du \\
&\leq \int_0^\infty 2\exp(-u^2) pu^{p-1} du \\
&= p\Gamma\left(\frac{p}{2}\right) \leq 3p\left(\frac{p}{2}\right)^{p/2}
\end{aligned}$$

where the last inequality follows from $\Gamma(x) \leq 3x^x$ for all $x \geq \frac{1}{2}$ and $p \geq 1$. Taking pth root, we have

$$K_2 = \sup_{p\geq 1}\left[(3p)^{\frac{1}{p}}\left(\frac{1}{2}\right)^{\frac{1}{2}}\right] = \frac{e^{\frac{3}{e}}}{\sqrt{2}} \leq 3$$

2 $\implies$ 3 : Again, assume $K_2 = 1$. By Taylor expansion, we obtain

$$\mathbb{E}\exp(\lambda^2 X^2) = \mathbb{E}\left[1 + \sum_{p=1}^\infty \frac{(\lambda^2 X^2)^p}{p!}\right] = 1 + \sum_{p=1}^\infty \frac{\lambda^{2p}\mathbb{E}\left[X^{2p}\right]}{p!}$$

Note that 2 implies that $\mathbb{E}[X^{2p}] \leq (2p)^p$. Then Stirling's approximation $p! \geq (p/e)^p$ yields:

$$\mathbb{E}\exp(\lambda^2 X^2) \leq 1 + \sum_{p=1}^\infty \frac{(2\lambda^2 p)^p}{(p/e)^p} = \sum_{p=0}^\infty (2e\lambda^2)^p$$

which converge to $\frac{1}{1-2e\lambda^2}$ provided $2e\lambda^2 < 1$, moreover, note $\frac{1}{1-x} \leq e^{2x}$ provided $x \in [0, \frac{1}{2}]$. It follows that

$$\mathbb{E}\exp(\lambda^2 X^2) \leq \exp(4e\lambda^2)$$

for all $|\lambda| \leq \frac{1}{2\sqrt{e}}$, this complete the proof with $K_3 = 2\sqrt{e}$.

3 $\implies$ 4 : Let $K_3 = 1$ and $\lambda = 1/2$, then $K_4 = 2$ and 3 implies 4.

4 $\implies$ 1: Suppose $K_4 = 1$. Then

$$\begin{aligned}
\mathbb{P}\left\{|X| \geq t\right\} &= \mathbb{P}\left\{e^{X^2} \geq e^{t^2}\right\} \\
&\leq e^{-t^2}\,\mathbb{E}\,e^{X^2} \text{ by Markov's inequality} \\
&\leq 2e^{-t^2}
\end{aligned}$$

This prove 1 with $K_1 = 1$.

Now assume $\mathbb{E}\,X = 0$ and show that 3 $\implies$ 5 and 5 $\implies$ 1.

3 $\implies$ 5: Note $e^x \leq x + e^{x^2}$ and assume $K_3 = 1$. Then $\forall\,|\lambda| \leq 1$,

$$\mathbb{E}\,e^{\lambda X} \leq \mathbb{E}\left[\lambda X + e^{\lambda^2 X^2}\right] = \overset{e^{\lambda^2 X^2}}{\mathbb{E}} \leq e^{\lambda^2}$$

For $\lambda > 1$, note $2\lambda x \le \lambda^2 + x^2$:

$$\mathbb{E}\, e^{\lambda x} \le e^{\frac{\lambda^2}{2}}\, \mathbb{E}\, e^{\frac{X^2}{2}} \le e^{\frac{\lambda^2}{2}+\frac{1}{2}} \le e^{\lambda^2}$$

This prove 5 with $K_5 = 1$.

$5 \implies 1$. Assume property 5 holds and $K_5 = 1$. Then

$$\mathbb{P}\{X \ge t\} = \mathbb{P}\left\{e^{\lambda X} \ge e^{\lambda t}\right\} \le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda X} \le e^{-\lambda t} e^{\lambda^2}$$

hold for any $\lambda > 0$. Then we conclude that

$$\mathbb{P}\{X \ge t\} \le \inf_{\lambda} e^{\lambda^2 - \lambda t} = e^{-\frac{t^2}{4}}$$

which prove 1 with $K_1 = 2$.

□

Exercise 2.6. Show that the property 5 in proposition 2.7 implies $\mathbb{E}\, X = 0$.

Solution. By Jensen's inequality **??**, we have

$$\exp(\lambda\, \mathbb{E}\, X) \le \mathbb{E} \exp(\lambda X) \le \exp(K^2\lambda^2)$$

thus $\lambda\, \mathbb{E}\, X \le K^2\lambda^2$ for any λ and thus $\mathbb{E}\, X = 0$.

Exercise 2.7. 1. Function $\lambda \mapsto \mathbb{E} \exp(\lambda^2 \mathfrak{Z}^2)$ is only finite in some interval $[-b, b]$.

2. Suppose X satisfy $\mathbb{E} \exp(\lambda^2 X^2) \le \exp(K\lambda^2)$ for **all** $\lambda \in \mathbb{R}$, then X is bounded *a.s.*, *i.e.*, $\|X\|_\infty < \infty$.

Solution. 1. Note that

$$\mathbb{E} \exp(\lambda^2 \mathfrak{Z}^2) = \left(\frac{1}{1-2\lambda^2}\right)^{\frac{1}{2}}$$

which is finite only if $2\lambda^2 < 1$.

2. TODO

2.5.2 Definition and examples of sub-gaussian distributions

Definition 2.3 (sub-gaussian distribution)**.** A *r.v.* X that satisfies one of the properties in 2.7 is called a sub-gaussian *r.v.*

Suppose that X, Y are sub-gaussian r.v. then $\|X\|_p \le K_2\sqrt{p}$, $\|Y\|_p \le W_2\sqrt{p}$, then

$$\|\alpha X + \beta Y\|_p \le |\alpha| \cdot \|X\|_p + |\beta| \cdot \|Y\|_p \le (|\alpha|\, K_2 + |\beta|\, W_2)\sqrt{p}$$

So all of the sub-gaussian r.v. form a vector space.

Proposition 2.8 (sub-gaussian space)**.** *There exists a norm called sub-gaussian norm of sub-gaussian space, denoted* $\|X\|_{\psi_2}$*, is defined as:*

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E} \exp(X^2/t^2) \le 2\right\}$$

Proof. We show that

$$\|X\| = \inf\left\{t > 0 : \|X\|_p \leq t\sqrt{p}\right\}$$

define a norm on sub-gaussian space. We only show the triangle inequality:

$$\|x+y\|_p \leq \|x\|_p + \|y\|_p \leq (\|x\| + \|y\|)\sqrt{p} \implies \|x+y\| \leq \|x\| + \|y\|$$

That justify $\|\cdot\|_{\psi_2}$ as K_2 and K_4 differ each by some constant c.

□

Example 2.2 (Examples of sub-Gaussian $r.v.'s$).

- (Gaussian): Note we have show that (by extend to $X = \sigma\mathfrak{Z}$):

  $$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-\frac{t^2}{2\sigma^2})$$

  that is $K_2 = \sqrt{2}\sigma$ and thus $\|X\|_{\psi_2} \leq C\sigma$.

- (Bernoulli): Let X is a symmetric Bernoulli distribution, then $X^2 = 1$ and consequently:

  $$\mathbb{E}\exp\left(\frac{X^2}{t^2}\right) = e^{\frac{1}{t^2}} \leq 2 \iff \frac{1}{\ln 2} \leq t^2$$

  it follows that $\|X\|_{\psi_2} = \frac{1}{\ln 2}$.

- (Bounded): Any bounded $r.v.$ is sub-gaussian. Suppose X is bounded by $b = \|X\|_\infty$

  $$\mathbb{E}\exp\left(\frac{X^2}{t^2}\right) \leq e^{\frac{b^2}{t^2}} \leq 2$$

  it follows that $\|X\|_{\psi 2} = \frac{1}{\ln 2}\|X\|_\infty$.

Exercise 2.8. Check that Poisson, exponential, Pareto and Cauchy distribution are not sub-Gaussian.

Solution. We show they are not by proving their tail $\mathbb{P}\{X \geq t\}$ cannot be bounded by gaussian tail $\exp(-ct^2)$.

Exercise 2.9. Let $(X_i)_{i\in\mathbb{N}^*}$ be sequence of sub-gaussian $r.v.'s$, then

$$\mathbb{E}\max_i \frac{|X_i|}{\sqrt{1+\ln i}} \leq CK$$

where $K = \max_i \|X_i\|_{\psi_2}$. For $n \geq 2$, we have

$$\mathbb{E}\max_{i\leq n}|X_i| \leq CK\sqrt{\ln n}$$

Solution. By proposition 2.7, we have $\forall t \geq 0$:

$$\mathbb{P}\{|X_i| > t\} \leq 2\exp\left(-c\frac{t^2}{\|X_i\|^2_{\psi_2}}\right) \leq 2\exp\left(-\frac{ct^2}{K^2}\right)$$

Note

$$\begin{aligned}\frac{1}{K}\mathbb{E}\max_i \frac{|X_i|}{\sqrt{1+\ln i}} &= \frac{1}{K}\int_0^\infty \mathbb{P}\left\{\max_i \frac{|X_i|}{\sqrt{1+\ln i}} > t\right\} dt \\ &\le \frac{1}{K}\int_0^{t_0} 1dt + \frac{1}{K}\int_{t_0}^\infty \sum_i \mathbb{P}\left\{|X_i| > t\sqrt{1+\ln i}\right\} dt\end{aligned}$$

Then we should find some t_0 *s.t.* the second integral is finite, it follows that

$$\begin{aligned}\int_{t_0}^\infty \sum_i \mathbb{P}\left\{|X_i| > t\sqrt{1+\ln i}\right\} dt &\le 2\int_{t_0}^\infty \sum_i \exp\left[-\frac{ct^2(1+\ln i)}{K^2}\right] dt \\ &= 2\int_{t_0}^\infty \exp\left[-\frac{ct^2}{K^2}\right]\sum_i \exp\left[-\frac{ct^2\ln i}{K^2}\right] dt \\ &\le 2\int_{t_0}^\infty \exp\left[-\frac{ct^2}{K^2}\right]\sum_i \exp\left[-\frac{ct_0^2\ln i}{K^2}\right] dt \\ &= 2\int_{t_0}^\infty \exp\left[-\frac{ct^2}{K^2}\right] dt \sum_i i^{-\frac{ct_0^2}{K^2}}\end{aligned}$$

Change variable $s = \frac{t}{K}$, then we have

$$\begin{aligned}\int_{t_0}^\infty \sum_i \mathbb{P}\left\{|X_i| > t\sqrt{1+\ln i}\right\} dt &\le 2\int_0^\infty \exp(-cs^2)ds + \sum_i i^{-\frac{ct_0^2}{K^2}} \\ &= \frac{\sqrt{\pi}}{\sqrt{c}} + N\end{aligned}$$

where we can let $t_0 = \frac{nK}{\sqrt{c}}$ for some $n > 1$ *s.t.* the summation is finite.

The second inequality follows from $\max_i \left\{\sqrt{1+\log i}\right\} = O(\log n)$.

2.6 General Hoeffding's and Khintchine's inequalities

Recall that a sum of independent normal *r.v.* X_i is normal, that is a form of the **rotation invariance** of normal distribution.
The rotation invariance property extends to general sub-gaussian distributions, albeit up to an absolute constant.

Proposition 2.9 (Sums of independent sub-gaussians)**.** *Let $X_i, i = 1, 2, \dots, n$ be independent, mean zero sub-gaussian r.v. Then $\sum_{i=1}^n X_i$ is also sub-gaussian and*

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2 \le C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$$

where C is an absolute constant.

Proof. Consider the MGF of the sum $\sum_{i=1}^n X_i$.

$$\begin{aligned}\mathbb{E}\exp\left(\lambda\sum_{i=1}^n X_i\right) &= \prod_{i=1}^n \mathbb{E}\exp(\lambda X_i)\\ &\le \prod_{i=1}^n \exp(C\lambda^2\|X_i\|_{\psi_2}^2)\\ &= \exp(\lambda^2 K^2)\end{aligned}$$

where $K^2 = C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$. For a sub-gaussian r.v. $X = \sum_{i=1}^n X_i$ with $\mu_X = 0$, note that

$$\mathbb{E}\exp(\lambda X) \le \exp(\lambda^2 K^2)$$

implies that there exists $c \in \mathbb{R}$ s.t.

$$\mathbb{E}\exp\left(\frac{X^2}{(cK)^2}\right) \le 2$$

Then recall the definition of $\|\cdot\|_{\psi_2}$ we find that there exists $c \in \mathbb{R}$ s.t.

$$\|X\|_{\psi_2} \le c\sqrt{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}$$

□

Theorem 2.6 (General Hoeffding inequality)**.** *Let $X_1, \dots, X_n$ be independent, mean-zero, sub-gaussian r.v. and $a = (a_1, \dots, a_n) \in \mathbb{R}^n$. Then for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| \ge t\right\} \le 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$

Proof. Let $X = \sum_{i=1}^n X_i$ with $\mu_X = 0$, sub-gaussian r.v.

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^n a_i X_i \ge t\right\} &= \mathbb{P}\left\{\exp\left(\lambda\sum_{i=1}^n a_i X_i\right) \ge \exp(\lambda t)\right\}\\ &\le \exp(-\lambda t)\cdot\mathbb{E}\exp\left(\lambda\sum_{i=1}^n a_i X_i\right)\\ &= \exp(-\lambda t)\cdot\prod_{i=1}^n \mathbb{E}\exp(\lambda a_i X_i)\end{aligned}$$

Note that X_i are sub-gaussian r.v.

$$\mathbb{E}\exp(\lambda a_i X_i) \le \exp(K_i^2(a_i\lambda)^2)$$

then

$$\prod_{i=1}^n \mathbb{E}\exp(\lambda a_i X_i) \le \exp\left(\lambda^2\sum_{i=1}^n K_i^2 a_i^2\right) \le \exp\left(cK^2\lambda^2\sum_{i=1}^n a_i^2\right) = \exp(cK^2\lambda^2\|a\|_2^2)$$

where $K = \max_i K_i, K_i = \sqrt{c}\|X_i\|_{\psi_2}$, then

$$\mathbb{P}\left\{\sum_{i=1}^n a_i X_i \geq t\right\} \leq \exp(-\lambda t) \cdot \exp(cK^2\lambda^2\|a\|_2^2) \text{ for all } \lambda > 0$$

then we have

$$\mathbb{P}\left\{\sum_{i=1}^n a_i X_i \geq t\right\} \leq \exp(-\lambda t + \alpha\lambda^2) \text{ where } \alpha = cK^2\|a\|_2^2$$

$$\mathbb{P}\left\{\sum_{i=1}^n a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{4\alpha}\right) \text{ by letting } \lambda = \frac{t}{2\alpha}$$

so

$$\mathbb{P}\left\{\sum_{i=1}^n a_i X_i \geq t\right\} \leq \exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right) \text{ where } K = \max_i \|X_i\|_{\psi_2}$$

□

Exercise 2.10. Deduce Hoeffding's inequality for bounded *r.v.* 2.4 from theorem 2.6

Solution. Note for bounded *r.v.*, we have $\|X_i - \mu\|_{\psi_2} \leq \frac{1}{\ln 2}(M_i - \mu_i) \leq \frac{1}{\ln 2}(M_i - m_i)$, then

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mu_i) > t\right\} \leq \exp\left(-\frac{ct^2(\ln 2)^2}{(M_i - m_i)^2}\right)$$

Theorem 2.7 (Khintchine's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-gaussian r.v. with zero-mean and unit variances, and let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, then for any $p \in [2, \infty)$, we have*

$$\|a\|_2 \leq \left\|\sum_{i=1}^n a_i X_i\right\|_p \leq CK\sqrt{p}\,\|a\|_2$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.

And for $p \in (0, 2)$, we have

$$c(K)\,\|a\|_2 \leq \left\|\sum_{i=1}^n a_i X_i\right\|_p \leq \|a\|_2$$

where $c(K)$ is some function of K.

Proof. **Step 1**. We first show that for $p \in [2, \infty)$. Note $\operatorname{Var} X_i = 1$, the first inequality is just consequence of $\|\cdot\|_2 \leq \|\cdot\|_p$. For another, by Hoeffding's inequality 2.6

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^p &= \int_0^\infty pt^{p-1}\,\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| > t\right\} dt \\
&\leq 2p\int_0^\infty t^{p-1}\exp\left(-\frac{ct^2}{K^2\,\|a\|_2^2}\right) dt \\
&= p\left(\frac{K\,\|a\|_2}{\sqrt{c}}\right)^p \Gamma\left(\frac{p}{2}\right)
\end{aligned}$$

Taking pth root and claim follows by Stirling's formula

$$\left[p\Gamma\left(\frac{p}{2}\right)\right]^{\frac{1}{p}} = O(p^{\frac{1}{p}}\left(\frac{p}{2}\right)^{\frac{p-1}{2}}) = O(\sqrt{p})$$

Step 2 Then we show that for $p \in (0,2)$. By Cauchy-Schwartz inequality (where $\mathbf{u} = \mathbb{E}\left|\sum\right|^{\frac{p}{2}}$ and $\mathbf{v} = \mathbb{E}\left|\sum\right|^{2-\frac{p}{2}}$):

$$\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^p \geq \left(\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^2\right)^2 \Big/ \mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^{4-p}$$

Note $4-p \geq 2$ and we have

$$\text{LHS} \geq \frac{\|a\|_2^4}{\left(CK\sqrt{4-p}\,\|a\|_2\right)^{4-p}} = c(K)^p \|a\|_2^p$$

where $c(K) = (CK\sqrt{4-p})^{1-\frac{4}{p}}$.

□

2.6.1 Centering

In results like Hoeffding's inequality, we typically assume X_i has zero mean. We can do so as the centering $X_i - \mu_i$ inherit most of properties of X_i. For example, $X_i - \mu_i$ keep in the L^2 space.

Proposition 2.10 (L^2 bound)**.**

$$\|X - \mathbb{E}X\|_2 \leq \|X\|_2$$

Proof. Suppose X is bounded. Note that

$$\int (X-\mu)^2 = \int (X^2 - 2\mu X + \mu^2) = \mathbb{E}X^2 - \mu^2 \leq \mathbb{E}X^2$$

□

And there is a similar results for sub-gaussian norm.

Proposition 2.11 (centering)**.** *If X is sub-gaussian, then $X - \mathbb{E}X$ is sub-gaussian and*

$$\|X - \mathbb{E}X\|_{\psi_2} \leq C\|X\|_{\psi_2}$$

where C is absolute constant.

Proof. Triangle inequality yield:

$$\|X - \mathbb{E}X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}X\|_{\psi_2}$$

□

Exercise 2.11. Show that the centering inequality above does not hold with $C = 1$.

Solution. Construct *r.v. s.t.*:

$$X = \begin{cases} a & \text{with probability } p \\ -a & \text{with probability } 1-p \end{cases}$$

where $a = \sqrt{\ln 2}$, it's clear that

$$\mathbb{E}\exp(X^2) = 2 \implies \|X\|_{\psi_2} = 1$$

Then we claim that $\|X-\mu\|_{\psi_2} > \|X\|_{\psi_2}$. Note $\mu = (2p-1)a$, we have

$$X - \mu = \begin{cases} 2(1-p)a & \text{with probability } p \\ -2pa & \text{with probability } 1-p \end{cases}$$

and thus

$$\mathbb{E}\exp\left(|X-\mu|^2\right) = p2^{4(1-p)^2} + (1-p)2^{4p^2}$$

which cannot bounded by 2 certainly.

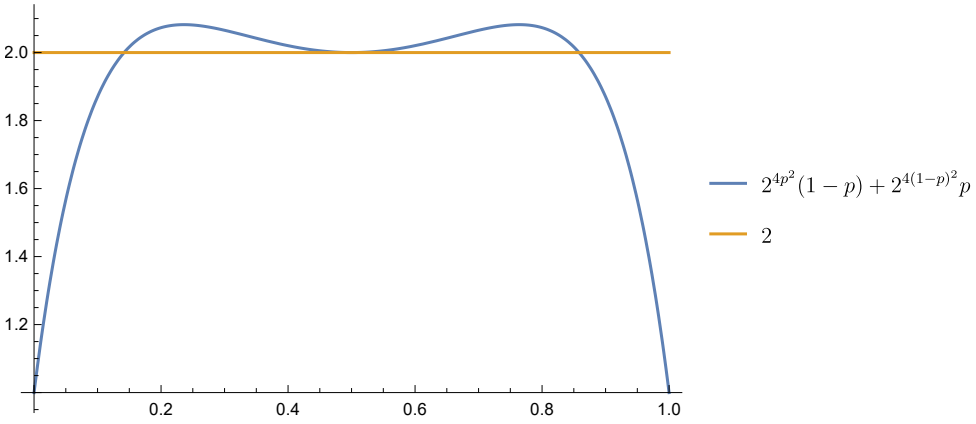

Figure 2.5: MGF of centering square

2.6.2 sub-exponential distribution

Proposition 2.12. *Let X be a r.v., TFAE*

1. *The tail of X satisfy:*
$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-\frac{t}{K_1}) \text{ for all } t \geq 0$$
2. *The moment of X satisfy:*
$$\|X\|_{L^p} \leq K_2 p \text{ for all } p \geq 1$$
3. *The MGF of $|X|$ satisfies:*
$$\mathbb{E}\exp(\lambda|X|) \leq \exp(K_3\lambda)$$
for all λ s.t. $0 \leq \lambda \leq 1/K_3$.

4. The MGF of $|X|$ is bounded at some point, namely

$$\mathbb{E}\exp\left(\frac{|X|}{K_4}\right) \leq 2$$

5. Moreover, if $\mathbb{E}X = 0$ then the MGF of X satisfies:

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2)$$

for all λ s.t. $|\lambda| \leq 1/K_5$.

Proof. $1 \implies 2$: Suppose $K_1 = 1$, then

$$\mathbb{P}\left\{|X| \geq t\right\} \leq 2\exp(-t)$$

and

$$\begin{aligned}
\|X\|_{L^p} &= (\mathbb{E}\,|X|^p)^{1/p} \\
&= \left(\int_0^\infty \mathbb{P}\left\{|X|^p \geq t\right\} dt\right)^{1/p} \\
&= \left(\int_0^\infty \mathbb{P}\left\{|X| \geq u\right\} pu^{p-1}du\right)^{1/p} \\
&\leq \left(\int_0^\infty 2e^{-u}pu^{p-1}du\right)^{1/p} \\
&= (2p\Gamma(p))^{1/p} \leq (6p\cdot p^p)^{1/p} = (6p)^{1/p}\cdot p \\
&\leq cp \text{ since } p^{\frac{1}{p}} \leq 2 \text{ for all } p \geq 1
\end{aligned}$$

$2 \implies 3$:

$$\begin{aligned}
\mathbb{E}\exp(\lambda\,|X|) &= \mathbb{E}\left[1 + \sum_{i=1}^\infty \frac{(\lambda\,|X|)^i}{i!}\right] \\
&= 1 + \sum_{i=1}^\infty \mathbb{E}\,\frac{(\lambda\,|X|)^i}{i!} \\
&= 1 + \sum_{i=1}^\infty \frac{\lambda^i\,\mathbb{E}\,|X|^i}{i!}
\end{aligned}$$

Note that $\mathbb{E}\,|X|^i \leq i^i$ and $p! \geq (p/e)^p$ then

$$\mathbb{E}\exp(\lambda\,|X|) \leq 1 + \sum_{i=1}^\infty \frac{\lambda^i i^i}{(i/e)^i} = \sum_{i=0}^\infty (\lambda e)^i$$

We should notice that if we don't assume $K_2 = 1$, then the series has the form

$$\sum_{i=0}^\infty \left(\frac{\lambda e}{K_2}\right)^i < \infty$$

which implies that $\lambda < K_2/e$ and

$$\sum_{i=0}^\infty \left(\frac{\lambda e}{K_2}\right)^i = \frac{1}{1-\alpha} \leq \exp(2\alpha) \text{ where } \alpha = \frac{\lambda e}{K_2} \in [0, \frac{1}{2}]$$

implies that

$$\lambda \leq \frac{K_2}{2e} \text{ and } \frac{1}{K_3} = \frac{2e}{K_2}$$

$2 \implies 5$:

$$\mathbb{E}\exp(\lambda X) = \mathbb{E}\left[1 + \lambda X + \sum_{i=2}^{\infty} \frac{(\lambda X)^i}{i!}\right]$$

then similar with above, we get

$$\mathbb{E}\exp(\lambda X) \leq 1 + \sum_{i=2}^{\infty} \frac{(\lambda i)^i}{(i/e)^i} = 1 + \frac{(e\lambda)^2}{1 - e\lambda}$$

if $|e\lambda| \leq 1/2$, then

$$1 + \frac{(e\lambda)^2}{1 - e\lambda} \leq 1 + 2e^2\lambda^2 \leq \exp(2e^2\lambda^2)$$

which implies

$$\mathbb{E}\exp(\lambda X) \leq \exp(2e^2\lambda^2) \text{ for } |\lambda| \leq \frac{1}{2e}$$

□

Definition 2.4. A *r.v.* satisfy above properties are called **sub-exponential**, and denoted

$$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}\exp\left(\frac{|X|}{t}\right) \leq 2\right\}$$

that is a norm for sub-exponential space.

Proposition 2.13. *A r.v. X is sub-gaussian iff X^2 is sub-exponential. Moreover,*

$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$$

Proof. Follows easily from the definition.

□

Proposition 2.14. *Let X and Y be sub-gaussian r.v. then XY is sub-exponential. Moreover,*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$$

Proof. Assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. Then

$$\mathbb{E}\exp(X^2) \leq 2 \text{ and } \mathbb{E}\exp(Y^2) \leq 2$$

Note that

$$\begin{aligned}
\mathbb{E}\exp(|XY|) &\leq \mathbb{E}\exp\left(\frac{X^2}{2} + \frac{Y^2}{2}\right) \\
&= \mathbb{E}\left[\exp\left(\frac{X^2}{2}\right)\exp\left(\frac{Y^2}{2}\right)\right] \\
&\leq \frac{1}{2}\,\mathbb{E}\left[\exp(X^2) + \exp(Y^2)\right] \\
&\leq \frac{1}{2}(2+2) = 2
\end{aligned}$$

So $\|\|XY|\|_{\psi_1} \leq 1$.

□

Similar for ψ_2, we have

$$\left\|X - \mathbb{E}\, X\right\|_{\psi_1} \leq C \|X\|_{\psi_1}$$

2.6.3 Orlicz space

Definition 2.5 (Orlicz space)**.** A function $\psi : [0, \infty) \to [0, \infty)$ is called an Orlicz function if

- ψ is convex
- ψ is increasing
- $\psi(0) = 0$
- $\psi(x) \to \infty$ when $x \to \infty$.

For a given Orlicz function ψ, the Orlicz norm of a $r.v.$ X is defined as:

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

The Orlicz space $L_\psi = L_\psi(\Omega, \mathcal{F}, \mathbb{P})$ consists of all $r.v.$ X on $(\Omega, \mathcal{F}, \mathbb{P})$ with finite Orlicz norm, *i.e.*

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

Exercise 2.12. Show that $\|\cdot\|_{\psi}$ is indeed a norm on L_ψ.

Solution. **Positive definite**. $\|X\|_{\psi} \geq 0$ clearly and if $\|X\|_{\psi} = 0$ then we can pick $t_n \to 0$ *s.t.*

$$1 \geq \lim_{n\to\infty} \mathbb{E}\, \psi\left(\frac{|X|}{t_n}\right) = \mathbb{E} \lim_{n\to\infty} \psi\left(\frac{|X|}{t_n}\right) = \infty \cdot \mathbb{P}\{|X| > 0\}$$

that force $X = 0$ *a.s.*

Absolutely homogeneous. Note

$$\|cX\|_{\psi} = \inf_{t>0}\left\{\mathbb{E}\, \psi\left(\frac{|cX|}{t}\right) \leq 1\right\} = \inf_{t>0}\left\{\mathbb{E}\, \psi\left(\frac{|X|}{t/|c|}\right) \leq 1\right\} = c\|X\|_{\psi}$$

Subadditive Since ψ is increasing and convex,

$$\begin{aligned}
\mathbb{E}\, \psi\left(\frac{|X+Y|}{\|X\|_{\psi} + \|Y\|_{\psi}}\right) &\leq \mathbb{E}\, \psi\left(\frac{|X| + |Y|}{\|X\|_{\psi} + \|Y\|_{\psi}}\right) \\
&\leq \mathbb{E}\left[\frac{|X|_{\psi}}{\|X\|_{\psi} + \|Y\|_{\psi}} \psi\left(\frac{|X|}{\|X\|_{\psi}}\right) + \frac{|Y|_{\psi}}{\|X\|_{\psi} + \|Y\|_{\psi}} \psi\left(\frac{|Y|}{\|Y\|_{\psi}}\right)\right] \\
&\leq 1(\text{ as } \mathbb{E}\, \psi\left(\frac{|X|}{\|X\|_{\psi}}\right) \leq 1)
\end{aligned}$$

and that implies $\|X + Y\|_{\psi} \leq \|X\|_{\psi} + \|Y\|_{\psi}$ by definition.

And it can be shown L_ψ is a Banach space. Also, $\|\cdot\|_{\psi_2}$ is just $\|\cdot\|_\psi$ when $\psi(x) = e^{x^2} - 1$ and L^p space is L_ψ when $\psi(x) = x^p$. Moreover, we have

$$L^\infty \subset L_{\psi_2} \subset L^p$$

2.7 Bernstein's inequality

Theorem 2.8 (Bernstein's inequality)**.** *Let $X_1, \dots, X_n$ be independent, mean-zero, sub-exponential r.v.. Then for each $t \geq 0$, we have:*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right]$$

where $c > 0$ is an absolute constant.

Proof. As before, let $S = \sum_{i=1}^n X_i$ then by Chernoff bound:

$$\mathbb{P}\{S \geq t\} \leq e^{-\lambda t} \cdot \prod_{i=1}^n \mathbb{E}\exp(\lambda X_i)$$

Recall that for a sub-exponential distribution,

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2 \lambda^2)$$

when λ is small enough so that $|\lambda| \leq \frac{c}{\max_i \|X_i\|_{\psi_1}}$ and then

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(C\lambda^2 \|X\|_{\psi_1}^2\right)$$

Substitute that, we have

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\lambda t + C\lambda^2\sigma^2\right) \text{ where } \sigma^2 = \sum_{i=1}^n \|X_i\|_{\psi_1}^2$$

Now we minimize the expression in λ subject to the constraint. Note the bound for λ yield the optimal choice of λ is:

$$\lambda = \min\left(\frac{t}{2C\sigma^2}, \frac{c}{\max_i \|X_i\|_{\psi_1}}\right)$$

Note $C\lambda^2\sigma^2 \leq \frac{\lambda t}{2} \iff |\lambda| \leq \frac{t}{2C\sigma^2}$, we have $-\lambda t + C\lambda^2\sigma^2 \leq -\frac{\lambda t}{2}$. And thus:

$$\mathbb{P}\{S \geq t\} \leq \exp\left[-\min\left(\frac{t^2}{4C\sigma^2}, \frac{ct}{2\max_i \|X_i\|_{\psi_1}}\right)\right]$$

For same argument of $-X$ instead of X, we got the same bound. So the Bernstein's inequality holds.

□

The following put theorem 2.8 in a more convenient form:

Theorem 2.9. *Let $X_1, \dots, X_n$ be independent, mean-zero and sub-exponential r.v. and $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, then for every $t \geq 0$, we have:*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_1}$.

In the special case where $a_i = \frac{1}{N}$, we have

Corollary 2.2. *Let $X_1, \dots, X_n$ be independent, mean-zero and sub-exponential r.v., then for $t \geq 0$, we have*

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

where $K = \max_i \|X_i\|_{\psi_1}$.

Remark. We may note that theorem 2.8 has two tails, as if the sum were mixture of sub-gaussian and sub-exponential distributions.
The sub-gaussian tail is of course expected from the CLT. But the sub-exponential tails of the terms X_i are too heavy to be able to produce a sub-gaussian tail everywhere, so the sub-exponential tail should be expected, too.
That is, sub-gaussian for small deviation while sub-exponential for the large:

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right\} \leq \begin{cases} 2\exp(-\frac{t^2}{K^2/cn}) & t \leq K \\ 2\exp(-\frac{t}{K/cn}) & t > K \end{cases}$$

Let us mention the following strengthening of Bernstein's inequality under the stronger assumption that the random variables Xi are bounded.

Theorem 2.10 (Bernstein's inequality for bounded *r.v.*)**.** *Let $X_1, \dots, X_n$ be independent, mean-zero and bounded r.v. s.t. $|X_i| \leq K$ for some $K \in [0, \infty)$ holds for every i.*
Then we have:

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

for every $t \geq 0$ and $\sigma^2 = \sum_{i=1}^n \mathbb{E}\, X_i^2$.

Proof. Similar with the proof of Bernstein's inequality, we have:

$$\mathbb{P}\left\{\sum_{i=1}^n X_i \geq t\right\} = \mathbb{P}\left\{\exp\left(\lambda\sum_{i=1}^n X_i\right) \geq \exp(\lambda t)\right\} \leq \frac{\mathbb{E}\exp\left(\lambda\sum_{i=1}^n X_i\right)}{\exp(\lambda t)}$$

Now consider $\mathbb{E}\exp(\lambda X_i)$.

First consider the result if X is a bounded *r.v. s.t.* $|X| \leq K$ then

$$\mathbb{E}\exp(\lambda X) \leq \exp(g(\lambda)\,\mathbb{E}\, X^2) \text{ where } g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|\, K/3}$$

when $|\lambda| < K/3$.

To prove the result, note that $e^x \leq 1 + x + \frac{x^2/2}{1-|x|/3}$, then replace x with λX, we have:

$$\mathbb{E}\exp(\lambda X) \leq 1 + \lambda K + \mathbb{E}\frac{\lambda^2 X^2/2}{1-|\lambda X|/3} \leq 1 + \lambda K + g(\lambda)\,\mathbb{E}\,X^2 \leq \exp(g(\lambda)\,\mathbb{E}\,X^2)$$

Then we continue to analysis $\mathbb{E}\exp(\lambda X_i)$

$$\mathbb{E}\exp(\lambda X_i) \leq \exp(g(\lambda)\,\mathbb{E}\,X_i^2) \text{ for } |\lambda| \leq K/3$$

So

$$\mathbb{E}\exp\left(\lambda\sum_{i=1}^n X_i\right) \leq \prod_{i=1}^n \exp(g(\lambda)\,\mathbb{E}\,X_i^2) = \exp\left(g(\lambda)\sum_{i=1}^n \mathbb{E}\,X_i^2\right) = \exp(g(\lambda)\sigma^2)$$

Thus

$$\mathbb{P}\{S \geq t\} \leq e^{-\lambda t}\cdot\exp(g(\lambda)\sigma^2) = \exp\left(-\lambda t + \frac{\lambda^2/2}{1-|\lambda|\,K/3}\sigma^2\right)$$

Note that $|\lambda| \leq K/3$, $\sigma^2 \geq 0$ and $t \geq 0$, let

$$\lambda = \frac{t}{\sigma^2 + tK/3}$$

we can get the inequality.

□

2.8 Tensorization and bounded difference

In this chapter *r.v.* denotes random variable or random vectors.

Proposition 2.15. *Let X be any r.v. or random vector. Then*

$$Var(f(X)) \leq \frac{1}{4}(\sup f - \inf f)$$

and

$$Var(f(X)) \leq \mathbb{E}\left[(f(X) - \inf f)^2\right]$$

Proof. Recall that

$$\mathrm{Var}\,[f(X)] = \mathbb{E}\,(f(X) - \mathbb{E}\,f(X))^2$$

so

$$\mathrm{Var}\,[f(X)] = \mathrm{Var}\,[f(X) - a] \leq \mathbb{E}\,(f(X) - a)^2 \text{ for any } a \in \mathbb{R}$$

Let $a = (\sup f + \inf f)/2$, then

$$|f(X) - a| \leq \frac{\sup f - \inf f}{2}$$

which implies

$$\mathrm{Var}\,[f(X)] = \mathrm{Var}\left[f(X) - \frac{\sup f + \inf f}{2}\right] \leq \mathbb{E}\left[f(X) - \left(\frac{\sup f + \inf f}{2}\right)\right]^2 \leq \frac{1}{4}(\sup f - \inf f)^2$$

and for the next one, let $a = \inf f$.

□

Theorem 2.11 (tensorization of var)**.** *Given independent r.v.* $X_1, X_2, \dots, X_n$*, we have*

$$Var[f(X_1, \dots, X_n)] \leq \mathbb{E}\left[\sum_{i=1}^{n} \underset{i}{Var}[f(X_1, \dots, X_n)]\right]$$

Proof. □

Chapter 3

Random Vector

3.1 Concentration of Norm

Suppose $X = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ are independent with zero mean and unit variance, we have

$$\mathbb{E}\,\|X\|_2^2 = \mathbb{E}\sum_{i=1}^n X_i^2 = \sum_{i=1}^n \mathbb{E}\,X_i^2 = n$$

so we expect $\|X\|_2 \approx \sqrt{n}$. We will see X is indeed X is very close to $\sqrt{n}$ with high probability.

Theorem 3.1 (concentration of norm)**.** *Let $X = (X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-gaussian coordination X_i that satisfy $\mathbb{E}\,X_i^2 = 1$. Then*

$$\left\|\|X\|_2 - \sqrt{n}\right\|_{\psi_2} \le CK^2$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is absolute constant.

Proof. WLOG, assume that $K \ge 1$, note that

$$\frac{1}{n}\|X\|_2^2 - 1 = \frac{1}{n}\sum_{i=1}^n (X_i^2 - 1)$$

Since X_i is sub-gaussian then $X_i^2 - 1$ is sub-exponential, with

$$\begin{aligned}\left\|X_i^2 - 1\right\|_{\psi_1} &\le C\left\|X_i^2\right\|_{\psi_1} \\ &= C\,\|X_i\|_{\psi_2}^2 \\ &\le CK^2\end{aligned}$$

Applying Bernstein's inequality 2.2, we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\|x\|_2^2 - 1\right| \ge u\right\} \le 2\exp\left(\frac{-cn}{K^4}\min(u^2, u)\right)$$

since we have assumed that $K \ge 1$ and thus $K^4 \ge K^2$.

Observe that for all $z \geq 0$

$$|z-1| \geq \delta \implies \left|z^2-1\right| \geq \max(\delta, \delta^2)$$

then

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{1}{\sqrt{n}}\left\|X\right\|_2 - 1\right| \geq \delta\right\} &\leq \left\{\mathbb{P}\left|\frac{1}{n}\left\|X\right\|_2 - 1\right| \geq \max(\delta, \delta^2)\right\} \\ &\leq 2\exp\left(\frac{-cn}{K^4}\cdot\delta^2\right) \text{ with } u = \max(\delta, \delta^2)\end{aligned}$$

changing $t = \delta\sqrt{n}$, we obtain

$$\mathbb{P}\left\{\left|\left\|X\right\|_2 - \sqrt{n}\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{K^4}\right) \text{ for all } t \geq 0$$

Then $\left\|X\right\|_2 - \sqrt{n}$ is a sub-gaussian $r.v.$ and

$$\left\|\left\|X\right\|_2 - \sqrt{n}\right\|_{\psi_2} \leq CK^2$$

□

Remark. X even even stays within constant distance from that sphere of radius $\sqrt{n}$ with high probability, say 0.99. To see this, note $S_n = \left\|X\right\|_2^2$ has mean n and standard deviation $O(\sqrt{n})$:

$$\begin{aligned}\operatorname{Var}\left(\left\|X\right\|_2^2\right) = \operatorname{Var}\left(\sum_{i=1}^n X_i^2\right) &= \mathbb{E}\left[\left(\sum_{i=1}^n X_i^2\right) - n\right]^2 \\ &= \mathbb{E}\sum_i\sum_j (X_iX_j)^2 - n^2 \\ &= \sum_{i=1}^n \mathbb{E}\, X_i^4 + \sum_{i\neq j}\mathbb{E}\, X_i^2X_j^2 - n^2\end{aligned}$$

Note $\mathbb{E}\, X_i^4 = O(1)$ by definition of sub-gaussian 2.7 and $\mathbb{E}\, X_i^2X_j^2 = 1$ when $i \neq j$. Thus

$$\operatorname{Var}\left(\left\|X\right\|_2^2\right) = nO(1) - n = O(n)$$

then the claim follows. That implies $\left\|X\right\|_2 = \sqrt{S_n}$ ought to deviate by $O(1)$ around $\sqrt{n}$. This is because

$$\sqrt{n \pm O(\sqrt{n})} = \sqrt{n} \pm O(1)$$

Exercise 3.1 (Expectation of the norm)**.** In the setting of theorem 3.1, show that

1. $\sqrt{n} - CK^2 \leq \mathbb{E}\left\|X\right\|_2 \leq \sqrt{n} + CK^2$
2. CK^2 can be replaced by $o(1) \to 0$ as $n \to \infty$.
3. $\operatorname{Var}\left\|X\right\|_2 \leq CK^4$

Solution. 1. Note that

$$\left|\mathbb{E}\left[\left\|X\right\|_2 - \sqrt{n}\right]\right| \leq \mathbb{E}\left|\left\|X\right\|_2 - \sqrt{n}\right| = \left\|\left\|X\right\|_2 - \sqrt{n}\right\|_1 \leq C\left\|\left\|X\right\|_2 - \sqrt{n}\right\|_{\psi_2}$$

thus

$$\left|\mathbb{E}\left\|X\right\|_2 - \sqrt{n}\right| \leq CK^2 \text{ where } K = \max_i \left\|X_i\right\|_{\psi_2}$$

It follows that

$$\sqrt{n} - CK^2 \leq \mathbb{E}\left\|X\right\|_2 \leq \sqrt{n} + CK^2$$

2. For the CK^2 in the upper bound, we have

$$\mathbb{E}\left\|X\right\|_2 \leq \sqrt{\mathbb{E}\left\|X\right\|_2^2} = \sqrt{n}$$

by Jensen's inequality. For the lower bound, we have

$$f(u) = \sqrt{u} - \frac{1}{2}(1 + u - (u-1)^2),\ u \geq 0$$

then changing u with $\left\|X\right\|_2^2/n$ leads to

$$\left\|X\right\|_2 \geq \frac{\sqrt{n}}{2}\left(\frac{n + \left\|X\right\|_2^2}{n} - \left(\frac{\left\|X\right\|_2^2 - \mathbb{E}\left\|X\right\|_2^2}{n}\right)^2\right)$$

Take expectations, we find that

$$\mathbb{E}\left\|X\right\|_2 \geq \sqrt{n} - \frac{\mathrm{Var}\left[\left\|X\right\|_2^2\right]}{2n\sqrt{n}}$$

and

$$\frac{\mathrm{Var}\left[\left\|X\right\|_2^2\right]}{n} = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left[X_i^2\right] \leq \max_i \mathrm{Var}\left[X_i^2\right] \leq \max_i \mathbb{E}\, X_i^4 \leq \max_i \left\|X_i\right\|_{\psi_2}^4 = K^4$$

thus

$$\mathbb{E}\left\|X\right\|_2 \geq \sqrt{n} - O(K^4/\sqrt{n}) = \sqrt{n} - o(1)$$

3. Note that

$$\begin{aligned}\mathrm{Var}\left\|X\right\|_2 &\leq \mathbb{E}\left(\|X\| - \sqrt{n}\right)^2 \\ &= \mathbb{E}\left\|X\right\|_2^2 - 2\sqrt{n}\,\mathbb{E}\left\|X\right\|_2 + n \\ &= 2\sqrt{n}(\sqrt{n} - \mathbb{E}\left\|X\right\|_2)\end{aligned}$$

In the proof of statement 2, we have $\mathbb{E}\left\|X\right\|_2 \geq \sqrt{n} - O\left(\frac{K^4}{\sqrt{n}}\right)$ and thus $\mathrm{Var}\left\|X\right\|_2 \leq O(K^4)$.

Exercise 3.2 (Small ball probabilities)**.** Let $X_1, \ldots, X_n$ be non-negative independent $r.v.'s$ with continuous distribution. Assume that the densities of X_i are bounded by 1. Show that, for any $\epsilon > 0$, we have

$$\mathbb{P}\left\{\left\|X\right\|_2 \leq \epsilon\sqrt{n}\right\} \leq (C\epsilon)^n$$

Solution. Note $\forall t > 0$,

$$\mathbb{E}\exp\left(-tX_i^2\right) = \int_0^\infty e^{-tx^2} f_{X_i}(x)dx \leq \int_0^\infty e^{-tx^2}dx = \frac{1}{2}\sqrt{\frac{\pi}{t}}$$

By Chernoff bound, we have

$$\begin{aligned}\mathbb{P}\left\{\left\|X\right\|_2 \leq \epsilon\sqrt{n}\right\} &= \mathbb{P}\left\{\left\|X\right\|_2^2 \leq \epsilon^2 n\right\} \\ &\leq e^{\epsilon^2 nt}\prod_{i=1}^n M_{X_i^2}(-\epsilon^2) \\ &\leq e^{\epsilon^2 nt}\left(\frac{1}{2}\sqrt{\frac{\pi}{t}}\right)^n\end{aligned}$$

Then $t = \epsilon^{-2}$ yield the desired result.

3.2 Covariance matrices and PCA

$$\operatorname{Cov}[X] = \mathbb{E}(X-\mu)(X-\mu)^t = \mathbb{E}\,XX^t - \mu\mu^t \text{ where } \mu = \mathbb{E}\,X$$

$$\Sigma[X] = \mathbb{E}\,XX^t$$

where $\Sigma[X]$ is somewhere like a one-dim $r.v.$'s second moment.

Note that $\Sigma[X]$ is symmetric and so its eigenvalues are all non-negative, then $\Sigma[X]$ can be decomposed as

$$\Sigma = \sum_{i=1}^{n} s_i u_i u_i^t$$

where u_i are eigenvectors of Σ and s_i are eigenvalues of Σ according to spectral theorem.

3.2.1 Isotropy

Definition 3.1 (isotropy)**.** A random vector $\mathbf{X} \in \mathbb{R}^n$ is called isotropic if

$$\mathbb{E}\,\mathbf{X}\mathbf{X}' = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix of $\mathbb{R}^n$.

Note that if X has positive variance, then we can reduce it

$$Z = \frac{X-\mu}{\sqrt{\operatorname{Var}(X)}}$$

Then $\mathbb{E}\,Z = 0$, $\operatorname{Var}(Z) = 1$, which called the **standard score** of a $r.v.$ X.

The following exercise gives a high-dimensional version of standard score.

Exercise 3.3. 1. Let $\mathbf{Z}$ be a mean-zero, isotropic vector in $\mathbb{R}^n$. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ be a fixed vector and $\boldsymbol{\Sigma}$ be a fixed $n \times n$ positive-semidefinite matrix. Then the vector

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{Z}$$

has mean $\boldsymbol{\mu}$ and covariance $\operatorname{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$.

2. Similarly, for a $r.v.$ $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and $\operatorname{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, let

$$\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$$

we have $\mathbf{Z}$ is isotropic with mean zero.

Proof. **1**

$$\mathbb{E}\,\mathbf{X} = \mathbb{E}\,\boldsymbol{\mu} = \boldsymbol{\mu}$$

and

$$\mathbb{E}\,(\mathbf{X}-\boldsymbol{\mu})^2 = \mathbb{E}\,\mathbf{X}\mathbf{X}' - \boldsymbol{\mu}\boldsymbol{\mu}' = \mathbb{E}\,\mathbf{X}\mathbf{X}' = \boldsymbol{\Sigma}$$

2

$$\mathbb{E}\,\mathbf{Z} = \mathbb{E}\begin{pmatrix} \sum_{i=1}^n a_{1i}(x_1-\mu_1) \\ \vdots \\ \sum_{i=1}^n a_{ni}(x_n-\mu_n) \end{pmatrix} = \mathbf{0}$$

where $a_{ij} = \operatorname{Cov}(\mathbf{X})_{ij}$.

$$\operatorname{Cov}(\mathbf{Z}) = \mathbb{E}\,\mathbf{Z}\mathbf{Z}' = \mathbb{E}\left(\boldsymbol{\Sigma}^{-\mathbf{1/2}}(\mathbf{X}-\boldsymbol{\mu})\right)\left(\boldsymbol{\Sigma}^{-\mathbf{1/2}}(\mathbf{X}-\boldsymbol{\mu})\right)' = \mathbb{E}\,\mathbf{I} = \mathbf{I}$$

by decomposing $\operatorname{Cov}(\mathbf{X}) = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})'$.

□

3.2.2 Isotropic Distribution

Proposition 3.1. *A r.v. $\mathbf{X} \in \mathbb{R}^n$ is isotropic iff*

$$\mathbb{E}\langle \mathbf{X}, x\rangle^2 = \|x\|_2^2 \text{ for all } x \in \mathbb{R}^n$$

Proof.

Lemma 3.1. *Suppose $\mathbf{A}$ and $\mathbf{B}$ are symmetric, then $\mathbf{A} = \mathbf{B}$ iff for any $\mathbf{x} \in \mathbb{R}^n$,*

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{B}\mathbf{x}$$

Proof. Suppose $\mathbf{A} = \mathbf{B}$, then it's obvious. For the converse, find some $\mathbf{x} \in \mathbb{R}^n$ *s.t.* $\mathbf{x}\mathbf{x}' = \mathbf{I}$, then

$$\mathbf{x}\mathbf{x}'\mathbf{A}\mathbf{x}\mathbf{x}' = \mathbf{x}\mathbf{x}'\mathbf{B}\mathbf{x}\mathbf{x}' \implies \mathbf{I}\mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{B}\mathbf{I} \implies \mathbf{A} = \mathbf{B}$$

□

Thus $\mathbf{X}$ is isotropic iff

$$\mathbf{x}'\left(\mathbb{E}\,\mathbf{X}\mathbf{X}'\right)\mathbf{x} = \mathbf{x}'\mathbf{I}\mathbf{x} \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

where the left side is precisely $\mathbb{E}\langle \mathbf{X}, \mathbf{x}\rangle^2$ while the right is $\|x\|_2^2$.

□

Proposition 3.2. *Let $\mathbf{X}$ be an isotropic r.v. in $\mathbb{R}^n$. Then*

$$\mathbb{E}\,\|\mathbf{X}\|_2^2 = n$$

Moreover, if $\mathbf{X}$ and $\mathbf{Y}$ are two independent isotropic r.v. in $\mathbb{R}^n$, then

$$\mathbb{E}\langle \mathbf{X}, \mathbf{Y}\rangle^2 = n$$

Proof. To show the first part, we have

$$\begin{aligned} \mathbb{E}\,\|\mathbf{X}\|_2^2 &= \mathbb{E}\,\mathbf{X}'\mathbf{X} = \mathbb{E}\operatorname{tr}(\mathbf{X}'\mathbf{X}) \\ &= \mathbb{E}\operatorname{tr}(\mathbf{X}\mathbf{X}') \\ &= \operatorname{tr}\left(\mathbb{E}\,\mathbf{X}\mathbf{X}'\right) \\ &= \operatorname{tr}\mathbf{I} = n \end{aligned}$$

For the second part, fix a realization of $\mathbf{Y}$ and compute the condition expectation. Note that

$$\begin{aligned}\mathbb{E}\langle \mathbf{X}, \mathbf{Y}\rangle^2 &= \mathop{\mathbb{E}}_{\mathbf{Y}} \mathop{\mathbb{E}}_{\mathbf{X}} \left[\langle \mathbf{X}, \mathbf{Y}\rangle^2 \mid \mathbf{Y}\right] \\ &= \mathop{\mathbb{E}}_{\mathbf{Y}} \|\mathbf{Y}\|_2^2 \,(\text{use previous proposition by } \mathbf{x} = \mathbf{Y}) \\ &= n\end{aligned}$$

□

Remark (Almost orthogonality of independent vectors). Normalize X and Y, we have

$$\overline{\mathbf{X}} = \frac{\mathbf{X}}{\|\mathbf{X}\|_2} \text{ and } \overline{\mathbf{Y}} = \frac{\mathbf{Y}}{\|\mathbf{Y}\|_2}$$

we have $\|\mathbf{X}\|_2 \asymp \|\mathbf{Y}\|_2 \asymp \langle \mathbf{X}, \mathbf{Y}\rangle \asymp \sqrt{n}$, thus

$$\left\langle \overline{\mathbf{X}}, \overline{\mathbf{Y}} \right\rangle \asymp \frac{1}{\sqrt{n}} \to 0 \text{ as } n \to \infty$$

That implies in high-dimensional spaces, independent isotropic random vectors tend to be almost orthogonal.

Exercise 3.4 (Distance between independent isotropic vectors)**.** Let $\mathbf{X}$ and $\mathbf{Y}$ be independent, mean-zero, isotropic *r.v.* in $\mathbb{R}^n$, then

$$\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|_2^2 = 2n$$

Solution.

$$\begin{aligned}\mathbb{E}\|\mathbf{X} - \mathbf{Y}\|_2^2 &= \mathbb{E}\langle \mathbf{X} - \mathbf{Y}, \mathbf{X} - \mathbf{Y}\rangle \\ &= \mathbb{E}\|\mathbf{X}\|_2^2 + \mathbb{E}\|\mathbf{Y}\|_2^2 + 2\,\mathbb{E}\langle \mathbf{X}, \mathbf{Y}\rangle \\ &= 2n + 2\,\mathbb{E}\langle \mathbf{X}, \mathbf{Y}\rangle\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\,\mathbf{X}'\mathbf{Y} &= \mathop{\mathbb{E}}_{\mathbf{Y}} \mathop{\mathbb{E}}_{\mathbf{X}} (\mathbf{X}\mathbf{Y} \mid \mathbf{Y}) \\ &= \mathop{\mathbb{E}}_{\mathbf{Y}} \left(\mathbf{Y} \mathop{\mathbb{E}}_{\mathbf{X}} \mathbf{X}\right) \\ &= \mathop{\mathbb{E}}_{\mathbf{Y}} \mathbf{0} = 0\end{aligned}$$

then claim follows.

3.3 Spherical Distribution

Definition 3.2 (spherical distribution)**.** A distribution is called a spherical distribution if a random vector X is uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at 0 and radius $\sqrt{n}$.

$$X \sim \text{Unif}\left(\sqrt{n} S^{n-1}\right)$$

where S^{n-1} is the sphere in $\mathbb{R}^n$.

Proposition 3.3. *A spherical distributed random vector X is isotropic.*

Proof.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \sim \text{Unif}\left(\sqrt{n}S^{n-1}\right)$$

then for any $x \in \mathbb{R}^n$, we have:

$$\langle \mathbf{X}, x \rangle = \left\langle \mathbf{X}, \|x\|_2 e \right\rangle \text{ in distribution}$$

this means that for every $x_1, x_2 \in \mathbb{R}^n$, $\langle \mathbf{X}, x_1 \rangle = \langle \mathbf{X}, x_2 \rangle$ in distribution. So

$$\mathbb{E}\langle \mathbf{X}, x \rangle^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\langle \mathbf{X}, \|x\|_2 e_i \right\rangle^2 = \frac{1}{n}\mathbb{E}\sum_{i=1}^n \left(\|x\|_2 X_i\right)^2 = \|x\|_2^2 \mathbb{E}\frac{1}{n}\sum_{i=1}^n X_i^2 = \|x\|_2^2$$

And in above proof, we used $\|X\|_2^2 = n$ which means X_i s are not independent.

□

Definition 3.3 (multivariate normal)**.** We say a random vector $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ is multivariate normal if $X_1, \ldots, X_n$ are independent *r.v.* of $\mathcal{N}(0,1)$, denoting as

$$\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_n)$$

Proposition 3.4. *Consider a random vector $X \sim \mathbf{N}(0, \mathbf{I}_n)$, then the density is*

$$f(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{2\pi} e^{-\|x\|_2^2/2}, x \in \mathbb{R}^n$$

3.4 Examples of high-dimensional distributions

3.4.1 Spherical and Bernoulli distributions

Definition 3.4 (uniformly distributed)**.** We say a *r.v.* is uniformly distributed on some subsets $T \subset \mathbb{R}^n$, such as $\sqrt{n}S^{n-1}$ if for every Borel subsets $E \subset T$, the probability $\mathbb{P}\{\mathbf{X} \in E\} = \mu E/\mu S^{n-1}$ in Lebesgue meaning in $\mathbb{R}^{n-1}$.

The coordinate of an isotropic random vector are always uncorrelated, but not necessarily independent, *e.g.*

Definition 3.5 (Spherical Distribution)**.** We say a *r.v.* $\mathbf{X}$ is the spherical distribution if $\mathbf{X}$ is uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:

$$\mathbf{X} \sim \text{Unif}\left(\sqrt{n}S^{n-1}\right)$$

Exercise 3.5. If $\mathbf{X} \sim \text{Unif}\left(\sqrt{n}S^{n-1}\right)$, then $\mathbf{X}$ is isotropic with zero mean.

Solution. Note that the sphere symmetry implies $\mathbb{E}\,\mathbf{X} = \mathbf{0}$. And for the variance, $\mathbb{E}\,\mathbf{X}\mathbf{X}'$, for any $\mathbf{x} \in \mathbb{R}^n$, again by symmetry we have

$$\langle \mathbf{X}, \mathbf{x} \rangle \overset{d}{=} \left\langle \mathbf{X}, \|\mathbf{x}\|_2 \mathbf{e} \right\rangle, \forall \mathbf{e} \in S^{n-1}$$

then let $\mathbf{e}_i$ denote the unit vector in the meaning of $\mathbb{R}^n$.

$$\mathbb{E}\,\langle \mathbf{X}, x \rangle = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\langle \mathbf{X}, \|x\|_2 \mathbf{e}_i \right\rangle^2 = \frac{1}{n}\,\mathbb{E}\sum_{i=1}^{n}\left(\|x\|_2 X_i\right)^2 = \|x\|_2^2\,\mathbb{E}\,\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \|x\|_2^2$$

where $\sum_{i=1}^n X_i^2 = \|\mathbf{X}\|_2^2 = n$ which shows $\mathbf{X}$ is isotropic.

Symmetric Bernoulli distribution is a good example of discrete isotropic vectors, formally, $\mathbf{X}$ is symmetric Bernoulli if

$$X \sim \text{Unif}(\{-1, 1\}^n)$$

Clearly, such $\mathbf{X}$ is isotropic.

More generally, any random vector $\mathbf{X}$ whose coordinate X_i are standard and independent is an isotropic vector.

3.4.2 Multivariate Normal

One of the most important high-dimensional distributions is Gaussian, or multivariate normal.

Definition 3.6 (Gaussian)**.** We say a *r.v.* $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if every coordinate of $\mathbf{X}$ are independent unit normal distribution, *i.e.* $X_i \sim \mathfrak{Z}$.

Multivariate normal distribution is isotropic by the previous discussion of general case. The density is

$$f_{\mathfrak{Z}}(\mathbf{x}) = (2\pi)^{-n/2} \exp\left\{-\sum_{i=1}^{n} x_i^2/2\right\} = (2\pi)^{-n/2} \exp(-\mathbf{x}'\mathbf{x}/2)$$

The mgf is

$$\begin{aligned} M_{\mathfrak{Z}}(\mathbf{t}) &= E\left\{\exp\left(\mathbf{t}^T\mathbf{Z}\right)\right\} = E\left\{\exp\left(\sum_{i=1}^{p} t_i Z_i\right)\right\} = \prod_{i=1}^{p} m_{z_i}(t_i) \\ &= \exp\left\{\sum_{i=1}^{p} t_i^2/2\right\} = \exp\left\{\mathbf{t}^T\mathbf{t}/2\right\} \end{aligned}$$

Thus the mgf for $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$ can be constructed:

$$m_{\mathbf{X}}(\mathbf{t}) = E\left[e^{\mathbf{t}^T\mathbf{X}}\right] = E\left[e^{\mathbf{t}^T\mu + \mathbf{t}^T\mathbf{A}\mathbf{Z}}\right] = e^{\mathbf{t}^T\mu} \times m_z\left(\mathbf{A}^T\mathbf{t}\right) = \exp\left\{\mathbf{t}^T\mu + \mathbf{t}^T\mathbf{A}\mathbf{A}^T\mathbf{t}/2\right\}$$

where we have $E[\mathbf{X}] = \mu$ and $\text{Cov}(\mathbf{X}) = \mathbf{A}\mathbf{A}^T$, which lead to

Definition 3.7. Random vector $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{V})$ iff the mgf satisfy

$$m_{\mathbf{X}}(\mathbf{t}) = \exp\left\{\mathbf{t}^T\mu + \mathbf{t}^T\mathbf{V}\mathbf{t}/2\right\}$$

Note that the shape of $\mathbf{X}$ is the same as μ, the we consider the transformation which turn $\mathbf{X}$ into different dimensions. Suppose $\mathbf{X} \sim N(\mu, \mathbf{V})$ where X is $p \times 1$ and $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$ where $\mathbf{a}$ is $q \times 1$ and thus $\mathbf{B}$ $q \times p$, then $\mathbf{Y} \sim N_q(\mathbf{a} + \mathbf{B}\mu, \mathbf{B}\mathbf{V}\mathbf{B}^T)$ since

$$\begin{aligned} m_{\mathbf{Y}}(\mathbf{t}) &= E\left[e^{\mathbf{t}^T\mathbf{Y}}\right] = E\left[e^{\mathbf{t}^T(\mathbf{a}+\mathbf{B}\mathbf{X})}\right] = e^{\mathbf{t}^T\mathbf{a}} \times m_{\mathbf{X}}(\mathbf{B}^T\mathbf{t}) \\ &= e^{\mathbf{t}^T\mathbf{a}} \times \exp\left\{\mathbf{t}^T\mathbf{B}\mu + \mathbf{t}^T\mathbf{B}\mathbf{V}\mathbf{B}^T\mathbf{t}/2\right\} \end{aligned}$$

Proposition 3.5. *If X is multivariate normal, then the joint distribution of any subset is multivariate normal.*

Proof. W.L.O.G, partition $\mathbf{X}, \boldsymbol{\mu}$ and $\mathbf{V}$ as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

Using $\mathbf{a} = \mathbf{0}$ and $\mathbf{B} = [\mathbf{I} \quad \mathbf{0}]$, we have $\mathbf{X_1} \sim N(\mu_1, \mathbf{V}_{11})$

□

Proposition 3.6 (Transformation of standard multivariate normal)**.** *If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{V}$ is nonsingular, then*

1. *A nonsingular matrix $\mathbf{A}$ exist s.t. $\mathbf{V} = \mathbf{A}\mathbf{A}'$*
2. *$\mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbf{I})$*
3. *The pdf is $(2\pi)^{-p/2}|\mathbf{V}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$.*

Proof. Covariance matrix must be semi positive definite, since $\mathbf{V}$ is nonsingular in this case, it's posdef. Employed Cholesky decomposition, we have 1. 2 is derived from last result and 3 is given by replacing $\mathbf{Z} = \mathbf{A^{-1}}(\mathbf{X} - \boldsymbol{\mu})$ in the pdf of standard case and note $|\det(\frac{\partial \mathbf{Z}}{\partial \mathbf{X}})| = |\det(\mathbf{A}^{-1})| = \det(V)^{-\frac{1}{2}}$.

□

Figure 3.1 shows examples of two densities of multivariate normal distributions.

Exercise 3.6 (Characterization of normal distribution)**.** Suppose $\mathbf{X}$ be random vector in $\mathbb{R}^n$. Show that $\mathbf{X}$ is multivariate normal iff every one-dimensional marginal $\langle \mathbf{X}, \boldsymbol{\theta} \rangle$ is normal.

Solution. Suppose $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then for any $\boldsymbol{\theta} \in \mathbb{R}^n$, we have

$$\langle \mathbf{Z}, \boldsymbol{\theta} \rangle = \boldsymbol{\theta}'\mathbf{Z} \sim \mathcal{N}(0, \boldsymbol{\theta}'\boldsymbol{\theta})$$

For the converse, suppose isotropic $\mathbf{Z_0}$ with zero mean, then

$$\begin{aligned} \mathbb{E}\,\boldsymbol{\theta}'\mathbf{Z_0} &= \boldsymbol{\theta}'\,\mathbb{E}\,\mathbf{Z_0} = 0 \\ \operatorname{Var}\boldsymbol{\theta}'\mathbf{Z_0} &= \mathbb{E}\,\boldsymbol{\theta}'\mathbf{Z_0}\mathbf{Z_0}'\boldsymbol{\theta} = \mathbb{E}\,\boldsymbol{\theta}'\boldsymbol{\theta} = \boldsymbol{\theta}'\boldsymbol{\theta} \end{aligned}$$

thus by Cram´er-Wold's theorem, we have

$$\boldsymbol{\theta}'\mathbf{Z_0} \overset{d}{=} \boldsymbol{\theta}'\mathbf{Z} \implies \mathbf{Z_0} \overset{d}{=} \mathbf{Z}$$

For general case, we just write $\mathbf{X}_{(\mathbf{0})} = \sqrt{\boldsymbol{\Sigma}}\mathbf{Z}_{(\mathbf{0})} + \boldsymbol{\mu}$ and then claim follows by direct algebra.

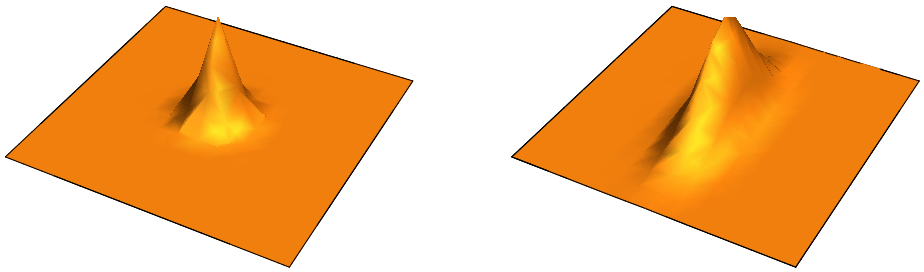

Figure 3.1: The densities of the isotropic distribution and a non-isotropic distribution

Exercise 3.7. Suppose $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, denote $\mathbf{X_u} := \langle \mathbf{X}, \mathbf{u} \rangle \sim \mathcal{N}(0, \|\mathbf{u}\|_2^2)$, show that

1. $\mathbb{E}\, \mathbf{X_u X_v} = \langle \mathbf{u}, \mathbf{v} \rangle$.
2. $\|X_u - X_v\|_2 = \|\mathbf{u} - \mathbf{v}\|_2$

Solution. For 1, note

$$\mathbb{E}\, \mathbf{u}'\mathbf{X}\mathbf{v}'\mathbf{X} = \mathbb{E}\, \mathbf{u}'\mathbf{X}\mathbf{X}'\mathbf{v} = \mathbf{u}'\, \mathbb{E}\, \mathbf{X}\mathbf{X}'\mathbf{v} = \mathbf{u}'\mathbf{v}$$

For 2, note

$$\mathbb{E}(\mathbf{u}'\mathbf{X} - \mathbf{v}'\mathbf{X})^2 = \mathbb{E}\left[(\mathbf{u} - \mathbf{v})'\mathbf{X}\right]^2 = \mathrm{Var}\left[(\mathbf{u} - \mathbf{v})'\mathbf{X}\right] = \|\mathbf{u} - \mathbf{v}\|_2^2$$

Exercise 3.8. Let $\mathbf{G} \in \mathbb{R}^{m \times n}$ be Gaussian random matrix where each entries are independent $\mathcal{N}(0, 1)$ $r.v.'s$. Let $\mathbf{u} \in \mathbb{R}^n$ be unit vector, then $\mathbf{Gu} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Moreover, $\mathbf{Gu}$ and $\mathbf{Gv}$ are independent if $\mathbf{u}$ and $\mathbf{v}$ are orthonormal.

Solution. Suppose $\mathbf{G} = (\mathbf{G_1}, \mathbf{G_2}, \dots, \mathbf{G_m})'$, where each $\mathbf{G_i}$ is isotropic with zero mean, then

$$\mathbf{Gu} = (\mathbf{G_1'u}, \mathbf{G_2'u}, \dots, \mathbf{G_m'u})'$$

where each coordinate is standard normal have been seen and thus distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$ by definition. Then

$$\mathrm{Cov}\left[\mathbf{Gu}, \mathbf{Gv}\right] = \mathbb{E}\, \mathbf{Guv}'\mathbf{G} = \mathbb{E}\, \mathbf{GG} = \mathbf{0}$$

3.4.3 Similarity of normal and spherical distributions

Recall the theorem 3.1, for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$, we have

$$\mathbb{P}\left\{\left|\|\mathbf{g}\|_2 - \sqrt{n}\right| \geq t\right\} \leq 2\exp(-ct^2)$$

for all $t \geq 0$. This observation suggests that the normal distribution should be quite similar to the uniform distribution on the sphere. Let us clarify the relation.

Proposition 3.7 (Normal and spherical distributions)**.** *Note we can represent* $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *in polar form as*

$$\mathbf{g} = r\boldsymbol{\theta}$$

where $r = \|\mathbf{g}\|_2$ *and* $\theta = \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$*. Show that*

1. r *and* $\boldsymbol{\theta}$ *are independent.*
2. $\boldsymbol{\theta} \sim Unif(S^{n-1})$

Proof. □

Thus the standard normal distribution in high dimensions is close to the uniform distribution on the sphere of radius $\sqrt{n}$:

$$\mathcal{N}(\mathbf{0}, \mathbf{I}) \approx \text{Unif}(\sqrt{n}S^{n-1})$$

3.4.4 Frames

For an example of an extremely discrete distribution, consider a **coordinate random vector**:

$$\mathbf{X} \sim \text{Unif}\left(\sqrt{n}\mathbf{e_i}\right)_{i=1}^{n}$$

Which often known as "the worst" distribution while gaussian is "the best".

A general class of discrete, isotropic distributions arises in the area of signal processing under the name of **frames**.

Definition 3.8. A **frame** is a set of vectors $\{\mathbf{u_i}\}_{i=1}^{N} \subset \mathbb{R}^n$ which obeys an approximate Parseval's identity, *i.e.*, there exist $c, C > 0$ called **frame bounds** $s.t. \forall \mathbf{x} \in \mathbb{R}^n$

$$c\|\mathbf{x}\|_2^2 \leq \sum_{i=1}^{N} \langle \mathbf{u_i}, \mathbf{x} \rangle^2 \leq C\|x\|_2^2$$

If $c = C$ the frame is **tight**.

Proposition 3.8. $\{\mathbf{u_i}\}_{i=1}^{N}$ *is tight with constant* c *iff*

$$\sum_{i=1}^{N} \mathbf{u_i}\mathbf{u_i}' = c\mathbf{I}$$

Proof. Construct $\mathbf{U} \in \mathbb{R}^{N \times n}$ by $\mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}, \dots, \mathbf{u_N})'$, then $\sum_{i=1}^{N} \mathbf{u_i}\mathbf{u_i}' = \mathbf{U}'\mathbf{U}$. Then claim follows by noting

$$\begin{aligned}
\mathbf{U}'\mathbf{U} = c\mathbf{I} &\iff \mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{x} = c\mathbf{x}'\mathbf{x} \quad \forall \mathbf{x} \\
&\iff \|\mathbf{U}\mathbf{x}\|^2 = c\|\mathbf{x}\|^2 \quad \forall \mathbf{x} \\
&\iff \sum_{i=1}^{N} \langle \mathbf{u_i}, \mathbf{x} \rangle^2 = c\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x}
\end{aligned}$$

□

Clearly, any orthonormal basis in $\mathbb{R}^n$ is a tight frame. But the independence requirement is not necessary, the Mercedez-Benz frame is an example. $N \geq n$ is still essential.

Now we are ready to connect the concept of frames to probability.

Proposition 3.9 (Tight frames and isotropic distributions)**.** *Suppose a tight frame* $\{\mathbf{u_i}\}_{i=1}^m \subset \mathbb{R}^n$ *with frame bounds* c*. Let* $\mathbf{X}$ *be a random vector that is uniformly distributed in which, i.e.*

$$\mathbf{X} \sim \mathit{Unif}(\mathbf{u_i} : i = 1, \dots, m)$$

then $\sqrt{\frac{m}{c}}\mathbf{X}$ *is an isotropic random vector in* $\mathbb{R}^n$*.*

On the other hand, suppose $\mathbf{X}$ *is isotropy with discrete distribution:* $\mathbb{P}\{\mathbf{X} = \mathbf{x_i}\} = p_i, \forall i = 1, 2, \dots, m$*. Then* $\mathbf{u_i} := \sqrt{p_i}\mathbf{x_i}$ *form a tight frame in* $\mathbb{R}^n$ *with bound* c*.*

Proof. Construct $\mathbf{U}$ follow the way above mentioned, WLOG, we can assume $m = c$ as $\sqrt{\frac{m}{c}}\mathbf{U}$ give a tight frame with constant m. Then claim follows by direct algebra:

$$\mathbb{E}\,\mathbf{X}\mathbf{X}' = \frac{1}{m}\sum_{i=1}^m \mathbf{u_i}\mathbf{u_i}' = \frac{1}{m}m\mathbf{I} = \mathbf{I}$$

On the other hand, employ $\mathbf{X}$ is isotropy, we have

$$\mathbb{E}\,\mathbf{X}\mathbf{X}' = \sum_{i=1}^m p_i\mathbf{x_i}\mathbf{x_i}' = \mathbf{I}$$

write $p_i\mathbf{x_i}\mathbf{x_i}' = (\sqrt{p_i}\mathbf{x_i})(\sqrt{p_i}\mathbf{x_i})'$, then the claim follows.

□

3.4.5 Isotropic convex sets

Now suppose $\mathbf{X}$ are uniformly distributed in a convex body K, which is compact and with $K^\circ \neq \varnothing$. WLOG, assume $\mathbb{E}\,\mathbf{X} = 0$(some translation on K achieve this) and $\operatorname{Cov}\mathbf{X} =: \mathbf{\Sigma}$.

Recall that $\mathbf{Z} := \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{X}$ is isotropic, $\mathbf{Z}$ is uniformly distributed in $\mathbf{\Sigma}^{-\frac{1}{2}}K$. Thus we found a operator $T := \mathbf{\Sigma}^{-\frac{1}{2}}$ which makes the uniform distribution on TK isotropic.

3.5 Sub-gaussian distributions in higher dimensions

Now we generalize sub-gaussian distributions to higher dimensions. Recall the characterization of multivariate normal distributions 3.6, it's natural to define

Definition 3.9 (Sub-gaussian random vectors)**.** A random vector X in $\mathbb{R}^n$ is called sub-gaussian if the marginal $\mathbf{u}'\mathbf{X}$ are sub-gaussian for all $\mathbf{u} \in \mathbb{R}^n$. The sub-gaussian norm is defined as

$$\|X\|_{\psi_2} = \sup_{\mathbf{u}\in S^{n-1}} \|\mathbf{u}'\mathbf{X}\|_{\psi_2}$$

Clearly, a random vector with independent, sub-gaussian coordinates is sub-gaussian by proposition 2.9, in such case, we have

$$\|\mathbf{X}\|_{\psi_2} \leq C \max_i \|X_i\|_{\psi_2}$$

as $\mathbf{u}$ should be $(0, 0, \dots, 1, \dots, 0)$.

If those coordinates are not independent, the $\|\cdot\|_{\psi_2}$ is not bounded.

Exercise 3.9. Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector with sub-gaussian coordinates X_i, show that X is a sub-gaussian.

Nevertheless, find an example of a random vector X with

$$\|\mathbf{X}\|_{\psi_2} \gg \max_i \|X_i\|_{\psi_2}$$

Proposition 3.10. *Let $\mathbf{X} \in \mathbb{R}^n$ be a r.v. with sub-gaussian coordinates X_i, then $\mathbf{X}$ is a sub-gaussian r.v..*

Proof. Note that $\langle x, \mathbf{X} \rangle = \sum_{i=1}^n x_i X_i$ where X_i are sub-gaussian distribution, so $\langle x, \mathbf{X} \rangle$ are sub-gaussian for every $x \in \mathbb{R}^n$ since the property of vector space.

□

3.5.1 Gaussian and Bernoulli distributions

Clearly, multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is sub-gaussian, moreover, $\mathcal{N}(0, \mathbf{I})$ has sub-gaussian norm of order $O(1)$:

$$\|\mathbf{X}\|_{\psi_2} \leq C$$

as all marginal $\mathbf{e}'\mathbf{X}$ are $\mathcal{N}(0, 1)$ as $\mathbf{e}'\mathbf{e} = \mathbf{1}$. Similarly, the multivariate symmetric Bernoulli distribution also has $O(1)$ norm.

3.5.2 Discrete distributions

Recall the "worst" distribution, coordinate distribution, where $\mathbf{X} \sim \text{Unif}\{\sqrt{n}\mathbf{e_i} : i = 1, 2, \dots, n\}$. It's sub-gaussian as it supported in a finite set, however, it has a larger norm:

Proposition 3.11. *Show that for coordinate distribution X:*

$$\|X\|_{\psi_2} \asymp \sqrt{\frac{n}{\ln n}}$$

Proof. Note $\left(\frac{nu_1^2}{t^2}, \frac{nu_2^2}{t^2}, \dots, \frac{nu_n^2}{t^2}\right)'$ is majored by $(\sum_{i=1}^n \frac{nu_i^2}{t^2}, 0, \dots, 0)'$, Karamata's inequality yield

$$\mathbb{E} \exp\left(\frac{\mathbf{u}'\mathbf{X}}{t^2}\right) = \frac{1}{n}\sum_{i=1}^n \left[\exp\left(\frac{nu_i^2}{t^2}\right)\right] \leq \frac{1}{n}\left[\exp\left(n\sum_{i=1}^n \frac{u_i^2}{t^2}\right) + n - 1\right]$$

Note $\|\mathbf{X}\|_{\psi_2}$ is just $\|\mathbf{e_i'X}\|_{\psi_2}$ for some i, and now the equity of Karamata's inequality holds. To find $\|\mathbf{X}\|_{\psi_2}$, we have

$$\frac{1}{n}\left[\mathbb{E}\exp\left(\frac{n}{t^2}\right)+(n-1)\right]\leq 2$$

Substitute t by $\|\mathbf{X}\|_{\psi_2}$, solve it lead to $\|\mathbf{X}\|_{\psi_2}=\sqrt{\frac{n}{\ln(n+1)}}\asymp\sqrt{\frac{n}{\ln n}}$.

□

More generally, discrete distributions do not make nice sub-gaussian distributions, unless they are supported on exponentially large sets:

Proposition 3.12. *Let* $\mathbf{X}$ *be isotropic supported in a finite set* $T\subset\mathbb{R}^n$. *Show that if* $\|X\|_{\psi_2}\ll 1$, *then* $|T|\geq e^{cn}$.

3.5.3 Uniform distribution on sphere

Good sub-gaussian random vectors not necessary have independent coordinates.

Theorem 3.2 (Uniform distribution on the sphere is sub-gaussian)**.** *Suppose* $\mathbf{X}\sim$ *Unif*$(\sqrt{n}S^{n-1})$, *then* $\mathbf{X}$ *is sub-gaussian and* $\|X\|_{\psi_2}\leq C$.

Proof. As proposition 3.7, we can represent X as

$$\mathbf{X}=\sqrt{n}\frac{\mathbf{g}}{\|\mathbf{g}\|_2}$$

where $\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$. It's sufficient to show that $\boldsymbol{\theta}'\mathbf{X}$ is sub-gaussian for all $\boldsymbol{\theta}$ and WLOG we may assume $\boldsymbol{\theta}=(1,0,\dots,0)$. In which case, we want to bounded

$$p(t):=\mathbb{P}\left\{|X_1|\geq t\right\}=\mathbb{P}\left\{\frac{|g_1|}{\|\mathbf{g}\|_2}\geq\frac{t}{\sqrt{n}}\right\}$$

As $\|\mathbf{g}\|_2-\sqrt{n}$ is sub-gaussian with order $O(1)$ by theorem 3.1, we reduce this to bounding $\mathbb{P}\left\{|g_1|\geq t\right\}$, and it's clearly sub-gaussian again.

Explicitly, take $\varepsilon:=\left\{\|\mathbf{g}\|_2\geq\frac{\sqrt{n}}{2}\right\}$, then

$$\begin{aligned}\mathbb{P}\left\{\varepsilon^c\right\}&=\mathbb{P}\left\{\|\mathbf{g}\|_2<\frac{\sqrt{n}}{2}\right\}=\mathbb{P}\left\{\|\mathbf{g}\|_2-\sqrt{n}<-\frac{\sqrt{n}}{2}\right\}\\&\leq\mathbb{P}\left\{\left|\|\mathbf{g}\|_2-\sqrt{n}\right|>\frac{\sqrt{n}}{2}\right\}\leq 2\exp\left(-cn\right)\end{aligned}$$

It follows that

$$\begin{aligned}p(t)&\leq\mathbb{P}\left\{\frac{|g_1|}{\|\mathbf{g}\|_2}\geq\frac{t}{\sqrt{n}}\cap\varepsilon\right\}+\mathbb{P}\,\varepsilon^c\\&\leq\mathbb{P}\left\{|g_1|\geq\frac{t}{2}\right\}+2\exp(-cn)\\&\leq 2\exp(-t^2/8)+2\exp(-cn)\end{aligned}$$

Note when $|X_1| \leq \|\mathbf{X}\|_2 = \sqrt{n}$ and that avoid $t < \sqrt{n}$. For $t \leq \sqrt{n}$, we clearly have $p(t) \leq C\exp(-c't^2)$ as desired. The second statement follows by all marginal distribution of sphere are identical.

□

Furthermore, we can extend sphere to a ball then claim remains true:

Proposition 3.13 (Uniform distribution on the Euclidean ball)**.** *Suppose* $\mathbf{Y} \sim$ *Unif*$(B(0, \sqrt{n}))$*, then it's sub-gaussian with* $O(1)$ *norm.*

Proof. Let $r := \frac{\|\mathbf{Y}\|_2}{\sqrt{n}} \leq 1$ and $\mathbf{X} := \frac{\mathbf{Y}}{\mathbf{r}}$, then $\mathbf{X} \sim \text{Unif}\left(\sqrt{n}S^{n-1}\right)$. By theorem 3.2, we have $\|\mathbf{X}\|_{\psi_2} \leq C$. Note

$$\mathbb{E}\exp\left(\frac{\langle \mathbf{Y}, \boldsymbol{\theta}\rangle^2}{t^2}\right) = \mathbb{E}\exp\left(\frac{r^2\langle \mathbf{X}, \boldsymbol{\theta}\rangle^2}{t^2}\right) \leq \mathbb{E}\exp\left(\frac{\langle \mathbf{X}, \boldsymbol{\theta}\rangle^2}{t^2}\right)$$

thus $\|\mathbf{Y}\|_{\psi_2} \leq \|\mathbf{X}\|_{\psi_2} \leq C$.

□

Theorem 3.3 (Projective limit theorem)**.** *If* $X \sim$ *Unif*$(\sqrt{n}S^{n-1})$*, then for any* $\boldsymbol{\theta} \in \mathbb{R}^n$*, as* $n \to \infty$*,*

$$\boldsymbol{\theta}'\mathbf{X} \xrightarrow{d} \mathcal{N}(0, 1)$$

Thus we can view theorem 3.2 as a concentration version of the Projective Limit Theorem, in the same sense as Hoeffding's inequality is a concentration version of the classical CLT.

3.5.4 Uniform distribution on convex sets

Related to Isotropic convex sets, we consider uniform distribution on a convex body K, as we seen, it can be an isotropy. Now is $\mathbf{X}$ always sub-gaussian?

Proposition 3.14. *Suppose a* ℓ_1 *norm ball in* $\mathbb{R}^n$ *and* $\mathbf{X_n} \sim$ *Unif*(K)*.*

$$K := \left\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 \leq r\right\}$$

1. $\mathbf{X}$ *is isotropic for some* $r \asymp n$*.*
2. $\|\mathbf{X}\|_{\psi_2}$ *can not bounded by some* C *as* n *grows.*

Proof. 1. Let $\mathbf{X} \sim \text{Unif}(K)$, then it is clear that

$$\mathbb{E}\, X_i = 0 \text{ and } \mathbb{E}\, X_i X_j = 0 \text{ for } i \neq j$$

for the sphere symmetry and to prove it is isotropic, just need to show that $\mathbb{E}\, X_i^2 = 1$ for each i. Note that

$$\mathbb{E}\, X_i^2 = \int_0^\infty \mathbb{P}\left\{X_i^2 \geq t\right\} dt = \int_0^\infty \mathbb{P}\left\{|X_i| \geq x\right\} \cdot 2x dx$$

Consider that

$$\mathbb{P}\left\{|X_i| \geq x\right\} = \frac{(r-x)^n}{r^n}$$

since the volume of the ball, and the probability vanishes when $x > r$

$$\mathbb{E}\, X_i^2 = \int_0^r 2x \cdot \frac{(r-x)^n}{r^n} dx = \frac{2r^2}{n^2+3n+2}$$

when $\mathbb{E}\, X_i^2 = 1$, we have

$$r = \sqrt{\frac{n^2+3n+2}{2}} = \Omega(n)$$

2. Consider the L^p norm of X_i.

$$\|X_i\|_{L^p}^p = \int_0^\infty p x^{p-1}\, \mathbb{P}\left\{|X_i| > x\right\} dx = p r^p \frac{\Gamma(p)\Gamma(n+1)}{\Gamma(n+p+1)}$$

so when $n \to \infty$ and $p \to \infty$, we have

$$\|X_i\|_{L^p} = \frac{r}{n} \cdot \Gamma(p)^{1/p} = \Omega(p)$$

so there does not exist a constant C *s.t.* $\|X_i\|_{L^p} \le C\sqrt{p}$ for any n.

□

3.6 Application: Grothendieck's inequality and semidefinite programming

Theorem 3.4 (Grothendieck's inequality)**.** *Suppose* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, assume that, for any* $x_i, y_i \in \{-1, 1\}$*, we have*

$$\left|\sum_{i,j} a_{ij} x_i y_j\right| \le 1$$

then, for any Hilbert space $\mathcal{H}$ *and any normal vectors* $u_i, v_j \in \mathcal{H}$*, we have*

$$\left|\sum_{i,j} a_{ij} \langle u_i, v_j\rangle\right| \le K$$

where K *is an absolute number.*

Remark. The assumption can be equivalently stated as

$$\left|\sum_{i,j} a_{ij} x_i y_i\right| \le \max|x_i| \cdot \max|y_j|$$

and the conclusion can be equivalently stated as

$$\left|\sum_{i,j} a_{ij} \langle u_i, v_j\rangle\right| \le K \max\|u_i\| \cdot \max\|v_j\|$$

for any $u_i, v_j \in \mathcal{H}$.

Proof. We show that $K \leq 288$.

Step 1. Choose $K = K(\mathbf{A})$ be the smallest number that makes conclusion valid for given $\mathbf{A}$, we can always do that as $K = \sum_{ij} |a_{ij}|$ is one of them. Our goal is to show that $K(\mathbf{A})$ does actually not depend on $\mathbf{A}$.

Note the Hilbert space are indifferential with the space spanned by the $m+n$ vectors $\{u_i, v_j\}$ and thus we can treated it as finite dimensional space and each norm are equivalent. It follows that we can specify $\mathcal{H} = \mathbb{R}^{m+n}$ equipped $\|\cdot\|_2$. Note the range of $\sum a_{ij}\mathbf{u_i'v_j}$ should be closed, we can fix $\mathbf{u_i}, \mathbf{v_j} \in \mathbb{R}^{m+n:=N}$ which realize K:

$$\sum_{i,j} a_{ij}\langle \mathbf{u_i}, \mathbf{v_j}\rangle = K$$

Step 2 Let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then let $U_i := \langle g, u_i\rangle$ and $V_j := \langle \mathbf{g}, \mathbf{v_j}\rangle$, we have

$$U_i \overset{d}{=} V_j \overset{d}{=} \mathfrak{Z} \text{ and } \mathbb{E}\, U_iV_j = \langle \mathbf{u_i}, \mathbf{v_j}\rangle$$

Thus $K = \mathbb{E}\sum_{i,j} a_{ij}U_iV_j$. Let $\mathbf{u} := (U_1, U_2, \dots, U_m)'$ and $\mathbf{v} := (V_1, V_2, \dots, V_n)'$ and similarly define $\mathbf{x}$ and $\mathbf{y}$, the assumption become $\mathbf{x'Ay} = 1$ while conclusion become $\mathbb{E}\,\mathbf{u'Av} \leq K$. If we can bound U_i by an constant r, then assumption yield $K^2 \leq r^2$.

Step 3 To this end, truncate $\mathfrak{Z}$ by an constant $r \geq 1$ to $\mathfrak{Z}^- = \mathfrak{Z}\mathbf{1}_{|\mathfrak{Z}|\leq r}$ and $\mathfrak{Z}^+ = \mathfrak{Z}\mathbf{1}_{|\mathfrak{Z}|>r}$, recall proposition 2.2. We have

$$\|\mathfrak{Z}^+\|_2^2 \leq 2\left(r + \frac{1}{r}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{r^2}{2}} < \frac{4}{r^2}$$

Step 4

Truncate each U_i and V_j similar to $\mathfrak{Z}$, we have

$$\begin{aligned} K &= \mathbb{E}\,\mathbf{u}'^{-}\mathbf{Av}^{-} + 2\,\mathbb{E}\,\mathbf{u}'^{+}\mathbf{Av}^{-} + \mathbb{E}\,\mathbf{u}'^{+}\mathbf{Av}^{+} \\ &\leq 2r + \frac{4K}{r} + \frac{4K}{r^2} \end{aligned}$$

where the inequality follows by all Hilbert space share the same K and Cauchy-Schwarz inequality. Choose r to minimize K, we have

$$K \leq 63.2865$$

□

When $\mathbf{A}$ is self-adjoint, we have

Proposition 3.15. *Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ is either positive-semidefinite or has zero diagonal, assume that, for any $x_i, y_i \in \{-1, 1\}$, we have*

$$\left|\sum_{i,j} a_{ij}x_ix_j\right| \leq 1$$

then, for any Hilbert space $\mathcal{H}$ and any normal vectors $u_i, v_j \in \mathcal{H}$, we have

$$\left|\sum_{i,j} a_{ij}\langle u_i, v_j\rangle\right| \leq 2K$$

where K is the Grothendieck constant.

3.6.1 Semidefinite programming

Definition 3.10. A semidefinite program(SDP) is an optimization problem that:

$$\begin{aligned} \max \quad & \langle \mathbf{A}, \mathbf{X} \rangle \\ s.t. \quad & \mathbf{X} \geq 0 \\ & \langle \mathbf{B_i}, \mathbf{X} \rangle = b_i, \forall i \end{aligned}$$

where $\mathbf{A}, \mathbf{B_i} \in \mathbb{R}^{n \times n}$.

Note that semidefinite matrices form a convex set and SDP is a convex program. It's a good news since there are variety of computationally efficient solvers available for convex programs. That motivate us to relax computationally hard problem to SDP. Such as:

$$\begin{aligned} \max \quad & \mathbf{x}'\mathbf{A}\mathbf{x} \\ s.t. \quad & \mathbf{x} = (x_1, x_2, \dots, x_n) \\ & x_i = \pm 1, \forall i \end{aligned} \tag{3.1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is self-adjoint. This is an integer optimization problem and known as NP-hard.

Nonetheless, we can relax this problem to a SDP by replacing $x_i = \pm 1$ by their higher-dimensional analogs-unit vectors $\mathbf{x_i} \in \mathbb{R}^n$. Now the problem is:

$$\begin{aligned} \max \quad & \sum_{i,j}^{n} A_{ij} \langle \mathbf{x_i}, \mathbf{x_j} \rangle \\ s.t. \quad & \|X_i\|_2 = 1, \forall i \end{aligned}$$

That is equivalent to the following SDP:

$$\begin{aligned} \max \quad & \langle \mathbf{A}, \mathbf{X} \rangle \\ s.t. \quad & \mathbf{X} \geq 0 \\ & X_{ii} = 1, \forall i \end{aligned} \tag{3.2}$$

As we can let $\mathbf{X}$ be the Gram matrix for vectors $\{\mathbf{x_i}\}_{i=1}^{n}$.

Grothendieck's inequality guarantees the accuracy of semidefinite relaxations:

Theorem 3.5. *Suppose $\mathbf{A} \geq \mathbf{0} \in \mathbb{R}^{n \times n}$, let INT($\mathbf{A}$) denote the maximum in the integer optimization problem (3.1) and SDP($\mathbf{A}$) denote the maximum in the SDP (3.2), then*

$$INT(\mathbf{A}) \leq SDP(\mathbf{A}) \leq 2K \cdot INT(\mathbf{A})$$

where K is the Grothendieck constant.

Proof. The first inequality follows with taking $\mathbf{x_i} = \mathbf{e_i}$ and the second follows from proposition 3.15.

□

Exercise 3.10. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, consider following optimization problem:

$$\begin{aligned} \max \quad & \sum_{i,j} A_{ij}\langle \mathbf{x_i}, \mathbf{y_j}\rangle \\ s.t. \quad & \|x_i\|_2 = \|x_j\|_2, \forall i,j \end{aligned}$$

over $\mathbf{x_i} \in \mathbb{R}^k$. Formulate this problem as a semidefinite program.

Proof. Note the target function is $\langle \tilde{\mathbf{A}}, \mathbf{ZZ}'\rangle$ where $\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ and $\mathbf{X}$ are constructed by $\mathbf{x_i}$. Then $\mathbf{ZZ}'$ is nothing more than a positive-semidefinite matrix whose diagonal entries equal 1.

□

Chapter 4

Random matrices

4.1 Preliminaries on matrices

4.1.1 Singular value decomposition

Note that $\mathbf{AA}'$ and $\mathbf{A}'\mathbf{A}$ share the same nonzero eigenvalues, and thus the singular values defined as:

$$s_i(\mathbf{A}) = \sqrt{\lambda_i(\mathbf{AA}')} = \sqrt{\lambda_i(\mathbf{A}'\mathbf{A})}$$

then we have singular value decomposition(SVD):

$$\mathbf{A} = \sum_{i=1}^{n} s_i \mathbf{u_i} \mathbf{v_i}' = \mathbf{U\Sigma V}^* \tag{4.1}$$

For self-adjoint $\mathbf{A}$, they have spectrum decomposition and $s_i(\mathbf{A}) = |\lambda_i(\mathbf{A})|$.

The quadratic form is something of the form $\mathbf{x}'\mathbf{Ax}$ as a function of $\mathbf{x} \neq \mathbf{0}$, where $\mathbf{A} \in \mathbb{R}^{m\times m}$ is symmetric. To avoid effect of scale, we often use **Rayleigh quotient**:

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\langle \mathbf{Ax}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}$$

Theorem 4.1. *For Hermitian* $\mathbf{A}$*,* $R(\mathbf{x}, \mathbf{A})$ *take minimum* λ_n *in* $S_{\mathbf{A}}(\lambda_n)$ *while maximum* λ_1 *in* $S_{\mathbf{A}}(\lambda_1)$*. Where* $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$*.*

Consequently, we have:

Theorem 4.2 (Courant–Fischer min–max theorem)**.** *Suppose* $\mathbf{A}$ *is Hermitian and* λ_i *are decreasing, then*

$$\lambda_i(\mathbf{A}) = \max_{dim\,\mathbf{U}=i} \min_{\mathbf{x}} R_{\mathbf{A}}(\mathbf{x})$$

Immediately, we have

$$s_i(\mathbf{A}) = \max_{\dim \mathbf{U}=i} \min_{\mathbf{x}} \|Ax\|_2$$

Proposition 4.1. *For given SVD (4.1),if* $\mathbf{A}$ *is invert, we have*

$$\mathbf{A^{-1}} = \sum_{i=1}^{n} \frac{1}{s_i} \mathbf{v_i u_i'}$$

In fact, for any $\mathbf{A}$*, we have*

$$\mathbf{A}^+ = \mathbf{V\Sigma}^+\mathbf{U}^*$$

Recall that we can see $\mathbf{A} \in \mathbb{K}^{m\times n}$ as an operator

$$\|\mathbf{A}\| = \max_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} = \max_{S^{n-1}} \|\mathbf{Ax}\|_2 = \max_{\mathbf{x}\in S^{n-1}, \mathbf{y}\in S^{m-1}} \|\mathbf{A}\|$$

Moreover, for symmetric matrices one can take $\mathbf{x} = \mathbf{y}$ in this formula. So we have $s_1(\mathbf{A}) = \|\mathbf{A}\|$, consequently,

$$s_n(\mathbf{A}) = \frac{1}{\|\mathbf{A}^+\|}$$

If now we see $\mathbf{A}$ as element in matrices space, the inner product defined as $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}\,(\mathbf{A'B})$.

Thus we have the Frobenius norm:

$$\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}\,(\mathbf{A}^*\mathbf{A})}$$

In terms of singular values, we have

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{r} s_i(\mathbf{A})^2}$$

Let's now compare the two norms, let $s = (s_n)$, we have

$$\|\mathbf{A}\| = \|s\|_\infty, \|\mathbf{A}\|_F = \|s\|_2$$

thus we have

$$\|\mathbf{A}\| \le \|\mathbf{A}\|_F \le \sqrt{r}\,\|\mathbf{A}\|$$

Proposition 4.2. *For any matrix* $\mathbf{A}$*, we have*

$$s_i \le \frac{1}{\sqrt{i}} \|\mathbf{A}\|_F$$

Proof. As $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n s_i^2}$, we have

$$s_i \le \sqrt{\sum_{j=1}^{i} \frac{s_j^2}{i}} \le \frac{1}{\sqrt{i}} \sqrt{\sum_{j=1}^{n} s_j^2} = \frac{1}{\sqrt{i}} \|\mathbf{A}\|_F$$

□

4.1.2 Low-rank approximation

Suppose we want to approximate a given matrix $\mathbf{A}$ by matrices with lower rank $\mathbf{A_k}$ where $k < r$. The **Eckart-Young-Mirsky's theorem** say the $\mathbf{A_k}$ is obtained by truncating the singular value decomposition of $\mathbf{A}$ at the kth term:

$$\mathbf{A}_k = \sum_{i=1}^{k} s_i \mathbf{u_i} \mathbf{v_i^*}$$

and $\mathbf{A_k}$ is called the best rank k approximation of $\mathbf{A}$.

Proposition 4.3 (Best rank k approximation)**.** *Let $\mathbf{A}_k$ be the best rank k approximation of a matrix $\mathbf{A}$, then*

$$\|\mathbf{A} - \mathbf{A_k}\|^2 = s_{k+1}$$

and

$$\|\mathbf{A} - \mathbf{A_k}\|_F = \sqrt{\sum_{i=k+1}^{n} s_i \mathbf{u_i} \mathbf{v_i^*}}$$

4.1.3 Approximate isometries

By theorem 4.1, for any matrix $\mathbf{A}$, we have

$$s_n \|x - y\| \leq \|Ax - Ay\| \leq s_1 \|x - y\|$$

thus operator $\mathbf{A}$ can only change distance by a factor that between s_n and s_1. That implies isometry matrix can only have 1 as singular value:

Proposition 4.4 (Isometries)**.** *Suppose $\mathbf{A} \in \mathbb{K}^{m \times n}$ and $m > n$, TFAE:*

1. $\mathbf{A}^*\mathbf{A} = \mathbf{I}$
2. $\mathbf{A}\mathbf{A}^*$ *is an orthogonal projection.*
3. $\mathbf{A}$ *is isometry.*
4. $s_n(\mathbf{A}) = s_1(\mathbf{A}) = 1$.

Quite often $\mathbf{A}^*\mathbf{A}$ only approximate $\mathbf{I}$, then $\mathbf{A}$ is regard as an approximate isometry.

Lemma 4.1 (Approximate isometries)**.** *Let $\mathbf{A} \in \mathbb{K}^{m \times n}$ with $\delta > 0$, suppose*

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{I}\| \leq \max(\delta, \delta^2)$$

then

$$(1 - \delta) \|\mathbf{x}\| \leq \|\mathbf{A}\mathbf{x}\| \leq (1 + \delta) \|\mathbf{x}\|$$

consequently, all the singular values are between $1 \pm \delta$.

Proof. WLOG, assume $\|\mathbf{x}\| = 1$. Then

$$\delta \vee \delta^2 \geq |\langle (\mathbf{A}^*\mathbf{A} - \mathbf{I})\mathbf{x}, \mathbf{x} \rangle| = \left| \|\mathbf{A}\mathbf{x}\|^2 - 1 \right|$$

and note $|z - 1| \vee |z - 1|^2 \leq \left|z^2 - 1\right|$, it follows that

$$\left| \|\mathbf{A}\mathbf{x}\|^2 - 1 \right| \leq \delta$$

□

The following partial converse holds:

Proposition 4.5. *If* $(1-\delta) \leq s_i(\mathbf{A}) \leq 1+\delta$*, then*

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{I}\| \leq 3\left(\delta \vee \delta^2\right)$$

Proof. Let $\|\mathbf{x}\| = 1$, we have $\|\mathbf{A}\mathbf{x}\| \in [1-\delta, 1+\delta]$ and

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{I}\| \geq \left|\|Ax\|^2 - 1\right| \geq \left|(1\pm\delta)^2 - 1\right| = \left|\delta^2 \pm 2\delta\right| \leq 3(\delta \vee \delta^2)$$

□

4.2 Nets, covering numbers and packing numbers

Proposition 4.6. *Let* N *be a maximal* ε *separated subset of* K*, then* N *is an* ε*-net of* K*.*

The covering and packing numbers are essentially equivalent:

Proposition 4.7. *For* $K \subset \tau$ *with metric* d*, we have*

$$P(K, 2\varepsilon) \leq N(K, \varepsilon) \leq P(K, \varepsilon)$$

Proposition 4.8. *The exterior covering numbers of* K *are equivalent to covering number:*

$$N^{ext}(K, \varepsilon) \leq N(K, \varepsilon) \leq N^{ext}\left(K, \frac{\varepsilon}{2}\right)$$

Proposition 4.9. *If* $A \subset B$*, then*

$$N(A, \varepsilon) \leq N\left(B, \frac{\varepsilon}{2}\right)$$

4.2.1 Volume

Theorem 4.3 (Covering numbers and volume)**.** *Let* $K \subset \mathbb{R}^n$ *and* $\varepsilon > 0$*, then*

$$\frac{\mu(K)}{\mu(\varepsilon B)} \leq N(K, \varepsilon) \leq P(K, \varepsilon) \leq \frac{\mu(K + (\frac{\varepsilon}{2})B)}{\mu\left(\frac{\varepsilon}{2}B\right)}$$

where B *is the unit ball in* $(\mathbb{R}^n, \|\cdot\|_2)$*.*

Consequently, the covering numbers are exponential in n:

Corollary 4.1. *For any* $\varepsilon > 0$*:*

$$\left(\frac{1}{\varepsilon}\right)^n \leq N(B, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^n$$

Proof. Let $K = B$ in theorem 4.3, the lower bound follows by noting $\mu(\varepsilon B) = \varepsilon^n \mu(B)$. The upper bound follows from

$$\frac{\mu(B + (\frac{\varepsilon}{2}B))}{\mu\left(\frac{\varepsilon}{2}B\right)} = \frac{(1+\frac{\varepsilon}{2})^n}{(\frac{\varepsilon}{2})^n} = \left(\frac{2}{\varepsilon} + 1\right)^n$$

□

Example 4.1. The Hamming cube $H = \{0,1\}^n$ consists of all binary strings of length n and we can identify each element x of them as a function from $1, 2, \dots, n$ to $\{0,1\}$, then

$$d(x,y) := |\{i : x(i) \neq y(i)\}|$$

define a metric on H. On space $(\mathcal{H}, d)$, we have

Proposition 4.10 (Covering and packing numbers of the Hamming cube)**.** *For every integer $m \in [0,n]$, we have*

$$\frac{2^n}{\sum_{i=0}^{m} \binom{n}{k}} \leq N(H,m) \leq P(H,m) \leq \frac{2^n}{\sum_{i=0}^{\lfloor \frac{m}{2} \rfloor} \binom{n}{k}}$$

4.3 Application: error correcting codes

4.4 Upper bounds on random sub-gaussian matrices

Recall the definition of $\mathbf{A}$, we fixed $\mathbf{x}$ on the unit sphere. In fact, it's enough for an ε net of it.

Proposition 4.11 (Computing the operator norm on a net)**.** *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\varepsilon \in [0,1)$, then for any ε net N of the sphere S^{n-1}, we have*

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

Proof. The lower bound is trivial and, to prove the upper bound, fix a vector $\mathbf{x} \in S^{n-1}$ for which

$$\|\mathbf{A}\| = \|\mathbf{A}\mathbf{x}\|_2$$

and $\mathbf{x_0} \in N$ near $\mathbf{x}$ at most ε, then we have

$$\|\mathbf{A}\mathbf{x_0}\| \geq \|\mathbf{A}\mathbf{x}\| - \|\mathbf{A}(\mathbf{x} - \mathbf{x_0})\| \leq \|\mathbf{A}\| (1 - \varepsilon)$$

□

Proposition 4.12. *For any $\mathbf{x} \in \mathbb{R}^n$, then*

$$\sup_{\mathbf{y} \in N} \langle \mathbf{x}, \mathbf{y} \rangle \leq \|x\| \leq \frac{1}{1-\varepsilon} \sup_{\mathbf{y} \in N} \langle \mathbf{x}, \mathbf{y} \rangle$$

Proof. To achieve upper bound, let $\mathbf{x_0} \in N$ near $\frac{\mathbf{x}}{\|\mathbf{x}\|}$, then

$$\varepsilon^2 \geq \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|} - x_0, \frac{x}{\|\mathbf{x}\|} - x_0 \right\rangle = 2 - 2 \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, x_0 \right\rangle$$

thus

$$\|\mathbf{x}\| \leq \frac{2}{2-\varepsilon^2} \langle \mathbf{x}, \mathbf{x_0} \rangle \leq \frac{1}{1-\varepsilon} \langle \mathbf{x}, \mathbf{x_0} \rangle$$

when $\varepsilon \leq 1$. And the lower bound is simply follows from Cauchy-Schwarz inequality:

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \sqrt{\|\mathbf{x}\|^2 \|\mathbf{y}\|^2} = \|\mathbf{x}\|$$

□

Proposition 4.13. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\varepsilon \in [0, \frac{1}{2})$, N a ε net of S^{n-1} and M of S^{m-1}, then*

$$\sup_{x\in N, y\in M} \langle \mathbf{Ax}, \mathbf{y}\rangle \le \|\mathbf{A}\| \le \frac{1}{1-2\varepsilon} \sup_{x\in N, y\in M} \langle \mathbf{Ax}, \mathbf{y}\rangle$$

Moreover, if $\mathbf{A}$ is symmetric, then we can take $\mathbf{x} = \mathbf{y}$ in above formula.

Proof. Pick $\mathbf{x}$ and $\mathbf{y}$ *s.t.* $\|\mathbf{A}\| = \langle \mathbf{Ax}, \mathbf{y}\rangle$, then select $\mathbf{x_0}$ and $\mathbf{y_0}$ near them respectively in N and M. Then

$$\begin{aligned}\langle \mathbf{Ax}, \mathbf{y}\rangle - \langle \mathbf{Ax_0}, \mathbf{y_0}\rangle &= \langle \mathbf{A}(\mathbf{x}-\mathbf{x_0}), \mathbf{y}\rangle + \langle \mathbf{Ax_0}, \mathbf{y}-\mathbf{y_0}\rangle \\ &\le \|\mathbf{A}\| \left(\|\mathbf{x}-\mathbf{x_0}\| \|\mathbf{y}\| + \|\mathbf{x_0}\| \|\mathbf{y}-\mathbf{y_0}\|\right) \\ &\le 2\varepsilon \|\mathbf{A}\|\end{aligned}$$

when $\mathbf{A}$ is symmetric, x_0 also approximate $\mathbf{y}$ and then claim follows.

□

Proposition 4.14. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\varepsilon \in [0, \frac{1}{2})$, N a ε net of S^{n-1}, we have*

$$\sup_{\mathbf{x}\in S^{n-1}} |\|\mathbf{Ax}\| - \mu| \ll \frac{1}{1-2\varepsilon} \sup_{\mathbf{x}\in N} |\|\mathbf{Ax}\| - \mu|$$

4.4.1 The norms of sub-gaussian random matrices

The following states that $\mathbf{A}$ with independent sub-gaussian entries satisfies

$$\|\mathbf{A}\| \ll \sqrt{m} + \sqrt{n}$$

with high probability.

Theorem 4.4. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ with independent, mean zero, sub-gaussian entries, then for any $t > 0$ we have*

$$\mathbb{P}\left\{\|\mathbf{A}\| \le CK(\sqrt{m} + \sqrt{n} + t)\right\} \ge 1 - 2e^{-t^2}$$

where $K = \max_{ij} \|A_{ij}\|_{\psi_2}$

Proof. **Step 1**. Take $\varepsilon = \frac{1}{4}$, then corollary 4.1 yields we can find an N for S^{n-1} and M for S^{m-1} with cardinalities less than 9^n and 9^m. By proposition 4.13, we have

$$\|\mathbf{A}\| \le 2 \max_{\mathbf{x}\in N, \mathbf{y}\in M} \langle \mathbf{Ax}, \mathbf{y}\rangle$$

Step 2. By proposition proposition 2.9

□