Notes of high-dimensional probability

Ruihan Liu

Contents

1 Concentration Inequalities

One core task in probability is to quantify how a random variable X deviates around its mean $\mathbb{E}[X]$, and concentration inequalities are the optimal tool to deal with this task. In the beginning, we consider a simple model and its relative concentration inequality.

1.1 Hoeffding's Inequality

Lemma 1.1.1 (Hoeffding lemma). *Assume a bounded random variable X taking value in $[m, M]$ and $\mathbb{E}[X] = 0$, then for $t > 0$*

$$\mathbb{E}[e^{tX}] \le \exp\left(\frac{t^2(M-m)^2}{8}\right) \tag{1}$$

Proof. By Jensen's inequality, since e^{tx} is convex, it concludes that

$$e^{tx} \le \frac{M-x}{M-m}e^{tm} + \frac{x-m}{M-m}e^{tM}$$

Hence

$$\mathbb{E}[e^{tX}] \le \frac{M}{M-m}e^{tm} - \frac{m}{M-m}e^{tM}$$

Let $h = t(M-m)$ and $p = -m/(M-m)$, denote $L(h) = -hp + \log(1-p+pe^h)$ and then

$$\frac{M}{M-m}e^{tm} - \frac{m}{M-m}e^{tM} = e^{-hp}\left(1-p+pe^h\right) = e^{L(h)}$$

In addition, $L(0) = L'(0) = 0$ and $L''(h) \le 1/4$, it concludes by Taylor expansion

$$L(h) \le \frac{1}{8}h^2$$

which is our conclusion. □

Theorem 1.1.1 (Hoeffding inequality). *Let X_i be independent bounded random variable such that $X_i \in [m_i, M_i]$ for $i = 1, \cdots, n$, then for $t > 0$*

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}[X_i] \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (M_i - m_i)^2}\right) \tag{2}$$

Proof. Let $Y_i := X_i - \mathbb{E}[X_i]$, which is bounded and taking value in $[m_i - \mathbb{E}[X_i], M_i - \mathbb{E}[X_i]]$, then for $\lambda > 0$ it concludes that

$$\mathbb{P}\left(\sum_{i=1}^n Y_i \ge t\right) = \mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^n Y_i\right) \ge e^{\lambda t}\right) \le e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n Y_i\right)\right]$$

Since X_i are independent with each other, so as Y_i. Combine with Lemma 1.1.1, we have

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}Y_i\right)\right]=\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda Y_i}\right]\leq\exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(M_i-m_i)^2\right)$$

As a result, we obtain

$$\mathbb{P}\left(\sum_{i=1}^{n}Y_i\geq t\right)\leq\exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(M_i-m_i)^2-\lambda t\right)$$

Since the above inequality is valid for all $\lambda>0$, then as $\lambda=4t/\sum_{i=1}^{n}(M_i-m_i)^2$, it attends its minimum

$$\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(M_i-m_i)^2}\right)$$

then we finish our proof. □

1.2 Chernoff Inequality

The general form of Hoeffding inequality is sometimes too conservative and does not give sharp results. This happens, for example, when the X_i are Bernoulli random variables with parameters p_i so small that we expect S_N to have an approximately Poisson distribution. However, Hoeffding inequality is not sensitive to the magnitudes of p_i , and the Gaussian tail bound that it gives is very far from the true, Poisson, tail. In this section we study Chernoff inequality, which is sensitive to the magnitudes of p_i.

Theorem 1.2.1 (Chernoff inequality)**.** *Let X_i be independent Bernoulli random variables with parameters p_i for $i=1,\cdots,N$, denote $S_N=\sum_{i=1}^{N}X_i$ and $\mathbb{E}[S_N]=\mu$, then for $t\neq\mu$, it concludes that*

$$\mathbb{P}\left(S_N\geq t\right)\leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t\quad\text{for } t>\mu \tag{3}$$

$$\mathbb{P}\left(S_N\leq t\right)\leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t\quad\text{for } t<\mu \tag{4}$$

Proof. First, consider $t>\mu$ and $\lambda>0$, by Chebyshev inequality, we obtain

$$\mathbb{P}\left(S_N\geq t\right)\leq e^{-\lambda}\prod_{i=1}^{N}\mathbb{E}\left[e^{\lambda X_i}\right]$$

Moreover

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

Hence

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

Since above inequality is valid for any $\lambda>0$ and the right side attains its minimum as $\lambda=\log(t/\mu)$, and we conclude (3). For $t<\mu$, we have

$$\mathbb{P}\left(-S_N\geq -t\right)$$

where $-S_N$ has the mean of $-\mu$, hence it concludes (4) by the same method as above. □

In fact, we have a better result in the next theorem.

Theorem 1.2.2 (Chernoff inequality: small deviations)**.** *With the same conditions in above theorem, let $\delta \in (0,1]$ and then*

$$\mathbb{P}\left(|S_N - \mu| \geq \delta\mu\right) \leq 2e^{-c\mu\delta^2} \tag{5}$$

where $c > 0$ is an absolute constant.

Proof. First, we have

$$\begin{aligned}\mathbb{P}\left(|S_N - \mu| \geq \delta\mu\right) &= \mathbb{P}\left(S_N \geq (1+\delta)\mu\right) + \mathbb{P}\left(S_N \leq (1-\delta)\mu\right) \\ &\leq e^{\delta\mu}\left(\frac{1}{1+\delta}\right)^{(1+\delta)\mu} + e^{-\delta\mu}\left(\frac{1}{1-\delta}\right)^{(1-\delta)\mu}\end{aligned}$$

Next, we want to show there exists a constant $c_1 > 0$ such that

$$f(\delta) := (1+\delta)\exp\left(-\frac{c_1\delta^2 + \delta}{1+\delta}\right) \geq 1$$

Taking derivative and obtain

$$f'(\delta) = \exp\left(-\frac{c_1\delta^2 + \delta}{1+\delta}\right)\left(-\frac{(c_1\delta + 2c_1 - 1)\delta}{1+\delta}\right)$$

If $c_1 \geq 1$, it concludes that $f'(\delta) > 0$ for $\delta \in (0,1]$, hence $f(\delta) > f(0) = 1$. On the other hand, we also need to show that there exists $c_2 > 0$ such that

$$f(-\delta) \geq 1$$

If $0 < c_2 < 0.5$, $f'(-\delta) > 0$ for $\delta \in (0,1]$, hence $f(-\delta) > f(0) = 1$. As a result, we conclude that

$$e^{\delta\mu}\left(\frac{1}{1+\delta}\right)^{(1+\delta)\mu} + e^{-\delta\mu}\left(\frac{1}{1-\delta}\right)^{(1-\delta)\mu} \leq e^{-c_1\mu\delta^2} + e^{-c_2\mu\delta^2} \leq 2e^{-\min\{c_1,c_2\}\mu\delta^2}$$

Then finish this proof. □

Remark 1.2.1. *Since $S_N \to \mathrm{Pois}(\mu) \sim X$ in distribution as $N \to \infty$ if $\mu := \sum_{i=1}^{\infty} p_i < \infty$, we can conclude the following results with the help of above theorems for $t > \mu$ and $s \in (0,\mu]$*

$$\mathbb{P}\left(X \geq t\right) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t \tag{6}$$

$$\mathbb{P}\left(|X - \mu| \geq s\right) \leq 2\exp\left(-\frac{cs^2}{\mu}\right) \tag{7}$$

1.3 Random graph

The study of random graph started by *Erdös − Rényi* model $G(n, p)$, that is, a graph with n vertexes and any of two is connected with probability p. Hence, the expectation of edges is $\binom{n}{2}p$ and the degree of vertexes $d := 2\binom{n}{2}p/n = (n-1)p$. And the following theorem shows that the degrees of all vertexes are close to d.

Theorem 1.3.1 (Dense graphs are almost regular)**.** *There is an absolute constant C such that the following holds. Consider a random graph $G \sim G(n, p)$ with expected degree satisfying $d \geq C \log n$. Then, with high probability the following occurs: all vertexes of G have degrees between $0.9d$ and $1.1d$.*

Proof. Let d_i be the degree of i-th vertex, by Theorem 1.2.2, it concludes that

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

where $\delta > 0$ is a small constant. Then we have

$$\mathbb{P}\left(\exists\ i \leq n : |d_i - d| \geq \delta d\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(|d_i - d| \geq \delta d\right) \leq 2ne^{-cd}$$

Since $d \geq C \log n$ for sufficient large constant C, we obtain

$$\mathbb{P}\left(\forall\ i \leq n : |d_i - d| < \delta d\right) \geq 1 - \delta$$

This completes the proof. □

1.4 Sub-Gaussian distributions

So far, we mainly focus on Bernoulli random variable and its relative concentration inequalities, but we expect some other distributions also have the same result, at least for Gaussian distribution, due to the central limit theorem. Remind the Hoeffding inequality, we have

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

Let $a_1 = 1$ and other $a_i = 0$, we obtain

$$\mathbb{P}\left(|X_1| \geq t\right) \leq 2\exp(-ct^2)$$

According to above result, we give the following definition.

Definition 1.4.1 (Sub-Gaussian distributions)**.** *A random variable X is called sub-Gaussian if it has the tail of*

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

where $c > 0$ is a constant.

This class of distributions deserves special attention. It is sufficiently wide as it contains Gaussian, Bernoulli, and all bounded distributions.

Proposition 1.4.1 (Sub-Gaussian property)**.** *Let X be a random variable. Then the following properties are equivalent, where the parameters $K_i > 0$.*

(i) The tails of X satisfy

$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp\left(-t^2/K_1^2\right) \text{ for } t \geq 0 \tag{8}$$

(ii) The moments of X satisfy

$$||X||_{L^p} \leq K_2\sqrt{p} \text{ for } p \geq 1 \tag{9}$$

(iii) The MGF of X^2 satisfies

$$\mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp\left(K_3^2\lambda^2\right) \text{ for } |\lambda| \leq \frac{1}{K_3} \tag{10}$$

(iv) The MGF of X^2 is bounded at some point, namely

$$\mathbb{E}\left[\exp\left(X^2/K_4^2\right)\right] \leq 2 \tag{11}$$

Moreover, if $\mathbb{E}[X] = 0$ then properties (i)–(iv) are also equivalent to the following property.

(v) The MGF of X satisfies

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(K_5^2\lambda^2\right) \text{ for } \lambda \in \mathbb{R} \tag{12}$$

Proof. (i)⇒(ii): Without losing generality, assume $K_1 = 1$ or consider X/K_1, we have

$$||X||_{L^p}^p = p\int_0^\infty t^{p-1}\mathbb{P}\left(|X| \geq t\right) dt \leq 2p\int_0^\infty t^{p-1}\exp\left(-t^2\right) dt = \int_0^\infty t^{p-1}e^{-t^2} dt = p\Gamma\left(\frac{p}{2}\right)$$

By Stirling approximation, we have $\Gamma(x) \leq x^x$ and finally obtain

$$||X||_{L^p} \leq \sqrt[p]{p}\sqrt{\frac{p}{2}} \leq K_2\sqrt{p}$$

where $K_2 \geq e^{1/e}\sqrt{2}$ is a constant.

(ii)⇒(iii): Consider Taylor expansion of $\exp\left(\lambda^2 X^2\right)$, it concludes that

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

By property (ii), we have $\mathbb{E}\left[X^{2p}\right] \leq (2p)^p$ by letting $K_2 = \sqrt{2}$, while Stirling approximation yields that $p! \geq (p/e)^p$, then for $2\lambda^2 e < 1$ it concludes that

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

Next, let $K_3 = \sqrt{1/(4e)}$, we can conclude that

$$\frac{1}{1-2\lambda^2 e} \leq \exp(K_3^2\lambda^2) \text{ for } |\lambda| \leq \frac{1}{K_3}$$

(iii)⇒(iv): trivial.

(iv)⇒(i): Without losing generality, assume $K_4 = 1$ and by Chebyshev inequality, we have

$$\mathbb{P}(|X| \geq t) = \mathbb{P}\left(X^2 \geq t^2\right) \leq e^{-t^2}\mathbb{E}[\exp(X^2)] \leq 2e^{-t^2}$$

Finally, suppose $\mathbb{E}[X] = 0$ and we show that (iii)⇒(v) and (v)⇒(i).

(iii)⇒(v): Assume $K_3 = 1$ and consider the inequality $e^x \leq x + e^{x^2}$ for $x \in \mathbb{R}$, we conclude that

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \mathbb{E}[\exp\left(\lambda^2 X^2\right)] \leq \exp(\lambda^2) \text{ for } |\lambda| \leq 1$$

For $|\lambda| > 1$, we have

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp(\lambda^2/2)\mathbb{E}\left[\exp(X^2/2)\right] \leq \exp\left((\lambda^2+1)/2\right) \leq \exp(\lambda^2)$$

(v)⇒(i): Assume $K_5 = 1$ and by Chebyshev inequality, we have

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\lambda^2 - \lambda t\right)$$

where $\lambda > 0$. Hence $\mathbb{P}(X \geq t) \leq e^{-t^2/4}$. On the other hand, consider $-X$ and still have $\mathbb{P}(X \leq -t) \leq e^{-t^2/4}$. Finally, we obtain $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/4}$. □

Remark 1.4.1. *In fact,* $\mathbb{E}[X] = 0$ *is necessary for condition (v). Suppose* $\mathbb{E}[X] = \delta > 0$ *and we have*

$$\mathbb{E}[\exp(\lambda X)] \geq 1 + \lambda\mathbb{E}[X] = 1 + \delta\lambda$$

Next, for any $M > 0$*, denote* $f(\lambda) = 1 + \delta\lambda - e^{M\lambda^2}$*, we have*

$$f'(\lambda) = \delta - 2M\lambda e^{M\lambda^2}$$

And there exits $\lambda_0 > 0$ *such that* $f(\lambda_0) = 0$ *and* $f(\lambda) > 0$ *if* $\lambda < \lambda_0$*. In addition,* $f(0) = 0$ *hence* $f(\lambda) > 0$ *for* $\lambda \in (0, \lambda_0)$*. Therefore, there exists* $\lambda > 0$ *such that for any* $M > 0$*,* $\mathbb{E}[\exp(\lambda X)] > \exp(M\lambda^2)$*, and condition (v) is not valid.*

Next, let's see the definition of sub-Gaussian random variable.

Definition 1.4.2. *A random variable* X *that satisfies one of the equivalent properties (i)–(iv) in last proposition is called a sub-Gaussian random variable. The sub-gaussian norm of* X*, denoted* $||X||_{\psi_2}$*, is defined to be the smallest* K_4 *in property (iv). In other words, we define*

$$||X||_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(X^2/t^2\right)\right] \leq 2\right\} \tag{13}$$

Here, we only check that $||X+Y||_{\psi_2} \leq ||X||_{\psi_2} + ||Y||_{\psi_2}$. According to above definition, for any $\delta > 0$, it has

$$2 \geq \sqrt{\mathbb{E}\left[\exp\left(\frac{X^2}{||X||_{\psi_2}+\delta}\right)\right]\mathbb{E}\left[\exp\left(\frac{Y^2}{||Y||_{\psi_2}+\delta}\right)\right]} \geq \mathbb{E}\left[\exp\left(\frac{1}{2}\left(\frac{X^2}{||X||_{\psi_2}+\delta}+\frac{Y^2}{||Y||_{\psi_2}+\delta}\right)\right)\right]$$
$$\geq \mathbb{E}\left[\exp\left(\frac{1}{2}\frac{X^2+Y^2}{||X||_{\psi_2}+||Y||_{\psi_2}+2\delta}\right)\right] \geq \mathbb{E}\left[\exp\left(\frac{(X+Y)^2}{||X||_{\psi_2}+||Y||_{\psi_2}+2\delta}\right)\right]$$

Hence we obtain $||X||_{\psi_2} + ||Y||_{\psi_2} + 2\delta \geq ||X+Y||_{\psi_2}$ for any $\delta > 0$, i.e. $||X+Y||_{\psi_2} \leq ||X||_{\psi_2} + ||Y||_{\psi_2}$.

Remark 1.4.2. *With the help of sub-Gaussian norm, we can restate conditions in Proposition 1.4.1 by*

(i) $\mathbb{P}(|X| \geq t) \leq 2\exp\left(-ct^2/||X||_{\psi_2}^2\right)$ *for all* $t \geq 0$.

(ii) $||X||_{L^p} \leq C||X||_{\psi_2}\sqrt{p}$ *for all* $p \geq 1$.

(iv) $\mathbb{E}\left[\exp\left(X^2/||X||_{\psi_2}^2\right)\right] \leq 2$.

(v) $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(C\lambda^2||X||_{\psi_2}^2\right)$ *if* $\mathbb{E}[X]=0$ *and* $\lambda \in \mathbb{R}$.

where $C, c > 0$ *are constant.*

Next, let's compute the sub-Gaussian norm of some special distributions.

- If $X \sim \mathcal{N}(0,\sigma)$, it has $||X||_{\psi_2} \leq C\sigma$, where $C > \sqrt{2}$ is a constant.
- If X is bounded, then $||X||_{\psi_2} \leq C||X||_\infty$, where $C > 1/\sqrt{\log 2}$ is a constant.

However, the Poisson, exponential, Pareto, and Cauchy distributions are not sub-Gaussian. In the end, let's see the following estimation about the sequence of sub-Gaussian random variables.

Proposition 1.4.2 (Maximum of sub-Gaussian)**.** *Let* $X_1, X_2, \cdots$ *be an infinite sequence of sub-Gaussian random variables which are not necessarily independent. Show that*

$$\mathbb{E}\left[\max_{i\in\mathbb{N}^+}\frac{|X_i|}{\sqrt{1+\log i}}\right] \leq CK \tag{14}$$

where $K := \max_{i\in\mathbb{N}^+}||X_i||_{\psi_2}$.

Proof. Consider

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in\mathbb{N}^+}\frac{|X_i|}{\sqrt{1+\log i}}\right] &= \int_0^\infty \mathbb{P}\left(\max_{i\in\mathbb{N}^+}\frac{|X_i|}{\sqrt{1+\log i}}\geq t\right)dt \\
&= \sqrt{2}K + \int_{\sqrt{2}K}^\infty \sum_{i=1}^\infty \mathbb{P}\left(\frac{|X_i|}{\sqrt{1+\log i}}\geq t\right)dt \\
&\leq \sqrt{2}K + 2\int_{\sqrt{2}K}^\infty \sum_{i=1}^\infty \exp\left(-\frac{t^2(\log i+1)}{||X||_{\psi_2}^2}\right)dt \\
&\leq \sqrt{2}K + 2\int_{\sqrt{2}K}^\infty \sum_{i=1}^\infty \left(\frac{1}{i}\right)^{\frac{t^2}{K^2}} \exp\left(-\frac{t^2}{K^2}\right)dt \\
&\leq \sqrt{2}K + \frac{\pi}{3}\int_{\sqrt{2}K}^\infty \exp\left(-\frac{t^2}{K^2}\right)dt \leq \left(\sqrt{2}+\frac{\pi}{2}\sqrt{\frac{\pi}{2}}\right)K
\end{aligned}$$

where the first $\leq$ is concluded from (i) in Proposition 1.4.1. □

Remark 1.4.3. *As a consequence of above Proposition, we deduce for $N \geq 2$, it has*

$$\mathbb{E}\left[\max_{1\leq i\leq N}|X_i|\right] \leq CK\sqrt{\log N} \tag{15}$$

Actually, we can show the bound is sharp. Let $X_1,\cdots,X_N$ be independent $\mathcal{N}(0,1)$, we can show that

$$\mathbb{E}\left[\max_{1\leq i\leq N}|X_i|\right] \geq c\sqrt{\log N} \tag{16}$$

where $c > 0$ is a constant.

1.5 General Hoeffding and Khintchine Inequalities

Until now, we only discuss Hoeffding inequality related to Bernoulli distribution and bounded random variables, in this subsection, we will extend Hoeffding inequality to general sub-Gaussian distributions. The following proposition is the preliminary for our goal.

Proposition 1.5.1 (Sums of independent sub-Gaussian)**.** *Let $X_1,\cdots,X_N$ be independent mean-zero sub-Gaussian random variables, then $\sum_{i=1}^N X_i$ is also sub-Gaussian and satisfies*

$$\left|\left|\sum_{i=1}^N X_i\right|\right|_{\psi_2}^2 \leq C\sum_{i=1}^N ||X_i||_{\psi_2}^2 \tag{17}$$

where $C > 0$ is a constant.

Proof. For $\lambda \in \mathbb{R}$, by condition (v) in Proposition 1.4.1, it has

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^N X_i\right)\right] = \prod_{i=1}^N \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \leq \exp\left(C\lambda^2\sum_{i=1}^N ||X_i||_{\psi_2}^2\right)$$

Hence $\sum_{i=1}^N X_i$ is sub-Gaussian, in addition, we have

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^N X_i\right)\right] \leq \exp\left(C\lambda^2\Big|\Big|\sum_{i=1}^N X_i\Big|\Big|_{\psi_2}^2\right)$$

where $\big|\big|\sum_{i=1}^N X_i\big|\big|_{\psi_2}^2$ is the smallest number such that above inequality is valid, so we finish our proof. □

As a consequence of above proposition, we get general form of Hoeffding inequality.

Theorem 1.5.1 (General Hoeffding inequality)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero sub-Gaussian random variables, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^2||a||_2^2}\right) \tag{18}$$

where $K = \max_{1\leq i\leq N}||X_i||_{\psi_2}$.

As a conclusion of general Hoeffding inequality, we can quickly derive the classical Khintchine inequality for the L^p norms.

Theorem 1.5.2 (Khintchine inequality)**.** *Let $X_1, \cdots, X_N$ be independent sub-Gaussian random variables with zero means and unit variances, and let $a = (a_1, \cdots, a_N) \in R^N$. Then for every $p \in [2, \infty)$ we have*

$$\left(\sum_{i=1}^N a_i^2\right)^{1/2} \leq \left|\left|\sum_{i=1}^N a_i X_i\right|\right|_{L^p} \leq CK\sqrt{p}\left(\sum_{i=1}^N a_i^2\right)^{1/2} \tag{19}$$

where $K = \max_{1\leq i\leq N}||X_i||_{\psi_2}$ and $C > 0$ is a constant.

Proof. First,

$$\begin{aligned}
\mathbb{E}\left[\Big|\sum_{i=1}^N a_i X_i\Big|^p\right] &= p\int_0^\infty t^{p-1}\mathbb{P}\left(\Big|\sum_{i=1}^N a_i X_i\Big| \geq t\right)dt \\
&\leq 2p\int_0^\infty t^{p-1}\exp\left(-\frac{ct^2}{K^2||a||_2^2}\right)dt \\
&= 2p\left(\frac{K||a||_2}{\sqrt{c}}\right)^p\int_0^\infty x^{p-1}\exp(-x^2)dx \\
&= p\left(\frac{K||a||_2}{\sqrt{c}}\right)^p\Gamma\left(\frac{p}{2}\right) \leq p\left(\frac{K||a||_2}{\sqrt{c}}\right)^p\left(\frac{p}{2}\right)^{p/2}
\end{aligned}$$

Hence, we obtain $\big|\big|\sum_{i=1}^N a_i X_i\big|\big|_{L^p} \leq CK\sqrt{p}||a||_2$. For the other side, suppose $p > 2$ and by Hölder inequality, we have

$$\mathbb{E}\left[\Big|\sum_{i=1}^N a_i X_i\Big|^p\right]^{\frac{2}{p}} \geq \mathbb{E}\left[\Big|\sum_{i=1}^N a_i X_i\Big|^2\right]$$

Since $\text{var}(X_i) = 1$ and all X_i are independent to each other and mean-zero, it concludes that

$$\mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|^2\right] = ||a||_2^2$$

Hence we conclude that

$$\left|\left|\sum_{i=1}^{N} a_i X_i\right|\right|_{L^p} \geq ||a||_2$$

As for $p = 2$, we have shown it before. □

Corollary 1.5.1 (Khintchine inequality for p = 1)**.** *With the same conditions as above theorem, let $p = 1$, then we have*

$$c(K)\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \leq \left|\left|\sum_{i=1}^{N} a_i X_i\right|\right|_{L^1} \leq \left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \tag{20}$$

where $K = \max_{1\leq i\leq N} ||X_i||_{\psi_2}$ and $c(K) > 0$ is a constant only depended on K.

Proof. First,

$$\mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|\right] \leq \mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|^2\right]^{\frac{1}{2}} = ||a||_2$$

On the other hand, by Hölder inequality, we have

$$\mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|^3\right]^{\frac{3}{4}} \mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|\right]^{\frac{1}{4}} \geq \mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|^2\right] = ||a||_2^2$$

Moreover,

$$\mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|^3\right] \leq CK\sqrt{3}||a||_2$$

Hence,

$$\mathbb{E}\left[\Big|\sum_{i=1}^{N} a_i X_i\Big|\right] \geq \left(\frac{1}{CK\sqrt{3}}\right)^3 ||a||_2$$

Then finish our proof. □

Remark 1.5.1. *In fact, by similar method, we can show for $p \in (0, 2)$, there still has*

$$c(K,p)\left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \leq \left|\left|\sum_{i=1}^{N} a_i X_i\right|\right|_{L^p} \leq \left(\sum_{i=1}^{N} a_i^2\right)^{1/2} \tag{21}$$

where $K = \max_{1\leq i\leq N} ||X_i||_{\psi_2}$ and $c(K,p) > 0$ is a constant only depended on K, p.

Until now, we only focus on mean-zero sub-Gaussian random variables, for more general cases, we have the following result for preliminary.

Lemma 1.5.1 (Centering)**.** *If X is a sub-Gaussian random variable then $X\mathbb{E}[X]$ is sub-Gaussian too and*

$$||X - \mathbb{E}[X]||_{\psi_2} \leq C||X||_{\psi_2} \tag{22}$$

where $C > 0$ is a constant.

Proof. First,

$$||X - \mathbb{E}[X]||_{\psi_2} \leq ||X||_{\psi_2} + ||\mathbb{E}[X]||_{\psi_2} = ||X||_{\psi_2} + \frac{|\mathbb{E}[X]|}{\sqrt{\log 2}}$$

In addition, $|\mathbb{E}[X]| \leq ||X||_{L^1} \leq C||X||_{\psi_2}$ by (ii) in Proposition 1.4.1. □

1.6 Sub-Exponential distributions

Although the class of sub-Gaussian distributions is quite large, it still leaves some useful and meaningful distributions. For instance, consider X_i are independent $\mathcal{N}(0,1)$ for $i = 1, \cdots, N$ and denote

$$S_N = \left(\sum_{i=1}^{N} X_i^2\right)^{1/2}$$

On the one hand, S_N is a sum of independent random variables X_i^2, so we should expect some concentration to hold. On the other hand, although the X_i are sub-Gaussian random variables, the X_i^2 are not. In fact, we have

$$\mathbb{P}\left(X_i^2 \geq t\right) = \mathbb{P}\left(|X_i| \geq \sqrt{t}\right) \leq \exp\left(-t/2\right)$$

The tails of X_i^2 are like those for the exponential distribution and are strictly heavier than sub- Gaussian. This prevents us from using Hoeffding inequality if we want to study the concentration of S_N. Hence, it is necessary to introduce sub-exponential distributions., but we first focus on the following properties

Proposition 1.6.1 (Sub-exponential property)**.** *Let X be a random variable. Then the following properties are equivalent, where the parameters $K_i > 0$.*

(i) The tails of X satisfy

$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp\left(-t/K_1\right) \text{ for } t \geq 0 \tag{23}$$

(ii) The moments of X satisfy

$$||X||_{L^p} \leq K_2 p \text{ for } p \geq 1 \tag{24}$$

(iii) The MGF of $|X|$ satisfies

$$\mathbb{E}\left[\exp\left(\lambda|X|\right)\right] \leq \exp\left(K_3\lambda\right) \text{ for } 0 \leq \lambda \leq \frac{1}{K_3} \tag{25}$$

(iv) The MGF of $|X|$ is bounded at some point, namely

$$\mathbb{E}\left[\exp\left(|X|/K_4\right)\right] \leq 2 \tag{26}$$

Moreover, if $\mathbb{E}[X] = 0$ then properties (i)–(iv) are also equivalent to the following property.

(v) The MGF of X satisfies

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(K_5^2\lambda^2\right) \text{ for } |\lambda| \leq \frac{1}{K_5} \tag{27}$$

Proof. (i)⇒(ii): assume $K_1 = 1$ and we have

$$\mathbb{E}\left[|X|^p\right] = p\int_0^\infty t^{p-1}\mathbb{P}\left(|X| \geq t\right)dt \leq 2p\int_0^\infty t^{p-1}e^{-t}dt = 2p\Gamma(p) \leq 2p^{p+1}$$

(ii)⇒(iii): let $K_2\lambda e < 1$, it concludes that

$$\mathbb{E}\left[\exp\left(\lambda|X|\right)\right] = 1 + \sum_{k=1}^\infty \frac{\lambda^k ||X||_{L^k}^k}{k!} \leq 1 + \sum_{k=1}^\infty \frac{(\lambda K_2)^k k^k}{(k/e)^k} = \frac{1}{1 - K_2\lambda e}$$

If $K_3 \geq 2eK_2$, we have $(1 - eK_2\lambda)\exp(K_3\lambda) \geq 1$ for $\lambda \in [0, 1/K_3]$.

(iii)⇒(iv): trivial.

(iv)⇒(i): by Chebyshev inequality.

If $\mathbb{E}[X] = 0$ and assume $K_2 = 1$, we have

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] = 1 + \sum_{k=2}^\infty \frac{\lambda^k ||X||_{L^k}^k}{k!} \leq 1 + \sum_{k=2}^\infty \frac{(\lambda)^k k^k}{(k/e)^k} = 1 + \frac{(\lambda e)^2}{1 - \lambda e}$$

If $K_5 \leq \left(1 + \sqrt{1 + 16e}\right)/4e$, we have

$$1 + \frac{(\lambda e)^2}{1 - \lambda e} \leq \exp\left(K_5^2\lambda^2\right)$$

That is (ii)⇒(v). As for the inverse, assume that $K_5 = 1$ and consider the numerical inequality $|x|^p \leq p^p(e^x + e^{-x})$ for $p > 0$ and $x \in \mathbb{R}$. We have

$$\mathbb{E}\left[|X|^p\right] \leq p^p\left(\mathbb{E}\left[\exp(X)\right] + \mathbb{E}\left[\exp(-X)\right]\right) \leq 2p^p e$$

That is (v)⇒(ii). □

Remark 1.6.1. *For the class of distributions whose tail decay is of the type* $\exp(-ct^\alpha)$ *or faster, where $\alpha > 0$. We still have similar results as $\alpha = 1, 2$.*

(i) The tails of X satisfy

$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp\left(-t^\alpha/K_1^\alpha\right) \text{ for } t \geq 0 \tag{28}$$

(ii) The moments of X satisfy

$$||X||_{L^p} \leq K_2 \sqrt[\alpha]{p} \text{ for } p \geq 1 \tag{29}$$

(iii) The MGF of $|X|^\alpha$ satisfies

$$\mathbb{E}\left[\exp\left(\lambda^\alpha |X|^\alpha\right)\right] \leq \exp\left(K_3^\alpha \lambda^\alpha\right) \text{ for } 0 \leq \lambda \leq \frac{1}{K_3} \tag{30}$$

(iv) The MGF of $|X|^\alpha$ is bounded at some point, namely

$$\mathbb{E}\left[\exp\left(|X|^\alpha/K_4^\alpha\right)\right] \leq 2 \tag{31}$$

Moreover, if $\mathbb{E}[X] = 0$ then properties (i)–(iv) are also equivalent to the following property.

(v) The MGF of X satisfies

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(K_5^2 \lambda^2\right) \text{ for } |\lambda| \leq \frac{1}{K_5} \tag{32}$$

Remark 1.6.2. *Moreover, $0 \leq \lambda \leq 1/K_3$ in condition (iii) of Proposition 1.6.1 can not be changed by $|\lambda| \leq 1/K_3$. In fact, suppose $\mathbb{E}[|X|] = \delta > 0$ and $\delta < K_3$, we have*

$$\mathbb{E}\left[\lambda|X|\right] \geq 1 + \lambda\delta > \exp\left(K_3\lambda\right)$$

For $\lambda \in (\log(\delta/K_3)/K_3, 0)$.

Definition 1.6.1 (Sub-exponential distribution)**.** *A random variable X that satisfies one of the equivalent properties (i)–(iv) in Proposition 1.6.1 is called a sub-exponential random variable. The sub-exponential norm of X, denoted $||X||_{\psi_1}$, is defined to be the smallest K_3 in property (iii). In other words,*

$$||X||_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(|X|/t\right)\right] \leq 2\right\} \tag{33}$$

From the definition of sub-exponential distribution, we know a sub-Gaussian random variable is also sub-exponential. Besides, we have the following lemmas.

Lemma 1.6.1. *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$||X^2||_{\psi_1} = ||X||_{\psi_2}^2 \tag{34}$$

Proof. Since

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

it concludes $||X^2||_{\psi_1} \leq ||X||_{\psi_2}^2$. On the other hand,

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

Hence $||X^2||_{\psi_1} \geq ||X||_{\psi_2}^2$. □

Lemma 1.6.2 (The product of sub-Gaussian is sub-exponential)**.** *Let X and Y be sub-Gaussian random variables. Then XY is sub-exponential. Moreover,*

$$||XY||_{\psi_1} \leq ||X||_{\psi_2}||Y||_{\psi_2} \tag{35}$$

Proof. Without loss of generality we may assume that $||X||_{\psi_2} = ||Y||_{\psi_2} = 1$, or consider $X/||X||_{\psi_2}$. As a result, we have

$$||XY||_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(|XY|/t\right)\right] \leq 2\right\}$$

In fact, for $\delta > 0$ we have

$$\mathbb{E}\left[\exp\left(\frac{|XY|}{1+\delta}\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{X^2+Y^2}{2(1+\delta)}\right)\right] \leq \sqrt{\mathbb{E}\left[\exp\left(\frac{X^2}{1+\delta}\right)\right]\mathbb{E}\left[\exp\left(\frac{Y^2}{1+\delta}\right)\right]} \leq 2$$

Hence $|XY||_{\psi_1} \leq 1 + \delta$ for any $\delta > 0$. □

Remark 1.6.3. *Notice that all condition (v) in Proposition 1.4.1 and 1.6.1 have the same bound for MGF near the origin. It is easy to see, consider the Taylor expansion of*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] = \mathbb{E}\left[1 + \lambda X + \frac{(\lambda X)^2}{2} + o((\lambda X)^2)\right] \approx 1 + \frac{\lambda^2}{2}\text{var}(X) \approx \exp\left(\frac{\lambda^2}{2}\text{var}(X)\right)$$

as $\lambda \to 0$

1.7 Orlicz Spaces

In this subsection, we will extend the concept of sub-Gaussian distributions to more general frameworks, which is called Orlicz Spaces. First, a function $\psi : [0, \infty) \to [0, \infty)$, where ψ is convex, increasing and satisfies

$$\psi(0) = 0,\ \ \psi(x) \to \infty \text{ as } x \to \infty \tag{36}$$

then $\psi(x)$ is called Orlicz function. Next, for a given ψ, the Orlicz norm of a random variable X is defined by

$$||X||_{\psi} := \inf\left\{t > 0 : \mathbb{E}\left[\psi\left(X/t\right)\right] \leq 1\right\} \tag{37}$$

Then the Orlicz space of ψ contains all random variables with finite Orlicz norm, i.e.

$$L_{\psi} : \{X : ||X||_{\psi} < \infty\} \tag{38}$$

In fact, we can show that L_{ψ} is a Banach space. Let $\{X_i\}_{i=1}^{\infty}$ be a Cauchy sequence in L_{ψ} such that

$$||X_{i+1} - X_i||_{\psi} \leq \frac{1}{2^i}$$

1.8 Bernstein Inequality

Theorem 1.8.1 (Bernstein Inequality)**.** *Let $X_1, \cdots, X_N$ be N independent mean-zero sub-exponential random variables, then for $t \geq 0$, it concludes that*

$$\mathbb{P}\left(\Big|\sum_{i=1}^{N} X_i\Big| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^{N}||X_i||_{\psi_1}^2}, \frac{t}{\max_{1\leq i\leq N}||X_i||_{\psi_1}}\right\}\right) \tag{39}$$

where $c > 0$ is a constant.

Proof. By Chebyshev inequality, it concludes that

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

where $\lambda > 0$ is a parameter. Then by condition (v) in Proposition 1.6.1, it concludes that

$$\mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \leq \exp\left(C||X_i||_{\psi_1}^2\lambda^2\right) \text{ for } i \in \{1, \cdots, N\}$$

where $|\lambda| \leq c/\max_{1\leq i\leq N}||X_i||_{\psi_1}$. Hence it concludes that

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i \geq t\right) \leq \exp\left(C\lambda^2\sum_{i=1}^{N}||X_i||_{\psi_1}^2 - \lambda t\right)$$

And the optimal choice of λ within domain $|\lambda| \leq c/\max_{1\leq i\leq N}||X_i||_{\psi_1}$ is

$$\min\left\{\frac{t}{2C\sum_{i=1}^{N}||X_i||_{\psi_1}^2}, \frac{c}{\max_{1\leq i\leq N}||X_i||_{\psi_1}}\right\}$$

In addition, we can still use similar method for

$$\mathbb{P}\left(-\sum_{i=1}^{N} X_i \geq t\right)$$

and finally obtain our conclusion. □

As a consequence of above theorem, we have the following corollaries.

Corollary 1.8.1. *Let $X_1, \cdots, X_N$ be independent, mean-zero sub-exponential random variables, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, it concludes that*

$$\mathbb{P}\left(\Big|\sum_{i=1}^{N} a_iX_i\Big| \geq t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2||a||_2^2}, \frac{t}{K||a||_\infty}\right\}\right) \tag{40}$$

where $K := \max_{1\leq i\leq N}||X_i||_{\psi_1}$ and $c > 0$ is a constant.

Corollary 1.8.2 (Bernstein inequality for bounded distributions)**.** *Let $X_1, \cdots, X_N$ be independent, mean-zero sub-exponential random variables such that $|X_i| \leq K$ for all i, then for any $t \geq 0$, it concludes that*

$$\mathbb{P}\left(\Big|\sum_{i=1}^{N} X_i\Big| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right) \tag{41}$$

where $\sigma^2 := \sum_{i=1}^{N} \mathbb{E}[X_i^2]$.

Proof. Consider the numerical inequality of

$$e^x \leq 1 + x + \frac{x^2/2}{1 - |x|/3}$$

for $|x| < 3$. Then we have

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

where $|\lambda| < 3/K$ and $g(\lambda) = \lambda^2/(2 - 2|\lambda|K/3)$. Then we can obtain

$$\mathbb{P}\left(\sum_{i=1}^{N} X_i \geq t\right) \leq \exp\left(-\lambda t\right) \prod_{i=1}^{N} \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \leq \exp\left(g(\lambda)\sigma^2 - \lambda t\right)$$

for all $0 \leq t \leq 3/K$. □

2 Random Vectors in High Dimension

We mainly consider a high-dimensional random variable $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$, where $n \sim 10^4$ or even larger. In fact, high dimensions will cause many difficulties, see notes of high-dimensional statistics. Hence we will introduce some tools to circumvent these difficulties in this section.

2.1 Concentration of the Norm

Suppose X_i are independent, mean-zero and unit variance random variables for $i = 1, \cdots, n$, then the norm of $X = (X_1, \cdots, X_n)$ can be computed by

$$\mathbb{E}\left[||X||_2^2\right] = \sum_{i=1}^{n} \mathbb{E}[X_i^2] = n$$

Hence we may expect that the norm of X is very close to $\sqrt{n}$ and we have the following result.

Theorem 2.1.1 (Concentration of the norm)**.** *Let $X = (X_1, \cdots, X_n)$ be a random vector with independent sub-Gaussian coordinates X_i that satisfy $\mathbb{E}[X_i^2] = 1$. Then*

$$\Big|\Big| ||X||_2 - \sqrt{n} \Big|\Big|_{\psi_2} \leq CK^2 \tag{42}$$

where $K = \max_{1 \leq i \leq n} ||X_i||_{\psi_2}$ and $C > 0$ is a constant.

Proof. Without, we assume $K \geq 1$ and since X_i is sub-Gaussian, then $X_i^2 - 1$ is sub-exponential and

$$\Big|\Big| X_i^2 - \mathbb{E}\left[X_i^2\right] \Big|\Big|_{\psi_1} \leq C||X_i^2||_{\psi_1} = C||X_i||_{\psi_2}^2 \leq CK^2$$

Hence by Bernstein inequality, it concludes that

$$\mathbb{P}\left(\Big|\frac{1}{n}||X||_2^2 - 1\Big| \geq t\right) \leq \exp\left(-\frac{cn}{K^4}\min\left\{t^2, t\right\}\right)$$

Moreover, due to

$$|x-1| \geq \delta \ \text{ implies } \ |x^2 - 1| \geq \max\left\{\delta, \delta^2\right\}$$

we have

$$\mathbb{P}\left(\Big|\frac{1}{\sqrt{n}}||X||_2 - 1\Big| \geq t\right) \leq \mathbb{P}\left(\Big|\frac{1}{n}||X||_2^2 - 1\Big| \geq \max\left\{t, t^2\right\}\right) \leq \exp\left(-\frac{cn}{K^4}t^2\right)$$

As a result, we deduce

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

The proof is complete. □

As a consequence of above theorem, we can give an estimation of the expectation and variance of $||X||_2$.

Corollary 2.1.1. *Deduce from Theorem 2.1.1, it concludes that*

$$\left|\mathbb{E}\left[||X||_2\right] - \sqrt{n}\right| \le C_1K^2 \tag{43}$$

$$\operatorname{var}\left(||X||_2\right) \le C_2K^4 \tag{44}$$

where $C_1, C_2 > 0$ are constants.

Proof. First,

$$\left|\mathbb{E}\left[||X||_2\right] - \sqrt{n}\right| \le \mathbb{E}\left[\left|||X||_2 - \sqrt{n}\right|\right] = \int_0^\infty \mathbb{P}\left(\left|||X||_2 - \sqrt{n}\right| \ge t\right) dt \le 2\int_0^\infty \exp\left(-\frac{ct^2}{K^4}\right) dt$$

Besides,

$$\operatorname{var}\left(||X||_2\right) = \mathbb{E}\left[||X||_2^2\right] - \mathbb{E}\left[||X||_2\right]^2 \le n - \left(\sqrt{n} - C_1K^2\right)^2 \le C_2K^4$$

The proof is complete. □

Remark 2.1.1. *In fact, even though X_i is not sub-Gaussian, we still have*

$$\operatorname{var}\left(||X||_2\right) \le K^4 \tag{45}$$

if $X = (X_1, \cdots, X_n)$ be a random vector with independent coordinates X_i that satisfy $\mathbb{E}[X_i^2] = 1$ and $\mathbb{E}[X_i^4] \le K^4$. It is easy to see since

$$\mathbb{E}\left[\left(||X||_2^2 - n\right)^2\right] = \sum_{i=1}^n \mathbb{E}\left[\left(X_i^2 - 1\right)^2\right] \le K^4n - n$$

On the other hand, due to

$$2\sqrt{n}||X||_2 + \frac{||X||_2^4}{n} = \sqrt{n}||X||_2 + \sqrt{n}||X||_2 + \frac{||X||_2^4}{n} \ge 3||X||_2^2$$

we obtain

$$\mathbb{E}\left[||X||_2 - \sqrt{n}\right]^2 \le \mathbb{E}\left[\left(||X||_2 - \sqrt{n}\right)^2\right] \le \mathbb{E}\left[\left(\frac{||X||_2^2}{\sqrt{n}} - \sqrt{n}\right)^2\right] \le K^4$$

Hence we conclude $\operatorname{var}\left(||X||_2\right) \le K^4$. *In addition, suppose X_i has continuous distribution and its density is uniformly bounded by 1, then for $\epsilon > 0$, we have*

$$\mathbb{P}\left(||X||_2 \le \epsilon\sqrt{n}\right) \le (C\epsilon)^n \tag{46}$$

In fact, since $\sqrt{n}||X||_2 \ge |\sum_{i=1}^n X_i|$, we only need to show that $\mathbb{P}\left(|\sum_{i=1}^n X_i| \le \epsilon n\right) \le (C\epsilon)^n$.

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

where $\lambda > 0$. In addition,

$$\int_{-\infty}^{\infty} e^{-\lambda x} p_i(x) dx = -\frac{1}{\lambda} \int_{-\infty}^{\infty} p_i(x) de^{-\lambda x} \leq -\frac{1}{\lambda} \int_{-\infty}^{\infty} de^{-\lambda x} = \frac{1}{\lambda}$$

Hence we obtain

$$\mathbb{P}\left(-\Big|\sum_{i=1}^{n} X_i\Big| \geq -\epsilon n\right) \leq \left(\frac{e^{\lambda\epsilon}}{\lambda}\right)^n$$

for any $\lambda > 0$. The right side attains its minimum as $\lambda = 1/\epsilon$ and we conclude that $\mathbb{P}\left(|\sum_{i=1}^{n} X_i| \leq \epsilon n\right) \leq (e\epsilon)^n$.

2.2 Covariance Matrices and Principal Component Analysis

For a high-dimensional random variable $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$, the covariance matrix of X is defined by

$$\mathrm{cov}\left(X\right) := \mathbb{E}\left[\left(X-\mu\right)\left(X-\mu\right)^T\right] \tag{47}$$

One important task in high-dimensional probability is Principal Component Analysis (PCA), which needs the spectral decomposition of $\mathrm{cov}\left(X\right)$. Suppose u_1 is the eigenvector corresponding to the largest eigenvalue s_1 defines the first principal direction.

Definition 2.2.1 (Isotropic random vectors)**.** *A random vector X in $\mathbb{R}^n$ is called isotropic if*

$$\Sigma(X) := \mathbb{E}\left[XX^T\right] = \boldsymbol{I}_n \tag{48}$$

where $\boldsymbol{I}_n$ denotes the identity matrix in $\mathbb{R}^n$.

Lemma 2.2.1 (Characterization of isotropy)**.** *A random vector X in $\mathbb{R}^n$ is isotropic if and only if*

$$\mathbb{E}\left[\langle X, x\rangle^2\right] = ||x||_2^2 \quad \text{for all } x \in \mathbb{R}^n \tag{49}$$

Proof. It is easy to see

$$\mathbb{E}\left[\langle X, x\rangle^2\right] = \mathbb{E}\left[x^T\left(XX^T\right)x\right] = x^T\mathbb{E}\left[XX^T\right]x = x^Tx = ||x||_2^2$$

This completes the proof. □

Lemma 2.2.2. *Let X be an isotropic random vector in $\mathbb{R}^n$. Then*

$$\mathbb{E}\left[||X||_2^2\right] = n \tag{50}$$

Moreover, if X and Y are two independent isotropic random vectors in $\mathbb{R}^n$, then

$$\mathbb{E}\left[\langle X, Y\rangle^2\right] = n \tag{51}$$

Proof. First,

$$\mathbb{E}\left[||X||_2^2\right] = \mathbb{E}\left[XX^T\right] = \mathbb{E}\left[\text{tr}\left(XX^T\right)\right] = \mathbb{E}\left[\text{tr}\left(X^TX\right)\right] = \text{tr}\left(\mathbb{E}\left[X^TX\right]\right) = n$$

The second equality due to XX^T is a 1×1 matrix. In addition, we have

$$\mathbb{E}\left[\langle X, Y\rangle^2\right] = \mathbb{E}\left[\mathbb{E}\left[\langle X, Y\rangle^2 \middle| Y\right]\right] = \mathbb{E}\left[YY^T\right] = n$$

This completes the proof. □

Remark 2.2.1. *From above lemma and Theorem 2.1.1, the isotropy random vector X satisfies that $||X||_2 \sim \sqrt{n}$ with high probability. In addition, we have*

$$\mathbb{P}\left(\left|\langle X, Y\rangle\right| \geq t\right) \leq t^{-2}\mathbb{E}\left[\langle X, Y\rangle^2\right] = nt^{-2}$$

hence $\langle X, Y\rangle \sim \sqrt{n}$ with high probability. If we denote $\bar{X} := X/||X||_2$ and $\bar{Y} := Y/||Y||_2$ and we conclude that

$$\langle \bar{X}, \bar{Y}\rangle \sim \frac{1}{\sqrt{n}}$$

Thus, in high-dimensional spaces independent and isotropic random vectors tend to be almost orthogonal. Besides, the distance between independent isotropic vectors have the following result.

$$\mathbb{E}\left[||X - Y||_2^2\right] = 2n \tag{52}$$

In fact, we have

$$\mathbb{E}\left[||X - Y||_2^2\right] = \mathbb{E}\left[\text{tr}\left((X - Y)(X - Y)^T\right)\right] = 2n - \text{tr}\left(\mathbb{E}\left[XY^T\right]\right) - \text{tr}\left(\mathbb{E}\left[YX^T\right]\right)$$

and

$$\mathbb{E}\left[XY^T\right] = \mathbb{E}\left[\mathbb{E}\left[XY^T|X\right]\right] = \mathbb{E}\left[X\mathbb{E}\left[Y^T|X\right]\right] = \mathbb{E}\left[X\mathbb{E}\left[Y^T\right]\right] = 0$$

2.3 Examples of High-Dimensional Distributions

The first example is the spherical distribution, where a random vector X is uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$

$$X \sim \text{Unif}\left(\sqrt{n}S^{n-1}\right) \tag{53}$$

We can show that the coordinates of X is not independent but it is still isotropic, in fact, since X is uniformly distributed on the Euclidean sphere, for any orthogonal matrix Q, we have

$$\mathbb{E}\left[\langle X, x\rangle^2\right] = \mathbb{E}\left[\langle QX, x\rangle^2\right] = \mathbb{E}\left[\langle X, Q^Tx\rangle\right]$$

Next let x_i for $i = 1, \cdots, n$ be an orthogonal system such that $Q_ix = x_i$, where Q_i is an orthogonal matrix and $||x_i||_2 = ||x||_2$, then it concludes that

$$n\mathbb{E}\left[\langle X, x\rangle^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}\langle X, x_i\rangle^2\right] = \mathbb{E}\left[||X||_2^2||x||_2^2\right] = n||x||_2^2$$

Moreover,

$$\mathbb{P}(X_1 > 0, X_2 > 0) = \frac{1}{2} > \frac{1}{4} = \mathbb{P}(X_1 > 0)\,\mathbb{P}(X_2 > 0)$$

Hence the coordinates of X are not independent. Another discrete isotropic distribution in $\mathbb{R}^n$ is the symmetric Bernoulli, that is, $X = (X_1, \cdots, X_n)$ where all X_i are independent, symmetric, Bernoulli random variables. Besides, the standard high-dimensional Gaussian distribution $X = (X_1, \cdots, X_n)$ is also isotropic, where $X_i \sim \mathcal{N}(0, 1)$ are independent. In fact, for an orthogonal matrix Q, we have $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, this property is called *rotation invariance*. In addition, we have

Proposition 2.3.1. *Let $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and $x \in \mathbb{R}^n$, it concludes that*

$$\langle X, x\rangle \sim \mathcal{N}(0, ||x||_2^2) \tag{54}$$

In addition, suppose $X_i \sim \mathcal{N}(0, \sigma_i)$, then

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(0, \sigma^2) \tag{55}$$

where $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$. Finally, let G be an $m \times n$ Gaussian random matrix, i.e.,the entries of G are independent $\mathcal{N}(0, 1)$ random variables. Let $u \in \mathbb{R}^n$ be a fixed unit vector. Then

$$Gu \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m) \tag{56}$$

Above propositions are easy to prove, hence we omit here. The general form of multivariate normal distribution in $\mathbb{R}^n$ is defined by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e. if $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it has $Z := \boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, the density function is

$$f_X(\boldsymbol{x}) := \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

Moreover, let X be a random vector in $\mathbb{R}^n$, then it has a multivariate normal distribution if and only if every one-dimensional marginal $\langle X, \theta\rangle$, $\theta \in \mathbb{R}^n$, has a (univariate) normal distribution.

Proof. Suppose $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have for any $\theta \in \mathbb{R}^n$

$$\langle X, \theta\rangle = \langle \boldsymbol{\Sigma} Z + \boldsymbol{\mu}, \theta\rangle = \langle Z, \boldsymbol{\Sigma}^T\theta\rangle + \langle \boldsymbol{\mu}, \theta\rangle \sim \mathcal{N}(\langle \boldsymbol{\mu}, \theta\rangle, ||\boldsymbol{\Sigma}^T\theta||_2^2)$$

On the other hand, denote another random vector $Y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \mathbb{E}[X], \boldsymbol{\Sigma} := \mathrm{cov}(X)$, then we have $\langle X, \theta\rangle \equiv \langle Y, \theta\rangle$ for any $\theta \in \mathbb{R}^n$, then by Cramér–Wold theorem, it concludes that X, Y have the same distribution. □

From Theorem 2.1.1, we guess that the density of $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ is concentrated in a thin spherical shell around the sphere of radius $\sqrt{n}$. First, if $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, denote $X = r\theta$ as the polar form expression, where $r := ||X||_2$ and $\theta := X/||X||_2$, then it has

Proposition 2.3.2 (Normal and spherical distributions)**.** *The length r and direction θ are independent random variables and θ is uniformly distributed on the unit sphere S_{n-1}.*

Proof. First, $\theta = (\theta_1, \cdots, \theta_n)$, where θ is the angle between X and x_i-th axis, hence the density function can be written

$$g(\theta, r) = \frac{r^{n-1}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} r^2 \cos^2(\theta_i)\right)$$

And the density function of θ is computed by

$$f(\theta) = \int_0^{\infty} g(\theta, r) dr = \frac{1}{(2\pi)^{n/2}} \int_0^{\infty} r^{n-1} e^{-r^2/2} dr = \frac{\Gamma(n/2)}{2\pi^{n/2}}$$

which is a constant and equals to reciprocal of S_{n-1} surface area. □

Since $||X||_2 \sim \sqrt{n}$ by Theorem 2.1.1 and it deduces that

$$\mathcal{N}(\mathbf{0}, \boldsymbol{I}_n) \approx \text{Unif}\left(\sqrt{n} S_{n-1}\right) \tag{57}$$

In other words, the standard normal distribution in high dimensions is close to the uniform distribution on the sphere of radius $\sqrt{n}$.

For an example of a discrete distribution, consider a coordinate random vector X uniformly distributed in the set $\{\sqrt{n} e_i\}_{i=1}^n$, where $\{e_i\}_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$:

$$X \sim \text{Unif}\{\sqrt{n} e_i : i = 1, \cdots, n\} \tag{58}$$

Then X is an isotropic random vector. Although high-dimensional Gaussian distributions are often the most convenient tools, it is still meaningful to consider a general class of discrete, isotropic, distributions arises in the area of signal processing under the name of *frames*.

Definition 2.3.1. *A frame is a set of vectors $\{u_i\}_{i=1}^N$ in $\mathbb{R}^n$ which obeys an approximate Parseval identity, i.e. there exist numbers $A, B > 0$, called frame bounds, such that*

$$A||x||_2^2 \leq \sum_{i=1}^{N} \langle x, u_i \rangle^2 \leq B||x||_2^2 \tag{59}$$

If $A = B$ then the set $\{u_i\}_{i=1}^N$ is called a tight frame.

In fact, we can think of tight frames as generalizations of orthogonal bases without the linear independence requirement. Any orthonormal basis in $\mathbb{R}^n$ is clearly a tight frame and we will show that tight frames correspond to isotropic distributions, and vice versa.

Lemma 2.3.1 (Tight frames and isotropic distributions)**.** *Consider a tight frame $\{u_i\}_{i=1}^N$ with frame bounds $A = B$. Let X be a random vector that is uniformly distributed in the set of frame elements, i.e.*

$$X \sim \text{Unif}\{u_i : i = 1, \cdots, N\}$$

Then $X\sqrt{N/A}$ is an isotropic random vector in $\mathbb{R}^n$. In addition, consider an isotropic random vector Y in $\mathbb{R}^n$ that takes a finite set of values y_i with probabilities p_i each, $i = 1, \cdots, N$. Then the vectors

$$u_i := \sqrt{p_i} y_i$$

form a tight frame in $\mathbb{R}^N$ with bounds $A = B = 1$.

Proof. First, compute

$$\frac{N}{A}\mathbb{E}[XX^T] = \frac{1}{A}\sum_{i=1}^{N} u_i u_i^T = \boldsymbol{I}_n$$

Besides, we have

$$\sum_{i=1}^{N}\langle x, u_i\rangle^2 = \sum_{i=1}^{N} p_i\langle x, y_i\rangle^2 = \mathbb{E}\left[\langle x, Y\rangle^2\right] = ||x||_2^2$$

where $x \in \mathbb{R}^n$ and the last equation due to isotropic of Y. □

Our last example of a high-dimensional distribution comes from convex geometry. Consider a bounded convex set K in $\mathbb{R}^n$ with non-empty interior; such sets are called convex bodies. Let X be a random vector uniformly distributed in K according to the probability measure given by the normalized volume in K:

$$X \sim \text{Unif}(K) \tag{60}$$

2.4 Sub-Gaussian Distributions in Higher Dimensions

We have introduced the concept of sub-Gaussian distributions in one dimension, it is natural to extend it into higher dimensional category. Recall one special property of multivariate normal distribution, i.e. it can be characterized through its one-dimensional marginals. Guided by this characterization, it is natural to define multivariate sub-Gaussian distributions as follows.

Definition 2.4.1 (Sub-Gaussian random vectors)**.** *A random vector X in $\mathbb{R}^n$ is called sub-Gaussian if the one-dimensional marginals $\langle X, x\rangle$ are sub-Gaussian random variables for all $x \in \mathbb{R}^n$. The sub-Gaussian norm of X is defined as*

$$||X||_{\psi_2} := \sup_{x\in S_{n-1}} ||\langle X, x\rangle||_{\psi_2} \tag{61}$$

Lemma 2.4.1 (Sub-Gaussian distributions with independent coordinates)**.** *Suppose $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates X_i. Then X is a sub-Gaussian random vector, and*

$$||X||_{\psi_2} \leq C \max_{1\leq i\leq n} ||X||_{\psi_2} \tag{62}$$

Proof. By Proposition 1.5.1, it concludes that for any $x \in S_{n-1}$

$$||\langle X, x\rangle||_{\psi_2}^2 \leq C\sum_{i=1}^{n} x_i^2||X_i||_{\psi_2}^2 \leq C \max_{1\leq i\leq n} ||X||_{\psi_2}^2$$

This completes our proof. □

Remark 2.4.1. *Let $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with sub-Gaussian coordinates X_i. Then X is a sub-Gaussian random vector. It is easy to see by*

$$\mathbb{E}\left[\exp\left(\langle X, x\rangle^2/t^2\right)\right] \leq \mathbb{E}\left[\exp\left(||X||_2^2/t^2\right)\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\exp\left(nX_i^2/t^2\right)\right]$$

where $x \in S_{n-1}$ and let $t = \sqrt{n}\max_{1\leq i\leq n}||X_i||_{\psi_2}$, it concludes that

$$\mathbb{E}\left[\exp\left(\langle X, x\rangle^2/t^2\right)\right] \leq 2$$

Hence X is a sub-Gaussian random vector and $||X||_{\psi_2} \leq \sqrt{n}\max_{1\leq i\leq n}||X_i||_{\psi_2}$.

Many important high-dimensional distributions are sub-Gaussian, but some are not. We now explore some basic distributions. First if X is symmetric Bernoulli distribution or multivariate normal random vectors, it is easy to show that

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

Hence they are sub-Gaussian. Next for a random vector X with the coordinate distribution is uniformly distributed in the set $\{\sqrt{n}e_i\}_{i=1}^n$, where e_i denotes the n-element set of the canonical basis vectors in $\mathbb{R}^n$, it has that

$$||X||_{\psi_2} \approx \mathcal{O}\left(\sqrt{\frac{n}{\log n}}\right) \tag{63}$$

In fact, we have

$$\mathbb{E}\left[\exp\left(\langle X, x\rangle^2/t^2\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{nx_i^2}{t^2}\right)$$

If $t = c\sqrt{n/\log n}$, where $c \geq \sqrt{\log n/\log 2n}$ is a constant, we have $\mathbb{E}\left[\exp\left(\langle X, x\rangle^2/t^2\right)\right] \leq 2$. Hence we conclude that

$$||X||_{\psi_2} \leq c\sqrt{\frac{n}{\log n}}$$

On the other hand, denote

$$f(t) := \frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{nx_i^2}{t^2}\right) \quad f'(t) = -2\sum_{i=1}^{n}\frac{x_i^2}{t^3}\exp\left(\frac{nx_i^2}{t^2}\right) < 0$$

Hence $f(t)$ is strictly monotone decreasing and the minimum t of $\mathbb{E}\left[\exp\left(\langle X, x\rangle^2/t^2\right)\right] \leq 2$ can only be$\sqrt{n/\log 2n}$. In conclusion, the coordinate distribution is indeed sub-Gaussian, but such a large norm makes it useless to think of X as a sub-Gaussian random vector.

Although sub-Gaussian random vectors that were useful had independent coordinates. This is not necessary, however. A good example is the uniform distribution on the sphere of radius $\sqrt{n}$.

Theorem 2.4.1 (Uniform distribution on the sphere is sub-Gaussian)**.** *Let X be a random vector uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:*

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

Then X is sub-Gaussian, and $||X||_{\psi_2} \leq C$.

Proof. Let $Y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and X has the same distribution as $\sqrt{n}Y/||Y||_2$, hence we have

$$\mathbb{P}\left(|X_1| \geq t\right) = \mathbb{P}\left(\frac{|Y_1|}{||Y||_2} \geq \frac{t}{\sqrt{n}}\right)$$

In addition, due to Theorem 2.1.1, we have

$$\mathbb{P}\left(\left|||Y||_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp\left(-t^2/K_1^2\right)$$

Denote $\mathcal{E} := \{||Y||_2 \geq \sqrt{n}/2\}$ and it concludes that

$$\mathbb{P}\left(\frac{|Y_1|}{||Y||_2} \geq \frac{t}{\sqrt{n}}\right) \leq \mathbb{P}\left(\frac{|Y_1|}{||Y||_2} \geq \frac{t}{\sqrt{n}} \text{ and } \mathcal{E}\right) + \mathbb{P}\left(\mathcal{E}^c\right) \leq \mathbb{P}\left(|Y_1| \geq \frac{t}{2}\right) + 2\exp\left(-nc\right)$$

As a result, we obtain $\mathbb{P}\left(|X_1| \geq t\right) \leq 2\exp\left(-nc\right) + 2\exp\left(-t^2/8\right)$. When $t \leq \sqrt{n}$, we have $\mathbb{P}(|X|_1 \geq t) \leq 4\exp(-Ct^2)$. For $t > \sqrt{n}$, this probability is zero since $|X_1| \leq ||X||_2 = \sqrt{n}$. In conclusion, X_1 is sub-Gaussian and $||X_1||_{\psi_2} \leq \sqrt{n/\log 2}$. Finally, due to the symmetric property, the above argument is valid for all coordinates, which deduces X is also sub-Gaussian and $||X||_{\psi_2} \leq n/\sqrt{\log 2}$. □

Theorem 2.4.2 (Uniform distribution on the Euclidean ball)**.** *Let X be a random vector uniformly distributed on the Euclidean ball in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:*

$$X \sim \text{Unif}\left(\mathcal{B}(0, \sqrt{n})\right)$$

Then X is sub-Gaussian, and $||X||_{\psi_2} \leq C$.

Proof. Let's compute the MGF of $\langle X, x\rangle^2$, where $x \in S_{n-1}$, for $\lambda \in \mathbb{R}$, it has

$$\mathbb{E}\left[\exp\left(\lambda^2\langle X, x\rangle^2\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda^2||X||_2^2\right)\right] \leq \exp\left(\lambda^2 n\right)$$

Hence $\langle X, x\rangle$ is sub-Gaussian for all $x \in S_{n-1}$ and $||\langle X, x\rangle||_{\psi_2} \leq \sqrt{n/\log 2}$, so $||X||_{\psi_2} \leq \sqrt{n/\log 2}$. □

Remark 2.4.2 (Projective limit theorem)**.** *In fact, above two theorems should be compared to the so- called projective central limit theorem. It states that the marginals of the uniform distribution on the sphere become asymptotically normal as n increases. Precisely, if $X \sim$* Unif$(\sqrt{n}S_{n-1})$*, then for any fixed unit vector x we have $\langle X, x\rangle \to \mathcal{N}(0, 1)$ as $n \to \infty$.*

3 Random Matrix

3.1 Preliminaries on Matrices

Since we will mainly deal with matrix in this section, it is necessary to discuss some backgrounds and notations about matrices. For a $m \times n$ matrix A, the *singular value decomposition* of A can be

$$A = \sum_{i=1}^{r} s_i u_i v_i^T \tag{64}$$

where $r = \text{rank}(A)$ and the non-negative s_i is called the *singular values* of A. Besides, $u_i \in \mathbb{R}^m$ ($v_i \in \mathbb{R}^n$) are the left (right) singular vectors of A related to s_i. In fact, s_i can also be regarded as the square root of eigenvectors of AA^T and A^TA. Next, if A is symmetric, by Courant–Fisher min–max theorem, we denote λ_i as the eigenvalues of A and it concludes

$$\lambda_i(A) = \max_{\dim E = i} \min_{x \in S(E)} \langle Ax, x \rangle \tag{65}$$

where the maximum is over all i-dimensional subspaces E of $\mathbb{R}^n$ and $S(E)$ denotes the unit Euclidean sphere in the subspace E. As a result, we obtain

$$s_i(A) = \max_{\dim E = i} \min_{x \in S(E)} ||Ax||_2 \tag{66}$$

Proposition 3.1.1. *Suppose A is invertible and has the following singular value decomposition*

$$A = \sum_{i=1}^{r} s_i u_i v_i^T$$

then

$$A^{-1} = \sum_{i=1}^{r} s_i^{-1} v_i u_i^T$$

In general, we usually order all singular values as $s_1 \geq s_2 \geq \cdots \geq s_r \geq 0$. Remind the operator norm of matrix $A : l_2^n \to l_2^m$ defined by

$$||A|| = \max_{x \in S_{n-1}} ||Ax||_2 \tag{67}$$

We can show that $s_1(A) \equiv ||A||$. For $m > n$, although the rank $r < n$ if A is not full rank, we usually denote $s_i(A) = 0$ for $r < i \leq n$. Therefore we still have $s_1(A) \geq \cdots \geq s_n(A) \geq 0$, and the smallest singular value $s_n(A)$ has special meaning, that is, $s_n(A) > 0$ is equivalent to A is full rank.

The Frobenius norm is also widely used, which is defined by

$$||A||_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{i,j}|^2 \right)^{1/2} \tag{68}$$

In terms of singular values, the Frobenius norm can be computed as

$$||A||_F = \left(\sum_{i=1}^{r} s_i(A)^2\right)^{1/2} \tag{69}$$

In fact, the canonical inner product on $\mathbb{R}^{m\times n}$ can be represented in terms of matrices as

$$\langle A, B\rangle = \operatorname{tr}\left(A^T B\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j}B_{i,j}$$

Hence we have $||A||_F^2 = \langle A, A\rangle$. As a result, we can conclude that $||A|| = ||s||_\infty$ and $||A||_F = ||s||_2$. In addition, since $||s||_\infty \leq ||s||_2 \leq \sqrt{r}||s||_\infty$, we obtain the best possible relation between the operator and Frobenius norms:

$$||A|| \leq ||A||_F \leq \sqrt{r}||A|| \tag{70}$$

Remark 3.1.1. *In fact, for any matrix A, we have the following bound for its k-th singular value, where $1 \leq k \leq r$.*

$$s_k(A) \leq \frac{1}{\sqrt{k}}||A||_F \tag{71}$$

From the definition of $||A||_F$, we have

$$||A||_F^2 = \sum_{i=1}^{r} s_i(A)^2 \geq \sum_{i=1}^{k} s_i(A)^2 \geq ks_k(A)^2$$

Next, we focus on the approximation of one matrix by a lower rank matrix with the same size. One essential question is how to choose the best approximation A_k of A, where $\operatorname{rank}(A_k) = k < r = \operatorname{rank}(A)$. We can use Frobenius norm to measure the difference between two matrices. In fact, the *Eckart–Young–Mirsky* theorem gives the answer to this low-rank approximation problem. It states that the minimizer A_k is obtained by truncating the singular value decomposition of A at the k-th term:

$$A_k := \sum_{i=1}^{k} s_i(A)u_i v_i^T$$

i.e. $||A - A_k|| = \min_{\operatorname{rank}(A')=k} ||A - A'||$.

Proof. For A' such that $\operatorname{rank}(A') = k$, there exists γ_i where $i = 1, \cdots, k+1$ such that

$$\sum_{i=1}^{k+1} \gamma_i A' v_i = 0 \quad \sum_{i=1}^{k+1} \gamma_i^2 = 1$$

Denote $w = \sum_{i=1}^{k+1} \gamma_i v_i$ and

$$||A - A'||^2 \geq ||(A - A')w||_2^2 = ||Aw||_2^2 = \sum_{i=1}^{k+1} \gamma_i^2 s_i(A)^2 \geq s_{k+1}(A)^2 = ||A - A_k||^2$$

This competes our proof. □

3.2 Nets, Covering Numbers, and Packing Numbers

Next, we will focus on a simple but powerful method–an ϵ-net argument, which is useful for the theory of random matrix. Let's see the definition first.

Definition 3.2.1 (ϵ-Net)**.** *Let (T, d) be a metric space. Consider a subset $K \subset T$ and let $\epsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ϵ-net of K if every point in K is within a distance of some point of $\mathcal{N}$, i.e.*

$$\forall x \in K \;\; \exists x_0 \in \mathcal{N} : \; d(x, x_0) < \epsilon$$

Equivalently, $\mathcal{N}$ is an ϵ-net of K if and only if K can be covered by balls with centers in $\mathcal{N}$ and radius ϵ.

Definition 3.2.2 (Covering numbers)**.** *The smallest possible cardinality of an ϵ-net of K is called the covering number of K and is denoted $\mathcal{N}(K, d, \epsilon)$. Equivalently, $\mathcal{N}(K, d, \epsilon)$ is the smallest number of closed balls with centers in K and radius ϵ whose union covers K.*

Remark 3.2.1 (Compactness)**.** *An important result in real analysis states that a subset K of a metric space (T, d) is pre-compact (i.e., the closure of K is compact) if and only if*

$$\mathcal{N}(K, d, \epsilon) < \infty \;\; \forall \epsilon > 0$$

Thus we can think of the magnitude $\mathcal{N}(K, d, \epsilon)$ as a quantitative measure of the compactness of K.

Definition 3.2.3 (Packing numbers)**.** *A subset $\mathcal{N}$ of a metric space (T, d, ϵ) is ϵ-separated if $d(x, y) > \epsilon$ for all distinct points $x, y \in \mathcal{N}$. The largest possible cardinality of an ϵ-separated subset of a given set $K \subset T$ is called the packing number of K and is denoted $\mathcal{P}(K, d, \epsilon)$.*

Remark 3.2.2. *Suppose that T is a normed space, then $P(K, d, \epsilon)$ is the largest number of closed disjoint balls with centers in K and radius $\epsilon/2$. However, for a general metric space, this conclusion is not valid.*

In fact, since the normed space satisfies the first axiom of countability, the number of closed disjoint balls with centers in K and radius $\epsilon/2$ is at most countable. First, it is easy to conclude this result from the definition of packing numbers if $P(K, d, \epsilon) < \infty$. So let $P(K, d, \epsilon) = \infty$, suppose we have $n < \infty$ such closed balls, we can indeed construct a new ball such that all these $n + 1$ balls are disjoint. Then we know that the number of closed disjoint balls with centers in K and radius $\epsilon/2$ is at least countable.

Lemma 3.2.1. *If $\mathcal{N}$ be a maximal ϵ-separated subset of K, then $\mathcal{N}$ is an ϵ-net of K.*

Proof. If $\mathcal{N}$ is not an ϵ-net of K, then there exists $x_0 \in K \backslash \mathcal{N}$ such that $d(x_0, y) > \epsilon$ for any $y \in \mathcal{N}$, which is contradiction since $\mathcal{N}$ is maximal ϵ-separated of K. □

Lemma 3.2.2 (Equivalence of covering and packing numbers)**.** *For any set $K \subset T$ and any $\epsilon > 0$, we have*

$$\mathcal{P}(K, d, 2\epsilon) \leq \mathcal{N}(K, d, \epsilon) \leq \mathcal{P}(K, d, \epsilon)$$

Proof. For the upper bound, it is just a consequence of above lemma. To prove the lower bound, let $\mathcal{P}_{2\epsilon} = \{x_i\}$ be the 2ϵ-separated subset of K and as well $\mathcal{N}_\epsilon = \{y_j\}$ be ϵ-net of K, then by definition, for any y_j, there must have a unique y_i such that $x_i \in \mathcal{B}(y_i, \epsilon)$. If not, suppose $x \in \mathcal{B}(y_1, \epsilon) \bigcap \mathcal{B}(y_2, \epsilon)$, which deduces that $d(y_1, y_2) < 2\epsilon$, it is a contradiction. Hence, $\mathcal{P}_{2\epsilon} \to \mathcal{N}_\epsilon$ is injective. □

Until now, we have define the covering numbers of a subset K in metric space T, however, we only allow the covering ball with its center in K, what happens if this restriction is relaxed? So we define the exterior covering number $\mathcal{N}^{\text{ext}}(K, d, \epsilon)$ similarly but without requiring that $x_i \in K$, and we conclude

$$\mathcal{N}^{\text{ext}}(K, d, \epsilon) \leq \mathcal{N}(K, d, \epsilon) \leq \mathcal{N}^{\text{ext}}(K, d, \epsilon/2) \tag{72}$$

It is clear to see the first inequality due to the restriction. Next, denote $\mathcal{N}^{\text{ext}}_{\epsilon/2} = \{x_i\}$ and $\mathcal{N}_\epsilon = \{y_j\}$, then for any x_i, there must exist a

3.3 Upper Bounds on Random Sub-Gaussian Matrices

In this subsection, we will focus on random matrix theory, the core work is to investigate the non-asymptotic property of random matrix, one useful method is to investigate the distribution of singular values. Before doing this, some background about the operator norm of a matrix and ϵ-net is necessary.

Lemma 3.3.1 (Computing the operator norm on a net)**.** *Let A be an $m \times n$ matrix and $\epsilon \in [0, 1)$. Then, for any ϵ-net $\mathcal{N}$ of the sphere S_{n-1}, we have*

$$\sup_{x \in \mathcal{N}} ||Ax||_2 \leq ||A|| \leq \frac{1}{1-\epsilon} \sup_{x \in S_{n-1}} ||Ax||_2 \tag{73}$$

Proof. The first inequality is easy to see. Since $||A|| := \max_{x \in S_{n-1}} ||Ax||_2$, then there exists $x_0 \in S_{n-1}$ such that $||Ax_0||_2 = ||A||$. In addition, there exists $y_0 \in \mathcal{N}$ such that $||y_0 - x_0||_2 < \epsilon$, hence

$$\big| ||Ay_0||_2 - ||Ax_0||_2 \big| \leq ||A(y_0 - x_0)||_2 \leq \epsilon ||A||$$

As a result, we obtain

$$\big| ||A|| - ||Ay_0||_2 \big| \leq \big| ||A|| - ||Ax_0||_2 \big| + \big| ||Ay_0||_2 - ||Ax_0||_2 \big| \leq \epsilon ||A||$$

In other words,

$$(1-\epsilon)||A|| \leq ||Ay_0||_2 \leq \sup_{x \in \mathcal{N}} ||Ax||_2$$

This completes our proof. □

Remark 3.3.1. *Similar as above lemma, we have the following result for $x \in \mathbb{R}^n$*

$$\sup_{y \in \mathcal{N}} \langle x, y \rangle \leq ||x||_2 \leq \frac{1}{1-\epsilon} \sup_{y \in \mathcal{N}} \langle x, y \rangle \tag{74}$$

where $\mathcal{N}$ is the ϵ-net of S_{n-1}. In addition, let A be a $m \times n$ matrix and $\epsilon \in [0, 0.5)$, we have

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \leq ||A|| \leq \frac{1}{1-2\epsilon} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \tag{75}$$

where $\mathcal{M}$ is the ϵ-net of S_{m-1}. If $m \equiv n$ and A is symmetric, then

$$\sup_{x \in \mathcal{N}} \big|\langle Ax, x \rangle\big| \leq ||A|| \leq \frac{1}{1-2\epsilon} \sup_{x \in \mathcal{N}} \big|\langle Ax, x \rangle\big| \tag{76}$$

Proof. It is easy to show the left side of these three inequalities, for the first, there exists $y_0 \in \mathcal{N}$ such that $||y_0 - x/||x||_2||_2 \leq \epsilon$, denote $y_0 := x/||x||_2 + e_\epsilon$, where $e_\epsilon \in \mathbb{R}^n$, hence

$$||x||_2 = \langle x, x/||x||_2 \rangle = \langle x, y_0 - e_\epsilon \rangle \leq \langle x, y_0 \rangle + \epsilon ||x||_2$$

For the second, due to the definition of $||A|| := \max_{x \in S_{n-1}, y \in S_{m-1}} \langle Ax, y \rangle$, denote $x_0 \in S_{n-1}$ and $y_0 \in S_{m-1}$ such that $||A|| = \langle Ax_0, y_0 \rangle$, then there exists $x_\epsilon := x_0 + e_1 \in \mathcal{N}$ and $y_\epsilon := y_0 + e_2 \in \mathcal{M}$ such that $||x_0 - x_\epsilon||_2 < \epsilon$ and $||y_0 - y_\epsilon||_2 < \epsilon$.

$$\langle Ax_0, y_0 \rangle = \langle A(x_\epsilon - e_1), y_\epsilon - e_2 \rangle \leq \langle Ax_\epsilon, y_\epsilon \rangle + ||Ae_1||_2 + \epsilon ||Ax_\epsilon||_2 \leq \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle + 2\epsilon ||A||$$

And we can obtain the third inequality by similar method. □

Next, we will focus on the estimation of sub-Gaussian matrices.

Theorem 3.3.1 (Norm of matrices with sub-Gaussian entries)**.** *Let A be an $m \times n$ random matrix whose entries $A_{i,j}$ are independent mean-zero sub-Gaussian random variables. Then, for any $t > 0$ we have*

$$||A|| \leq CK\left(\sqrt{m} + \sqrt{n} + t\right) \tag{77}$$

with probability at least $1 - 2\exp(-t^2)$ and $K = \max_{i,j} ||A_{i,j}||_{\psi_2}$.

Proof. First, we give an estimation of $\mathcal{N}(K, \epsilon)$. In fact, the following is valid.

$$\frac{|K|}{|\epsilon \mathcal{B}_2^n|} \leq \mathcal{N}(K, \epsilon) \leq \mathcal{P}(K, \epsilon) \leq \frac{|K + (\epsilon/2)\mathcal{B}_2^n|}{|(\epsilon/2)\mathcal{B}_2^n|} \tag{78}$$

where $K + (\epsilon/2)\mathcal{B}_2^n$ is Minkowski sum, which is defined by

$$A + B := \{a + b : a \in A, b \in B\} \tag{79}$$

where A, B are subsets of $\mathbb{R}^n$. The left side is easy to obtain, for the right side, due to the definition of $\mathcal{P}(K, \epsilon)$, we can find $\mathcal{P}(K, \epsilon)$ disjoint closed balls with center in K and radius of $\epsilon/2$, which all contained in $K + (\epsilon/2)\mathcal{B}_2^n$, then we obtain the right side.

Next, we choose K to be the ϵ-net of S_{m-1} and S_{n-1} separately, denoted by $\mathcal{M}$ and $\mathcal{N}$. If $\epsilon = 0.25$ and we have $|\mathcal{M}| \leq 9^m$, $|\mathcal{N}| \leq 9^n$. Then for $x \in \mathcal{M}, y \in \mathcal{M}$, we have

$$\big|\big|\langle Ax, y \rangle\big|\big|_{\psi_2}^2 = \Big|\Big| \sum_{i=1}^m \sum_{j=1}^n A_{i,j} x_i y_j \Big|\Big|_{\psi_2}^2 \leq C \sum_{i=1}^m \sum_{j=1}^n \big|\big|A_{i,j} x_i y_j\big|\big|_{\psi_2}^2 \leq CK^2$$

As a result, we conclude that for $u \geq 0$

$$\mathbb{P}\left(\langle Ax, y\rangle \geq u\right) \leq 2\exp\left(-cu^2/K^2\right)$$

Hence

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

Choose $u = CK\left(\sqrt{m}+\sqrt{n}+t\right)$ such that $cu^2/K^2 \geq 3(m+n)+t^2$, then we have

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

Finally, by above remark, $\mathbb{P}\left(||A|| \geq u\right) \leq 2\exp(-t^2)$. □

Corollary 3.3.1. *As a consequence of above theorem, we have*

$$\mathbb{E}\left[||A||\right] \leq CK\left(\sqrt{m}+\sqrt{n}\right) \tag{80}$$

If the entries $A_{i,j}$ have unit variances

$$\mathbb{E}\left[||A||\right] \geq C\left(\sqrt{m}+\sqrt{n}\right) \tag{81}$$

Proof. First, we have

$$\mathbb{E}\left[||A||\right] \leq CK\left(\sqrt{m}+\sqrt{n}\right) + \int_{CK\left(\sqrt{m}+\sqrt{n}\right)}^{\infty}\mathbb{P}\left(||A|| \geq u\right)du \leq CK\left(\sqrt{m}+\sqrt{n}\right)+\sqrt{\pi}$$

In addition, since

$$||A|| := \max_{x\in S_{n-1},y\in S_{m-1}}\langle Ax, y\rangle$$

choose $x = (1, 0, \cdots, 0)^T$ and $y = (1, 0, \cdots, 0)^T$, we have $||A|| \geq A_{1,1}$ □

3.4 Two-Sided Bounds on Sub-Gaussian Matrices

We have given an upper bound on the spectrum of an $m \times n$ matrix A with independent sub-Gaussian entries. In detail, it has high probability that

$$s_1(A) \leq C(\sqrt{m}+\sqrt{n})$$

We are going to show that the two-sides bound of $s_i(A)$ by

$$\sqrt{m} - C\sqrt{n} \leq s_i(A) \leq \sqrt{m}+C\sqrt{n}$$

In detail, the following result is valid.

Theorem 3.4.1 (Two-sided bound on sub-Gaussian matrices)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent mean-zero sub-Gaussian isotropic random vectors in $\mathbb{R}^n$. Then, for any $t \geq 0$ we have*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t) \tag{82}$$

with probability at least $1 - 2\exp(-t^2)$ *and* $K := \max_i ||A_i||_{\psi_2}$.

Proof. We will show a stronger conclusion in this proof, i.e.

$$\Big|\Big|\frac{1}{m}A^TA - \boldsymbol{I}_n\Big|\Big| \leq K^2 \max\left\{\delta, \delta^2\right\}, \text{ where } \delta := C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right) \tag{83}$$

Let $\mathcal{N}$ be the 0.25-net of S_{n-1}, then we have $|\mathcal{N}| \leq 9^n$. In addition,

$$\Big|\Big|\frac{1}{m}A^TA - \boldsymbol{I}_n\Big|\Big| \leq 2\sup_{x\in\mathcal{N}}\Big|\Big\langle\frac{1}{m}A^TA - \boldsymbol{I}_n x, x\Big\rangle\Big| = 2\sup_{x\in\mathcal{N}}\Big|\frac{1}{m}||Ax||_2^2 - 1\Big|$$

Since A_i be the row of A and $||Ax||_2^2 = \sum_{i=1}^n \langle A_i, x\rangle^2 := \sum_{i=1}^n X_i^2$, where $X_i := \langle A_i, x\rangle$ is also sub-Gaussian such that $\mathbb{E}[X_i^2] = 1$. Hence, $X_i^2 - 1$ is sub-exponential and mean-zero. Moreover, $||X_i||_{\psi_2} \leq K$, i.e. $||X_i^2 - 1||_{\psi_1} \leq CK$. As a result, we obtain from Theorem 1.8.1

$$\mathbb{P}\left(\Big|\frac{1}{m}||Ax||_2^2 - 1\Big| \geq \frac{\epsilon}{2}\right) \leq 2\exp\left(-cm\min\left\{\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2}\right\}\right) \leq 2\exp\left(-cC^2(n + t^2)\right)$$

where $\epsilon := K^2\max\{\delta^2, \delta\}$. Finally, we conclude that

$$\mathbb{P}\left(\sup_{x\in\mathcal{N}}\Big|\frac{1}{m}||Ax||_2^2 - 1\Big| \geq \frac{\epsilon}{2}\right) \leq 9^n 2\exp\left(-cC^2(n + t^2)\right) \leq 2\exp\left(-t^2\right)$$

The last inequality is valid when C is large enough. That completes our proof. □

Corollary 3.4.1. *As a consequence of above theorem, we have*

$$\mathbb{E}\left[\Big|\Big|\frac{1}{m}A^TA - \boldsymbol{I}_n\Big|\Big|\right] \leq CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right) \tag{84}$$

In addition, we have

$$\sqrt{m} - CK^2\sqrt{n} \leq \mathbb{E}\left[s_n(A)\right] \leq \mathbb{E}\left[s_1(A)\right] \leq \sqrt{m} + CK^2\sqrt{n} \tag{85}$$

Proof. From above theorem, we have

$$\begin{aligned}
\mathbb{E}\left[\Big|\Big|\frac{1}{m}A^TA - \boldsymbol{I}_n\Big|\Big|\right] &= \int_0^\infty \mathbb{P}\left(\Big|\Big|\frac{1}{m}A^TA - \boldsymbol{I}_n\Big|\Big| \geq s\right)ds \\
&\leq CK^2\sqrt{\frac{n}{m}} + \int_{CK^2\sqrt{n/m}}^\infty 2\exp\left(-\left(\frac{s\sqrt{m}}{CK^2} - \sqrt{n}\right)^2\right)ds \\
&= CK^2\sqrt{\frac{n}{m}} + \frac{\sqrt{\pi m}}{CK^2} \leq CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right)
\end{aligned}$$

The last inequality is valid when C is large enough. Moreover, we have

$$\begin{aligned}\mathbb{E}\left[s_n(A)\right] &= \int_0^\infty \mathbb{P}\left(s_n(A) \geq s\right) ds \geq \int_0^{\sqrt{m}-CK^2\sqrt{n}} \mathbb{P}\left(s_n(A) \geq s\right) ds \\ &\geq CK^2 \int_0^{\sqrt{m}/CK^2-\sqrt{n}} 1 - 2\exp\left(-t^2\right) dt = \sqrt{m} - CK^2\sqrt{n}\end{aligned}$$

□

4 Concentration Without Independence

In this section, we mainly focus on such concentration problem, as we have known in last section, the Gaussian random vector $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, it has

$$X \approx \mathbb{E}[X]$$

with high probability. But if we change X by $f(X)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a general function, do we still have such probability? In fact, if f is linear, it is clear to see. How about f is non-linear? In order to answer this question, we may guess that f does not oscillate too wildly, so it is necessary to introduce *Lipschitz function*.

4.1 Concentration of Lipschitz Functions for the Sphere

The main result of this subsection is the following theorem.

Theorem 4.1.1 (Concentration of Lipschitz functions on the sphere)**.** *Consider a random vector $X \sim \mathrm{Unif}(\sqrt{n}S_{n-1})$, i.e. X is uniformly distributed on the Euclidean sphere of radius $\sqrt{n}$. Consider a Lipschitz function $f : \sqrt{n}S_{n-1} \to \mathbb{R}$. Then*

$$||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq C||f||_{\mathrm{Lip}} \tag{86}$$

To prove about theorem, the following preliminaries are necessary.

Theorem 4.1.2 (Isoperimetric inequality on $\mathbb{R}^n$)**.** *Among all subsets $A \subset \mathbb{R}^n$ with given volume, Euclidean balls have minimal area. Moreover, for any ϵ, Euclidean balls minimize the volume of the ϵ-neighborhood of A, defined as*

$$A_\epsilon := \{x \in \mathbb{R}^n : \exists y \in A \text{ such that } ||x-y||_2 \leq \epsilon\} = A + \epsilon\mathcal{B}_2^n \tag{87}$$

Lemma 4.1.1 (Blow-up)**.** *Let A be a subset of the sphere $\sqrt{n}S_{n-1}$, and let σ denote the normalized area on that sphere. If $\sigma(A) \geq 0.5$ then, for every $t \geq 0$*

$$\sigma(A_t) \geq 1 - 2\exp(-ct^2) \tag{88}$$

where c is a constant.

Proof. Denote $H := \{x \in \sqrt{n}S_{n-1} : x_1 \leq 0\}$, by assumption, $\sigma(A) \geq 0.5 = \sigma(H)$, hence, $\sigma(A_t) \geq \sigma(H_t)$. Let $X \sim \mathrm{Unif}(\sqrt{n}S_{n-1})$, we have $\sigma(H_t) = \mathbb{P}(X \in H_t)$. Next, by the definition of H_t, we know $\{x \in \sqrt{n}S_{n-1} : x_1 \leq t/2\} \subset H_t$. In fact, suppose $y \in \sqrt{n}S_{n-1}$ and $0 < y_1 \leq t/2$. Denote $y' \in \sqrt{n}S_{n-1}$ such that $y'_i \equiv y_i$ for $i = 2, \cdots, n$ and $y'_1 = y_1$, then it concludes that

$$||y - y'||_2 = |y_1 - y'_1| \leq t$$

As a result, we obtain

$$\sigma(H_t) \geq \mathbb{P}\left(X_1 \leq \frac{t}{2}\right) \geq 1 - 2\exp\left(-ct^2\right)$$

The last inequality is valid due to $||X_1||_{\psi_2} \leq ||X||_{\psi_2} \leq C$ in Theorem 2.4.1. □

Corollary 4.1.1 (Blow-up of exponentially small sets)**.** *Let A be a subset of the sphere $\sqrt{n}S_{n-1}$ such that*

$$\sigma(A) \geq 2\exp\left(-cs^2\right) \quad \text{for some } s > 0$$

Then $\sigma(A_s) > 0.5$ and for any $t \geq s$, $\sigma(A_{2t}) \geq 1 - \exp(ct^2)$, where c is a constant deduced from above lemma.

Proof. Suppose $\sigma(A_s) < 0.5$, then $\sigma(A_s^c) > 0.5$ and by Lemma 4.1.1, we obtain $\sigma\left((A_s^c)_s\right) > 1 - \exp(-cs^2)$, i.e. $\sigma\left((A_s^c)_s^c\right) < 2\exp(-cs^2)$. However, for any $x \in A$ and $y \in A_s^c$, it concludes that $||x - y||_2 > s$, i.e. $x \notin (A_s^c)_s$, hence $x \in (A_s^c)_s^c$, i.e. $A \subset (A_s^c)_s^c$, which is contradiction since $\sigma(A) \geq 2\exp(-cs^2)$. Furthermore, □

Now we are ready to prove Theorem 4.1.1.

Proof. Without loss of generality, assume $||f||_{\text{Lip}} = 1$, then let M be the median of $f(X)$, i.e.

$$\mathbb{P}\left(f(X) \leq M\right) \geq 0.5 \ \text{ and } \mathbb{P}\left(f(X) \geq M\right) \geq 0.5$$

Denote

$$A := \left\{x \in \sqrt{n}S_{n-1} : f(x) \leq M\right\}$$

By Lemma 4.1.1, we obtain $\mathbb{P}\left(X \in A_t\right) \geq 1 - 2\exp(-ct^2)$. On the other hand, for $X \in A_t$, there exists $y \in A$ such that $||X - y||_2 \leq t$, as a result, we obtain

$$f(X) \leq f(y) + ||X - y||_2 \leq M + t$$

That is $\mathbb{P}\left(X \in A_t\right) \leq \mathbb{P}\left(f(X) \leq M + t\right)$, hence $\mathbb{P}\left(f(X) \leq M + t\right) \geq 1 - 2\exp(-ct^2)$. Replace f by $-f$, we still have $\mathbb{P}\left(f(X) \geq M - t\right) \geq 1 - 2\exp(-ct^2)$, i.e. $||f(X) - M||_{\psi_2} \leq C$. Finally, by Lemma 1.5.1, we have

$$||f(X) - \mathbb{E}\left[f(X)\right]||_{\psi_2} = ||f(X) - M + M - \mathbb{E}\left[f(X)\right]||_{\psi_2} \leq C'||f(X) - M||_{\psi_2} \leq C'C$$

This completes our proof. □

As a consequence of Theorem 4.1.1, we have

Corollary 4.1.2 (Concentration for the unit sphere)**.** *For $X \sim \text{Unif}(S_{n-1})$ and f is a Lipschitz function on S_{n-1}, then it has*

$$||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq \frac{C||f||_{\text{Lip}}}{\sqrt{n}} \tag{89}$$

In other words, we have

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2\exp\left(-\frac{cnt^2}{||f||_{\text{Lip}}^2}\right) \tag{90}$$

Proof. Assume $||f||_{\text{Lip}} = 1$ as in above theorem, since $X \sim \text{Unif}(S_{n-1})$, it deduces that $Y := \sqrt{n}X \sim \text{Unif}(\sqrt{n}S_{n-1})$. Next, denote g defined on $\sqrt{n}S_{n-1}$ such that $g(y) = f(x)$, where $y = \sqrt{n}x$. Then

$$||g(y_1) - g(y_2)||_2 = ||f(x_1) - f(x_2)||_2 \leq ||x_1 - x_2||_2 = \frac{1}{\sqrt{n}}||y_1 - y_2||_2$$

Hence, we know $||g||_{\text{Lip}} = n^{-1/2}$, by Theorem 4.1.1, it concludes that

$$||g(Y) - \mathbb{E}[g(Y)]||_{\psi_2} \leq \frac{C}{\sqrt{n}}$$

Finally, since $f(X) = g(Y)$, it completes our proof. □

Corollary 4.1.3 (Concentration about the expectation and concentration about the median are equivalent)**.** *Consider a random variable Z with median M, then*

$$c||Z - \mathbb{E}[Z]||_{\psi_2} \leq ||Z - M||_{\psi_2} \leq C||Z - \mathbb{E}[Z]||_{\psi_2} \tag{91}$$

where $c, C > 0$ are some absolute constants.

Proof. For both sides, it is easy to conclude by Lemma 1.5.1. □

Corollary 4.1.4 (Concentration and blow-up are equivalent)**.** *Consider a random vector X taking values in some metric space (T, d). Assume that there exists a $K > 0$ such that*

$$||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq K||f||_{\text{Lip}} \tag{92}$$

for every Lipschitz function $f : T \to \mathbb{R}$. For a subset $A \subset T$, define $\sigma(A) := \mathbb{P}(X \in A)$. (Then σ is a probability measure on T.) Then if $\sigma(A) \geq 0.5$, for every $t \geq 0$,

$$\sigma(A_t) \geq 1 - 2\exp\left(-ct^2/K^2\right) \tag{93}$$

where $c > 0$ is an absolute constant.

Proof. We choose a special Lipschitz function f defined by

$$f(x) = \text{dist}(x, A) := \inf_{z \in A} d(x, z)$$

It is easy to see f is Lipschitz, in fact,

$$f(x) \leq \inf_{z \in A} \{d(x, y) + d(y, z)\} \leq f(y) + d(x, y)$$

i.e. $f(x) - f(y) \leq d(x, y)$. Similarly, we can deduce $f(y) - f(x) \leq d(x, y)$. Hence, $||f||_{\text{Lip}} = 1$. Now, it is easy to see $A_t = \{y \in T : f(y) \leq t\}$. Moreover, since $\sigma(A) \geq 0.5$, the median M of $f(X)$ equals to zero. By above corollary, we obtain $||f(X)||_{\psi_2} \leq CK||f||_{\text{Lip}}$, i.e.

$$\sigma(A_t) = \mathbb{P}(|f(X)| \leq t) \geq 1 - 2\exp\left(-\frac{ct^2}{K^2||f||_{\text{Lip}}^2}\right)$$

That completes our proof. □

4.2 Concentration for Other Metric Measure Spaces

In this subsection, we will extend the concentration inequalities into other metric measure space.

Theorem 4.2.1 (Gaussian isoperimetric inequality)**.** *Let $\epsilon > 0$ and $A \subset \mathbb{R}^n$ with fixed Gaussian measure $\gamma_n(A)$, which is defined by*

$$\gamma_n(A) := \frac{1}{(2\pi)^{n/2}} \int_A \exp\left(-||x||_2^2/2\right) dx \tag{94}$$

the half-spaces minimize the Gaussian measure of the neighborhood $\gamma_n(A_\epsilon)$.

Proof. In detail, we need to show if $\gamma_n(A) = \gamma_n(H)$, where A is a Borel set and H is any half-space, then $\gamma_n(A_\epsilon) \geq \gamma_n(H_\epsilon)$ for any $\epsilon > 0$. First, choose $R \in SO(n)$ such that $RH := \{Rx : x \in H\} = \{x_1 < t : x \in \mathbb{R}^n\}$, hence

$$\gamma_n(H) = \gamma_n(RH) = \Phi(t)$$

where $\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^t e^{-x^2/2} dx$. And $\gamma_n(H_\epsilon) = \Phi(t+\epsilon)$. As a result,

$$\Phi^{-1}(\gamma_n(A)) + \epsilon = \Phi^{-1}(\gamma_n(H)) + \epsilon = t + \epsilon = \Phi^{-1}(\gamma_n(H_\epsilon))$$

□

Theorem 4.2.2 (Gaussian concentration)**.** *Let $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and a Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, it concludes that*

$$||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq C||f||_{\text{Lip}} \tag{95}$$

Proof. As before, we assume that $||f||_{\text{Lip}} = 1$. Let M be the median of $f(X)$ and denote $A := \{x \in \mathbb{R}^n : f(x) \leq M\}$, hence $\gamma_n(A) = 0.5$ then for $X \in A_\epsilon$, there exists $y \in A$ such that

$$f(X) \leq f(y) + ||f||_{\text{Lip}}||X - y||_2 \leq M + t$$

Hence $\mathbb{P}(X \in A_\epsilon) \leq \mathbb{P}(f(X) \leq M + t)$. By Theorem 4.2.1, it concludes that

$$\mathbb{P}(X \in A_\epsilon) = \gamma_n(A_\epsilon) \geq \Phi\left(\Phi^{-1}(\gamma_n(A)) + \epsilon\right) = \Phi(\epsilon)$$

As a result, we obtain

$$\mathbb{P}(|f(X) - M| \geq t) \leq 2(1 - \Phi(\epsilon)) \leq 2\exp\left(-\frac{\epsilon^2}{2}\right)$$

Hence, $||f(X) - M||_{\psi_2} \leq C$. That completes our proof □

Corollary 4.2.1 (Replacing expectation by L^p norm)**.** *In fact, we can replace $\mathbb{E}[f(X)]$ by $\mathbb{E}[f^p(X)]^{1/p}$, for any $p > 0$ and for any non-negative function f. The constants may depend on p.*

Proof. We only need to show, if X is sub-Gaussian, it concludes that

$$\mathbb{E}\left[f^p(X)\right]^{1/p} \leq C\sqrt{p}||X||_{\psi_2}$$

see Proposition 1.4.1. □

4.3 Matrix Bernstein Inequality

In this subsection, we will show that how to generalize concentration inequalities for sums of independent random variables $\sum_i X_i$ to sums of independent *random matrices.* As a preliminary, we first introduce some matrix calculus.

Theorem 4.3.1 (Golden–Thompson inequality)**.** *Let A and B be $n \times n$ symmetric matrix, then it concludes that*

$$\operatorname{tr}\left(e^{A+B}\right) \leq \operatorname{tr}\left(e^{A}e^{B}\right) \tag{96}$$

Theorem 4.3.2 (Lieb's inequality)**.** *Let H be an $n \times n$ symmetric matrix. Define the function on matrices*

$$f(X) := \exp\left(H + \log X\right) \tag{97}$$

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

Lemma 4.3.1 (Lieb's inequality for random matrices)**.** *Let H be a fixed $n \times n$ symmetric matrix and Z be a random $n \times n$ symmetric matrix. Then*

$$\mathbb{E}\left[\operatorname{tr}\left(\exp\left(H + Z\right)\right)\right] \leq \operatorname{tr}\left(\exp\left(H + \log \mathbb{E}\left[e^{Z}\right]\right)\right) \tag{98}$$

Proof. By Jensen's inequality and above two theorems, it concludes that

$$\mathbb{E}\left[\operatorname{tr}\left(\exp\left(H + \log e^{Z}\right)\right)\right] \leq \operatorname{tr}\left(\mathbb{E}\left[\exp\left(H + \log e^{Z}\right)\right]\right) \leq \operatorname{tr}\left(\mathbb{E}[e^{H}]\mathbb{E}[]\right)$$

□

Lemma 4.3.2 (Moment generating function)**.** *Let X be an $n \times n$ symmetric mean-zero random matrix such that $||X|| \leq K$ almost surely. Then*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \preceq \exp\left(g(\lambda)\mathbb{E}[X^2]\right) \quad \text{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3} \tag{99}$$

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

Proof. Similar as in Corollary 1.8.2, we have

$$e^{z} \leq 1 + z + \frac{1}{1 - |z|/3}\frac{z^2}{2}$$

As a result, we obtain $\exp\left(\lambda x\right) \leq 1 + \lambda x + g(\lambda)x^2$ for $|x| < K$. Then replace x by X, it concludes that $\exp\left(\lambda X\right) \preceq \boldsymbol{I} + \lambda X + g(\lambda)X^2$, which completes our proof. □

With the help of above results, we can deduce the following theorem now.

Theorem 4.3.3 (Matrix Bernstein inequality)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero $n \times n$ symmetric random matrices, such that $||X_i|| \leq K$ a.s. for all i, then for $t \geq 0$, it has*

$$\mathbb{P}\left(\Big|\Big|\sum_{i=1}^{N} X_i\Big|\Big| \geq t\right) \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right) \tag{100}$$

where $\sigma^2 := || \sum_{i=1}^{N} \mathbb{E}[X_i^2]||$ is the norm of the matrix variance of the sum.

Proof. First, denote $S := \sum_{i=1}^{N} X_i$ and $\lambda_{\max}(S) = \max_{1 \leq i \leq N} \lambda_i(S)$, then we have $||S|| = \max(\lambda_{\max}(S), \lambda_{\max}(-S))$, and for fixed $\lambda > 0$

$$\mathbb{P}(\lambda_{\max}(S) \geq t) = \mathbb{P}\left(\exp(\mu\lambda_{\max}(S)) \geq e^{\mu t}\right) \leq e^{-\mu t}\mathbb{E}\left[\exp(\mu\lambda_{\max}(S))\right]$$

In addition, $\mathbb{E}\left[\exp(\mu\lambda_{\max}(S))\right] = \mathbb{E}\left[\lambda_{\max}(\exp(\mu S))\right] \leq \mathbb{E}\left[\mathrm{tr}(e^{\mu S})\right]$.

Second, by Lemma 4.3.1, it shows that

$$\mathbb{E}\left[\mathrm{tr}\left(e^{\mu S}\right)\right] = \mathbb{E}\left[\mathrm{tr}\left(\exp\left(\mu\sum_{i=1}^{N-1} X_i + \mu X_N\right)\right)\right] \leq \mathbb{E}\left[\mathrm{tr}\left(\exp\left(\mu\sum_{i=1}^{N-1} X_i + \log\mathbb{E}\left[e^{\mu X_N}\right]\right)\right)\right]$$

And this process is also valid for X_{N-1}, hence we can finally obtain

$$\mathbb{E}\left[\mathrm{tr}\left(e^{\mu S}\right)\right] \leq \mathrm{tr}\left(\exp\left(\sum_{i=1}^{N}\log\mathbb{E}\left[e^{\mu X_i}\right]\right)\right)$$

Third, by Lemma 4.3.2, it deduces that

$$\mathbb{E}\left[e^{\mu X_i}\right] \preceq \exp\left(g(\mu)\mathbb{E}\left[X_i^2\right]\right)$$

Therefore, we have

$$\mathrm{tr}\left(\exp\left(\sum_{i=1}^{N}\log\mathbb{E}\left[e^{\mu X_i}\right]\right)\right) \leq \mathrm{tr}\left(\exp\left(g(\mu)\sum_{i=1}^{N}\mathbb{E}\left[X_i^2\right]\right)\right) \leq n\exp\left(g(\mu)\lambda_{\max}\left(\sum_{i=1}^{N}\mathbb{E}\left[X_i^2\right]\right)\right)$$

Finally, we get

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

Similarly, we can also get the same result for $\lambda_{\max}(-S)$, hence

$$\mathbb{P}(||S|| \geq t) \leq 2n\exp\left(-\mu t + g(\mu)\sigma^2\right)$$

Take $\mu = t/(\sigma^2 + Kt/3)$, then completes our proof □

Corollary 4.3.1 (Matrix Bernstein inequality: expectation)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero $n \times n$ symmetric random matrices such that $||X_i|| \leq K$ almost surely for all i. Then*

$$\mathbb{E}\left[\Big|\Big|\sum_{i=1}^{N} X_i\Big|\Big|\right] \leq \left|\left|\sum_{i=1}^{N}\mathbb{E}\left[X_i^2\right]\right|\right|^{1/2}\sqrt{\log n} + K\log n \tag{101}$$

Proof. Denote $S = \sum_{i=1}^{N} X_i$ and by Theorem 4.3.3, it concludes that

$$\mathbb{P}\left(||S|| \geq \sigma\sqrt{\log n + s} + K(\log n + s)\right) \leq 2n \exp\left(-\right)$$

□

Theorem 4.3.4 (Matrix Hoeffding inequality)**.** *Let $\epsilon_1, \cdots, \epsilon_n$ be independent symmetric Bernoulli random variables and let $A_1, \cdots, A_N$ be symmetric $n \times n$ matrices, then for any $t \geq 0$, it concludes that*

$$\mathbb{P}\left(\left|\left|\sum_{i=1}^{N} \epsilon_i A_i\right|\right| \geq t\right) \leq 2n \exp\left(-t^2/2\sigma^2\right) \tag{102}$$

where $\sigma^2 = ||\sum_{i=1}^{N} A_i^2||$.

Proof. First, for $\lambda \geq 0$,

$$\mathbb{E}\left[\exp\left(\lambda\epsilon_i A_i\right)\right] = \frac{1}{2}\left(\exp\left(\lambda A_i\right) + \exp\left(-\lambda A_i\right)\right) \preceq \exp\left(\lambda^2 A_i^2/2\right)$$

In fact, let $\Lambda_i = Q_i^T A_i Q_i$, where Q_i is an orthogonal matrix and Λ_i is diagonal whose components are the eigenvalues of A_i. Then

$$\frac{1}{2}Q_i^T\left(\exp\left(\lambda\Lambda_i\right) + \exp\left(-\lambda\Lambda_i\right)\right)Q_i \preceq Q_i^T \exp\left(\lambda^2\Lambda_i^2/2\right)Q_i$$

The last inequality due to $\cosh(\lambda) \leq \exp\left(\lambda^2/2\right)$. Now, denote $S = \sum_{i=1}^{N} \epsilon_i A_i$ and $\lambda_{\max}(S) = \max_{1\leq i\leq N} \lambda_i(S)$, then

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}(S) \geq t\right) &\leq e^{-\mu t}\mathbb{E}\left[\exp\left(\mu\lambda_{\max}(S)\right)\right] \leq e^{-\mu t}\mathbb{E}\left[\mathrm{tr}\left(e^{\mu S}\right)\right] \\
&\leq e^{-\mu t}\mathrm{tr}\left(\exp\left(\sum_{i=1}^{N} \log \mathbb{E}\left[e^{\mu\epsilon_i A_i}\right]\right)\right) \\
&\leq e^{-\mu t}\mathrm{tr}\left(\exp\left(\frac{\mu^2}{2}\sum_{i=1}^{N} A_i^2\right)\right) \\
&\leq n \exp\left(\frac{\sigma^2}{2}\mu^2 - \mu t\right)
\end{aligned}$$

Hence we obtain $\mathbb{P}\left(\lambda_{\max}(S) \geq t\right) \leq n \exp\left(-t/2\sigma^2\right)$ as $\mu = t/\sigma^2$, and we can get the same result for $\lambda_{\max}(-S)$, that completes our proof. □

Corollary 4.3.2 (Matrix Khintchine inequality)**.** *Let $\epsilon_1, \cdots, \epsilon_n$ be independent symmetric Bernoulli random variables and let $A_1, \cdots, A_N$ be symmetric $n \times n$ matrices, then it concludes that*

$$\mathbb{E}\left[\left|\left|\sum_{i=1}^{N} \epsilon_i A_i\right|\right|\right] \leq C\sigma\sqrt{\log n} \tag{103}$$

where $\sigma^2 = || \sum_{i=1}^{N} A_i^2||$. *More general, for* $p \geq 1$, *it has*

$$\mathbb{E}\left[\left|\left|\sum_{i=1}^{N} \epsilon_i A_i\right|\right|^p\right]^{1/p} \leq C\sigma\sqrt{p + \log n} \tag{104}$$

where $C > 0$ *is a constant.*

Proof. By Theorem 4.3.4, we have

$$\mathbb{P}\left(||S|| \geq \sigma\sqrt{2(\log n + s)}\right) \leq 2n \exp\left(-(\log n + s)\right) = 2e^{-s}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[||S||\right] &= \int_0^\infty \mathbb{P}\left(||S|| \geq t\right) dt \\
&\leq \sigma\sqrt{2\log n} + \int_0^\infty \mathbb{P}\left(||S|| \geq \sigma\sqrt{2(\log n + s)}\right) \frac{\sigma}{\sqrt{2(\log n + s)}} ds \\
&\leq \sigma\sqrt{2\log n} + \frac{\sigma}{\sqrt{2\log n}} \int_0^\infty 2e^{-s} ds = \sigma\sqrt{2\log n}\left(1 + \frac{1}{\log n}\right)
\end{aligned}$$

For $n \geq 2$, we can choose $C \geq 3\sqrt{2}$. In addition, we have

$$\begin{aligned}
\mathbb{E}\left[||S||^p\right] &= \int_0^\infty \mathbb{P}\left(||S||^p \geq t\right) dt = \int_0^\infty \mathbb{P}\left(||S|| \geq s\right) p s^{p-1} ds \\
&\leq \left(\sigma\sqrt{2\log n}\right)^p + p\sigma^p \int_0^\infty \mathbb{P}\left(||S|| \geq \sigma\sqrt{2(\log n + x)}\right) \left(\sqrt{2(\log n + x)}\right)^{p-2} dx \\
&\leq \left(\sigma\sqrt{2\log n}\right)^p + 2p\sigma^p \int_0^\infty e^{-x} \left(\sqrt{2(\log n + x)}\right)^{p-2} dx \\
&\leq \left(\sigma\sqrt{2\log n}\right)^p + np\sigma^p \int_0^\infty e^{-y/2} y^{p/2-1} dy = \left(\sigma\sqrt{2\log n}\right)^p + np\sigma^p \Gamma\left(p/2\right) \\
&\leq \left(\sigma\sqrt{2\log n}\right)^p + np\sigma^p \left(p/2\right)^{p/2} \leq C\sigma^p \left(\sqrt{p + \log n}\right)^p
\end{aligned}$$

where C is enough large such that the last inequality is valid. That completes our proof. □

Corollary 4.3.3 (Sharpness of matrix Bernstein inequality)**.** *Let* X *be an* $n \times n$ *random matrix that takes values* $e_k e_k^T$, $k = 1, \cdots, n$, *with probability* $1/n$ *each, where* (e_k) *denotes the standard basis in* $\mathbb{R}^n$. *Let* $X_1, \cdots, X_N$ *be independent copies of* X *and* $S := \sum_{i=1}^{N} X_i$

5 Quadratic Forms, Symmetrization, and Contraction

In this section, we will introduce some useful tools for high-dimensional probability and statistics.

5.1 Decoupling

Theorem 5.1.1 (Decoupling)**.** *Let A be an $n \times n$ diagonal-free matrix, i.e., the diagonal entries of A equal zero. Let $X = (X_1, \cdots, X_n)$ be a random vector with independent mean-zero coordinates X_i . Then, for every convex function $F : \mathbb{R} \to \mathbb{R}$, one has*

$$\mathbb{E}\left[F(X^TAX)\right] \le \mathbb{E}\left[F(4X^TAX')\right] \tag{105}$$

where X' is an independent copy of X.

In order to prove above theorem, the following lemma is necessary.

Lemma 5.1.1. *Let Y and Z be independent random variables such that $\mathbb{E}[Z] = 0$. Then, for every convex function F, one has*

$$\mathbb{E}\left[F(Y)\right] \le \mathbb{E}\left[F(Y+Z)\right] \tag{106}$$

Proof. For fixed $y \in \mathbb{R}$, by Jensen inequality, it concludes that

$$F(y) = F(y + \mathbb{E}[Z]) = F(\mathbb{E}\left[y+Z\right]) \le \mathbb{E}\left[F(y+Z)\right]$$

As a result,

$$\begin{aligned}\mathbb{E}\left[F(Y+Z)\right] &= \int_{\mathbb{R}^2} F(y+z)p_{Y,Z}(y,z)dzdy = \int_{\mathbb{R}^2} F(y+z)p_Y(y)p_Z(z)dzdy \\ &= \int_{\mathbb{R}} \mathbb{E}\left[F(y+Z)\right]p_Y(y)dy \ge \int_{\mathbb{R}} F(y)p_Y(y)dy = \mathbb{E}\left[F(Y)\right]\end{aligned}$$

where $p_{Y,Z}(y,z)$ is the joint density of Y, Z and the second equality is valid since Y and Z are independent. □

Now back to the proof of Theorem 5.1.1.

Proof. Denote $\delta_1, \cdots, \delta_n$ be n independent Bernoulli random variables and $\mathbb{E}\left[\delta_i(1-\delta_j)\right] = 1/4$ for $i \neq j$, hence

$$X^TAX = \sum_{i \neq j} a_{i,j}X_iX_j = 4\sum_{i \neq j} \mathbb{E}_\delta\left[\delta_i(1-\delta_j)\right]a_{i,j}X_iX_j = 4\mathbb{E}_I\left[\sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j\right]$$

where $I := \{i : \delta_i = 1\}$. In addition, by Jensen inequality and Fubini theorem, we have

$$\mathbb{E}_X\left[F(X^TAX)\right] = \mathbb{E}_X\left[F\left(4\mathbb{E}_I\left[\sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j\right]\right)\right] \le \mathbb{E}_I\left[\mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j\right)\right]\right]$$

Then there exists a realization of I such that

$$\mathbb{E}_X\left[F(X^TAX)\right] \le \mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j\right)\right] = \mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j'\right)\right]$$

The last inequality is valid since X_i and X_j are independent. As a result, we can obtain

$$\mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j'\right)\right] \le \mathbb{E}_X\left[F\left(4\sum_{i,j=1}^n a_{i,j}X_iX_j'\right)\right]$$

In fact, denote $Y := \sum_{(i,j)\in I\times I^c} a_{i,j}X_iX_j'$, $Z_1 := \sum_{(i,j)\in I\times I} a_{i,j}X_iX_j'$, $Z_2 := \sum_{(i,j)\in I^c\times[n]} a_{i,j}X_iX_j'$, the above inequality can be rewritten as

$$\mathbb{E}_X\left[F(4Y)\right] \le \mathbb{E}_X\left[F(4(Y+Z_1+Z_2))\right]$$

It is easy to see Z_1, Z_2 are mean-zero, hence the above inequality can be derived from Lemma 5.1.1. □

5.2 Hanson–Wright Inequality

In this subsection, we will focus on a more general concentration inequality for a chaos, as preliminary, we need some lemmas.

Lemma 5.2.1 (MGF of Gaussian chaos)**.** *Let $X, X' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ are independent and A be a $n\times n$ symmetric matrix, then*

$$\mathbb{E}\left[\exp\left(\lambda X^TAX\right)\right] \le \exp\left(C\lambda^2||A||_F^2\right) \tag{107}$$

for all $|\lambda| \le c/||A||$, where $c, C > 0$ are constant.

Proof. Consider the singular value decomposition of A by $A = \sum_{i=1}^n s_iu_iv_i^T$, where s_i are the singular value of A, hence we obtain

$$X^TAX' = \sum_{i=1}^n s_i\langle X, u_i\rangle\langle X', v_i\rangle$$

Next, denote $g := (\langle X, u_1\rangle, \cdots, \langle X, u_n\rangle)^T$ as well as g', let U, V be the matrices with u_i^T, v_i^T as their rows, then $g \sim \mathcal{N}(\mathbf{0}, U^TU)$ and $g' \sim \mathcal{N}(\mathbf{0}, V^TV)$, hence

$$\mathbb{E}\left[\exp\left(\lambda X^TAX'\right)\right] = \mathbb{E}\left[\prod_{i=1}^n \exp\left(\lambda s_ig_ig_i'\right)\right]$$

□

Lemma 5.2.2 (Comparison)**.** *Consider independent mean-zero sub-Gaussian random vectors* X, X *in* $\mathbb{R}^n$ *such that* $||X||_{\psi_2}, ||X'||_{\psi_2} \leq K$. *Consider also independent random vectors* $g, g' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$*. Let* A *be an* $n \times n$ *matrix. Then*

$$\mathbb{E}\left[\exp\left(\lambda X^T A X'\right)\right] \leq \mathbb{E}\left[\exp\left(CK^2 \lambda g^T A g'\right)\right] \tag{108}$$

For any $\lambda \in \mathbb{R}$.

Proof. Since $X^T A X'$ can be rewritten as $\langle X, AX' \rangle$, conditional expectation $\mathbb{E}_X\left[X^T A X'\right]$ on X is also sub-Gaussian whose sub-Gaussian norm is bounded by $K||AX'||_2$, in other words, by Proposition 1.4.1, it concludes that

$$\mathbb{E}_X\left[\exp\left(\lambda X^T A X'\right)\right] \leq \exp\left(C\lambda^2 K^2 ||AX'||_2^2\right) \quad \text{for } \lambda \in \mathbb{R}$$

Similarly, we also can obtain

$$\mathbb{E}_g\left[\exp\left(\mu g^T A X'\right)\right] = \exp\left(\mu^2 ||AX'||_2^2/2\right) \quad \text{for } \lambda \in \mathbb{R}$$

Hence we conclude that

$$\mathbb{E}_X\left[\exp\left(\lambda X^T A X'\right)\right] \leq \mathbb{E}_g\left[\exp\left(\sqrt{2C} K \lambda g^T A X'\right)\right]$$

In addition, we can conclude that following result by the same argument.

$$\mathbb{E}_{X'}\left[\exp\left(\mu^2 ||AX'||_2^2/2\right)\right] \leq \exp\left(\mu^2 K^2 ||A||_F^2\right)$$

and

$$\mathbb{E}_{g'}\left[\mathbb{E}_g\left[\exp\left(\tau g^T A g'\right)\right]\right] = \mathbb{E}_{g'}\left[\exp\left(\tau^2 ||Ag'||_2^2/2\right)\right] = \exp\left(\tau^2 ||A||_F^2/2\right)$$

This completes our proof. □

Theorem 5.2.1 (Hanson–Wright inequality)**.** *Let* $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ *be a random vector with independent mean-zero sub-Gaussian coordinates. Let* A *be an* $n \times n$ *matrix. Then, for every* $t \geq 0$*, we have*

$$\mathbb{P}\left(\left|X^T A X - \mathbb{E}\left[X^T A X\right]\right| \geq t\right) \leq 2\exp\left(-c \min\left\{\frac{t^2}{K^4 ||A||_F^2}, \frac{t}{K^2 ||A||}\right\}\right) \tag{109}$$

where $K := \max_{1 \leq i \leq n} ||X_i||_{\psi_2}$.

Proof. Without loss of generality, assume $K = 1$, denote

$$p_t := \mathbb{P}\left(X^T A X - \mathbb{E}\left[X^T A X\right] \geq t\right)$$

and

$$X^T A X - \mathbb{E}\left[X^T A X\right] = \sum_{i=1}^n a_{i,i}\left(X_i^2 - \mathbb{E}[X_i^2]\right) + \sum_{i \neq j} a_{i,j} X_i X_j := I_1 + I_2$$

Hence, we obtain

$$p_t \leq \mathbb{P}\left(I_1 \geq t/2\right) + \mathbb{P}\left(I_2 \geq t/2\right) := p_t^{(1)} + p_t^{(2)}$$

By centering property of sub-exponential norm and $X_i^2 - \mathbb{E}[X_i^2]$ be mean-zero and sub-exponential, it has

$$||X_i^2 - \mathbb{E}\left[X_i^2\right]||_{\psi_1} \leq ||X_i^2||_{\psi_1} = ||X_i||_{\psi_2}^2 \leq 1$$

By Theorem 1.8.1, it concludes that

$$p_t^{(1)} \leq \exp\left(-c \min\left\{\frac{t^2}{\sum_i a_{i,i}^2}, \frac{t}{\max_i a_{i,i}}\right\}\right) \leq \exp\left(-c' \min\left\{\frac{t^2}{||A||_F}, \frac{t}{||A||}\right\}\right)$$

In addition, by Chebyshev inequality, it concludes that

$$p_t^{(2)} \leq \exp\left(-\lambda t\right) \mathbb{E}\left[\exp\left(\lambda I_2\right)\right]$$

According to Lemma 5.2.2 and Theorem 5.1.1, we have

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda I_2\right)\right] &\leq \mathbb{E}\left[\exp\left(4 \sum_{i \neq j} a_{i,j} X_i X_j'\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(C_1 \lambda \sum_{i \neq j} a_{i,j} g_i g_j'\right)\right] \\ &\leq \exp\left(C \lambda^2 ||A||_F^2\right) \end{aligned}$$

As a result,

$$p_t^{(2)} \leq \exp\left(-c'' \min\left\{\frac{t^2}{||A||_F}, \frac{t}{||A||}\right\}\right)$$

This completes our proof. □

Corollary 5.2.1. *Consider a mean-zero sub-Gaussian random vector X in $\mathbb{R}^n$ with $||X||_{\psi_2} \leq K$. Let B be an $m \times n$ matrix, then*

$$\mathbb{E}\left[\exp\left(\lambda^2 ||BX||_2^2\right)\right] \leq \exp\left(CK^2\lambda^2||B||_F^2\right) \quad \text{for} \quad \lambda \leq \frac{c}{||B||} \tag{110}$$

Proof. By Theorem 5.1.1, it concludes that

$$\mathbb{E}\left[\exp\left(\lambda^2||BX||_2^2\right)\right] = \mathbb{E}\left[\exp\left(\lambda^2 X^T B^T B X\right)\right] \leq \mathbb{E}\left[\exp\left(4\lambda^2 X^T B^T B X'\right)\right]$$

where X' is an independent copy of X. Next denote $g, g' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and according to Lemma 5.2.2, it deduces that

$$\mathbb{E}\left[\exp\left(4\lambda^2 X^T B^T B X'\right)\right] \leq \mathbb{E}\left[\exp\left(C_1 K^2 \lambda^2 g^T B^T B g'\right)\right]$$

Denote $B^T B = \{b_{i,j}\}_{i,j=1}^n$, which is a symmetric matrix, and $g^T B^T B g' = \sum_{i,j=1}^n b_{i,j} g_i g_j'$, hence

$$\mathbb{E}\left[\exp\left(C_1 K^2 \lambda^2 g^T B^T B g'\right)\right] = \mathbb{E}\left[\exp\left(C_1 K^2 \lambda^2 \sum_{i,j=1}^n b_{i,j} g_i g_j'\right)\right] = \prod_{i,j=1}^n \mathbb{E}\left[\exp\left(C_1 K^2 \lambda^2 b_{i,j} g_i g_j'\right)\right]$$

□

5.3 Concentration for Anisotropic Random Vectors

In this subsection, we will focus on the concentration property of anisotropic random vectors, i.e. for B is a fixed $m \times n$ matrix and X is an isotropic n-dimensional vector, consider the behave of BX.

Lemma 5.3.1. *It can conclude that* $\mathbb{E}\left[||BX||_2^2\right] = ||B||_F^2$.

Proof. Since

$$\mathbb{E}\left[||BX||_2^2\right] = \mathbb{E}\left[\sum_{j=1}^{m}\left(\sum_{i=1}^{n} b_{j,i}X_i\right)^2\right] = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{j,i}^2 = ||B||_F^2$$

This completes our proof. □

Lemma 5.3.2. *Let D be $k \times m$ and B be $m \times n$ matrices, then* $||DB||_F \leq \|D\|\|B\|_F$.

Proof.

$$||DB||_F = \max_{||x||_2=1} ||DBx||_2 \leq ||D|| \max_{||x||_2=1} ||Bx||_2 = \|D\|\|B\|_F$$

□

Theorem 5.3.1 (Concentration for random vectors). *Let B be an $m \times n$ matrix, and let $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero unit-variance sub-Gaussian coordinates. Then*

$$\Big|\Big| ||BX||_2 - ||B||_F \Big|\Big|_{\psi_2} \leq CK^2||B|| \tag{111}$$

where $K := \max_{1\leq i\leq n} ||X_i||_{\psi_2}$.

Proof. Denote $A := B^TB$ and assume $K \geq 1$, then by above lemma

$$||A||_F \leq \|B^T\|\|B\|_F = \|B\|\|B\|_F$$

Next according to Theorem 5.2.1, it concludes that

$$\mathbb{P}\left(\big|||BX||_2 - ||B||_F\big| \geq u\right) \leq 2\exp\left(-\frac{c}{K^4}\min\left\{\frac{u^2}{||B||^2||B||_F^2}, \frac{u}{||B||^2}\right\}\right)$$

Let $u = \epsilon\|B\|_F^2$ for $\epsilon \geq 0$, we obtain

$$\mathbb{P}\left(\big|||BX||_2^2 - ||B||_F^2\big| \geq \epsilon\|B\|_F^2\right) \leq 2\exp\left(-c\min\left\{\epsilon^2, \epsilon\right\}\frac{\|B\|_F^2}{K^4\|B\|^2}\right)$$

In addition, if $\big|||BX||_2 - ||B||_F\big| \geq \delta||B||_F$, then $\big|||BX||_2^2 - ||B||_F^2\big| \geq \epsilon\|B\|_F^2$. In fact, denote $\delta^2 = \min\left\{\epsilon^2, \epsilon\right\}$, i.e. $\epsilon = \max\left\{\delta, \delta^2\right\}$, and if $\|BX\|_2 \geq \|B\|_F$, then $\|BX\|_2^2 \geq (1+\delta)^2\|B\|_F^2$ and $(1+\delta)^2 \geq 1 + \delta^2 \geq 1 + \epsilon$, hence

$$\|BX\|_2^2 \geq (1+\epsilon)\|B\|_F^2$$

On the other hand, if $\|BX\|_2 < \|B\|_F$, then $\|BX\|_2^2 < (1-\delta)^2\|B\|_F^2$ and $(1-\delta)^2 \leq 1-\epsilon$, hence

$$\|BX\|_2^2 < (1-\epsilon)\|B\|_F^2$$

As a result, we obtain

$$\mathbb{P}\left(\left|||BX||_2 - ||B||_F\right| \geq \delta\|B\|_F\right) \leq 2\exp\left(-c\delta^2\frac{\|B\|_F^2}{K^4\|B\|^2}\right)$$

Let $t = \delta\|B\|_F$ and hence conclude

$$\mathbb{P}\left(\left|||BX||_2 - ||B||_F\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^4\|B\|^2}\right)$$

So $\|BX\|_2 - \|B\|_F$ is sub-Gaussian with norm of $CK^4\|B\|^2$. □

Corollary 5.3.1 (Distance to a subspace)**.** *Let E be a subspace of $\mathbb{R}^n$ of dimension d. Consider a random vector $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ with independent mean-zero unit-variance sub-Gaussian coordinates. Then*

$$\mathbb{E}\left[\text{dist}\left(X, E\right)^2\right] = n - d \tag{112}$$

and for any $t \geq 0$

$$\mathbb{P}\left(\left|d(X,E) - \sqrt{n-d}\right| > t\right) \leq 2\exp\left(-ct^2/K^4\right) \tag{113}$$

where $K := \max_{1\leq i\leq n}||X_i||_{\psi_2}$.

Proof. Assume that the basis of E is $\{e_i\}_{i=1}^d$, where e_i be the standard coordinates in $\mathbb{R}^n$. In fact, if not, we can find an orthogonal matrix A such that AE is such subspace, besides, AX still has independent sub-Gaussian coordinates with zero-mean and unit-variance. Hence, $\text{dist}(X,E)^2 = \sum_{i=1}^d X_{n+1-i}^2$, therefore we conclude that $\mathbb{E}\left[\text{dist}\left(X,E\right)^2\right] = n - d$. Next, denote matrix B by

$$\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{n-d} \end{pmatrix}$$

Then $\text{dist}(X,E) = \|BX\|_2$ and $\|B\|_F = n - d$, by Theorem 5.3.1, it concludes that

$$\mathbb{P}\left(\left|d(X,E) - \sqrt{n-d}\right| > t\right) \leq 2\exp\left(-ct^2/K^4\right)$$

This completes our proof. □

Corollary 5.3.2 (Tails of sub-Gaussian random vectors)**.** *Let B be an $m \times n$ matrix, and let X be a mean-zero sub-Gaussian random vector in $\mathbb{R}^n$ with $||X||_{\psi_2} \leq K$. Then for any $t \geq 0$, we have*

$$\mathbb{P}\left(\|BX\|_2 \geq CK\|B\|_F + t\right) \leq \exp\left(-\frac{ct^2}{K^2\|B\|^2}\right) \tag{114}$$

Remark 5.3.1. *In fact, the independence of coordinates are essential, since we can give the following example. That is, there exists a mean-zero isotropic sub-Gaussian random vector $X \in \mathbb{R}^n$ such that*

$$\mathbb{P}\left(\|X\|_2 = 0\right) = \mathbb{P}\left(\|X\|_2 \geq 1.4\sqrt{n}\right) = \frac{1}{2}$$

In other words, Theorem 5.3.1 is actually weaker than Theorem 2.1.1.

5.4 Symmetrization

In this section we develop the simple and useful technique of symmetrization. It allows one to reduce arbitrary distributions to symmetric distributions and in some cases even to the symmetric Bernoulli distribution.

Proposition 5.4.1 (Constructing symmetric distributions). *Let X be a random variable and ξ be an independent symmetric Bernoulli random variable.*

(a) ξX and $\xi|X|$ are symmetric random variables and that they have the same distribution.

(b) If X is symmetric, show that the distributions of ξX and $\xi|X|$ are the same as that of X.

(c) Let X' be an independent copy of X, then $X - X'$ is symmetric.

Proof. For (a),

$$\mathbb{P}\left(\xi X \geq t\right) = \frac{1}{2}\left(\mathbb{P}\left(X \geq t|\xi = 1\right) + \mathbb{P}\left(X \leq -t|\xi = -1\right)\right) = \frac{1}{2}\left(\mathbb{P}\left(X \geq t\right) + \mathbb{P}\left(X \leq -t\right)\right)$$
$$\mathbb{P}\left(-\xi X \geq t\right) = \frac{1}{2}\left(\mathbb{P}\left(X \leq -t|\xi = 1\right) + \mathbb{P}\left(X \geq t|\xi = 1\right)\right) = \frac{1}{2}\left(\mathbb{P}\left(X \geq t\right) + \mathbb{P}\left(X \leq -t\right)\right)$$

For (b), due to symmetric property

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

For (c), denote $p_X(\cdot)$ be the density of X

$$\mathbb{P}\left(X - X' \geq t\right) = \int_{y-z\geq t} (y-z)p_X(y)p_X(z)dydz$$

This completes our proof. □

Throughout this section, we denote by

$$\epsilon_1, \epsilon_2, \epsilon_3 \cdots$$

a sequence of independent symmetric Bernoulli random variables. We assume that they are (jointly) independent not only of each other but also of any other random variables in question.

Lemma 5.4.1 (Symmetrization). *Let $X_1, \cdots, X_N$ be independent mean-zero random vector in a normed space, then*

$$\frac{1}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N} X_i\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right] \tag{115}$$

Proof. By Lemma 5.1.1, since the norm is convex, it concludes that

$$p := \mathbb{E}\left[\left\|\sum_{i=1}^{N} X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(X_i - X_i'\right)\right\|\right]$$

where X_i' be the independent copy of X_i. Since $X_i - X_i'$ are symmetric and they have the same distribution as $\epsilon_i\left(X_i - X_i'\right)$, it concludes that

$$p \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i\left(X_i - X_i'\right)\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right] + \mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i'\right\|\right] = 2\mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right]$$

For the lower bound, we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i\left(X_i - X_i'\right)\right\|\right] = \mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(X_i - X_i'\right)\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^{N} X_i\right\|\right]$$

This completes our proof. □

Lemma 5.4.2 (Removing the mean-zero assumption)**.** *If remove the mean-zero condition in Lemma 5.4.1, it concludes that*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N} X_i - \sum_{i=1}^{N}\mathbb{E}\left[X_i\right]\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right] \tag{116}$$

Proof. Similar as Lemma 5.4.1, it has

$$\mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(X_i - \mathbb{E}\left[X_i\right]\right)\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(X_i - X_i'\right)\right\|\right]$$

where X_i' is the independent copy of X_i and $X_i' - \mathbb{E}[X_i']$ is mean-zero, besides, $\mathbb{E}[X_i] = \mathbb{E}[X_i']$. Then this completes our proof. □

Remark 5.4.1. *For Lemma 5.4.1, we can replace the norm by any increasing, convex, function $F : \mathbb{R}^+ \to \mathbb{R}$, i.e. replace $\|\cdot\|$ by $F\left(\|\cdot\|\right)$, namely*

$$\mathbb{E}\left[F\left(\frac{1}{2}\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right)\right] \leq \mathbb{E}\left[F\left(\left\|\sum_{i=1}^{N} X_i\right\|\right)\right] \leq \mathbb{E}\left[F\left(2\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|\right)\right] \tag{117}$$

Remark 5.4.2. *Let $X_1, \cdots, X_N$ be independent random variables, then $\sum_{i=1}^{N} X_i$ is sub-Gaussian if and only if $\sum_{i=1}^{N}\epsilon_i X_i$ is sub-Gaussian and*

$$c\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|_{\psi_2} \leq \left\|\sum_{i=1}^{N} X_i\right\|_{\psi_2} \leq C\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|_{\psi_2} \tag{118}$$

Proof. First, since $F(x) := \exp\left(\lambda^2 x^2\right)$ is convex, by above remark, it concludes that

$$\mathbb{E}\left[\exp\left(\frac{\lambda^2}{4}\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|^2\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda^2\left\|\sum_{i=1}^{N} X_i\right\|\right)\right] \leq \mathbb{E}\left[\exp\left(4\lambda^2\left\|\sum_{i=1}^{N}\epsilon_i X_i\right\|^2\right)\right]$$

Then by Proposition 1.4.1, we complete this proof. □

5.5 Random Matrices With Non-I.I.D. Entries

As we have seen in last subsection, the general usage of symmetrization technique has two steps, first, replace X_i by $\epsilon_i X_i$, where ϵ_i is a symmetric Bernoulli distribution, next conditional on X_i, we only leave the randomness of ϵ_i, which is simpler to deal with. As a application this technique, we will give a general bound for random matrices with non-i.d.d entries.

Theorem 5.5.1 (Norms of random matrices with non-i.i.d. entries)**.** *Let A be an $n \times n$ symmetric random matrix whose entries on and above the diagonal are independent mean-zero random variables. Then*

$$\mathbb{E}\left[\|A\|\right] \leq C\sqrt{\log n}\,\mathbb{E}\left[\max_i \|A_i\|_2\right] \tag{119}$$

where A_i be the row of A.

Proof. Denote $\{e_i\}_{i=1}^n$ be the standard basis of $\mathbb{R}^n$ and define

$$Z_{i,j} = \begin{cases} A_{i,j} e_i e_i^T & \text{if } i = j \\ A_{i,j}(e_i e_j^T + e_j e_i^T) & \text{if } i < j \end{cases}$$

Hence $A = \sum_{i \leq j} Z_{i,j}$, then by Lemma 5.4.1 and Corollary 4.3.2 with conditional on $Z_{i,j}$, it concludes that

$$\mathbb{E}\left[\|A\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i \leq j} \epsilon_{i,j} Z_{i,j}\right\|\right] \leq C\sqrt{\log n}\,\mathbb{E}\left[\left\|\sum_{i \leq j} Z_{i,j}^2\right\|\right]^{1/2}$$

In addition, $\sum_{i \leq j} Z_{i,j}^2 = \sum_{i=1}^n \left(\sum_{j=1}^n A_{i,j}^2\right) e_i e_i^T = \sum_{i=1}^n \|A_i\|_2^2 e_i e_i^T$. As a result, we obtain $\left\|\sum_{i \leq j} Z_{i,j}^2\right\| \leq \max_i \|A_i\|_2^2$ □

5.6 Contraction Principle

Theorem 5.6.1 (Contraction Principle)**.** *Let $x_1, \cdots, x_N$ be (deterministic) vectors in some normed space, and let $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^N a_i \epsilon_i x_i\right\|\right] \leq \|a\|_\infty \mathbb{E}\left[\left\|\sum_{i=1}^N \epsilon_i x_i\right\|\right] \tag{120}$$

Proof. Without loss of generality, assume $\|a\|_\infty \leq 1$. Define

$$f(a) = \mathbb{E}\left[\left\|\sum_{i=1}^N a_i \epsilon_i x_i\right\|\right]$$

It is easy to see $f : \mathbb{R}^N \to \mathbb{R}$ is convex. Since $a \in [-1, 1]^N$ and the maximum of convex functions attains on the boundary of its domain, that is, $a_i \pm 1$. In addition, $-\epsilon_i$ are also Bernoulli distributions, this completes our proof. □

Corollary 5.6.1. *Let $X_1, \cdots, X_N$ be independent mean-zero random vectors in a normed space, and let $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^N a_i X_i\right\|\right] \leq 4\|a\|_\infty \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \tag{121}$$

Proof. By Lemma 5.4.1 for the first and last inequalities, Theorem 5.6.1 for the second, we conclude that

$$\mathbb{E}\left[\left\|\sum_{i=1}^N a_i X_i\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^N a_i \epsilon_i X_i\right\|\right] \leq 2\|a\|_\infty \mathbb{E}\left[\left\|\sum_{i=1}^N \epsilon_i X_i\right\|\right] \leq 4\|a\|_\infty \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right]$$

This completes our proof. □

Lemma 5.6.1 (Symmetrization with Gaussian)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero random vectors in a normed space. Let $g_1, \cdots, g_N \sim \mathcal{N}(0,1)$ be independent Gaussian random variables which are also independent of X_i. Then*

$$\frac{c}{\sqrt{\log N}}\mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq 3\mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right] \tag{122}$$

Proof. For the upper bound, by Lemma 5.4.1

$$\mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^N \epsilon_i X_i\right\|\right] = \sqrt{2\pi}\mathbb{E}_X\left[\left\|\sum_{i=1}^N \mathbb{E}_g\left[|g_i|\right] \epsilon_i X_i\right\|\right]$$

Then by Jensen's inequality, it concludes that

$$\mathbb{E}_X\left[\left\|\sum_{i=1}^N \mathbb{E}_g\left[|g_i|\right] \epsilon_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^N |g_i| \epsilon_i X_i\right\|\right] = \mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right]$$

For the lower bound, since g_i and $\epsilon_i g_i$ has the same distributions, it concludes that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^N \epsilon_i g_i X_i\right\|\right] \\
&\leq \mathbb{E}_X\left[\mathbb{E}_g\left[\|g\|_\infty \mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^N \epsilon_i X_i\right\|\right]\right]\right] \\
&= \mathbb{E}_g\left[\|g\|_\infty \mathbb{E}_\epsilon\left[\mathbb{E}_X\left[\left\|\sum_{i=1}^N \epsilon_i X_i\right\|\right]\right]\right] \\
&\leq 2\left(\mathbb{E}\left[\|g\|_\infty\right]\right)\left(\mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right]\right)
\end{aligned}$$

By Proposition 1.4.2, it concludes that $\mathbb{E}\left[\|g\|_\infty\right] \leq C\sqrt{\log N}$, which completes our proof. □

6 Random Processes

In this section, we will focus on the concentration property of random process.

6.1 Preliminary for random process

Definition 6.1.1 (Gaussian process)**.** *Suppose $X = (X_t)_{t\in[0,T]}$ is a random process such that for any $0 \le t_1 \le \cdots \le t_n \le T$, $(X_{t_1}, \cdots, X_{t_n})$ is a Gaussian vector, i.e. for any $a_1 \cdots, a_n \in \mathbb{R}$, $\sum_{i=1}^n a_i X_{t_i}$ is a normal random variable, then X is called the Gaussian process.*

An important example of Gaussian process is the standard Brownian motion $B = (B_t)_{t\ge 0}$ such that $\mathbb{E}[B_t] = 0$ and $\mathrm{Var}(B_t) = t$. Moreover, suppose $g \in \mathbb{R}^n$ is a standard normal random vector, we can define the canonical Gaussian process of g by

$$X_t := \langle g, t\rangle$$

where $t \in T \subset \mathbb{R}^n$. As a result,

Lemma 6.1.1 (Gaussian random vectors)**.** *Let Y be a mean-zero Gaussian random vector in $\mathbb{R}^n$. Then there exist points $t_1, \cdots, t_n \in \mathbb{R}^n$ such that*

$$Y \equiv (\langle g, t_i\rangle)_{i=1}^n \tag{123}$$

where $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, and the equivalence symbol means that the distributions of the two random vectors are the same.

Proof. Denote Σ be the covariance of Y and hence $Y \equiv \Sigma^{-1/2} g$, where t_i are denoted by the i-th row of $\Sigma^{-1/2}$, so the coordinate of Y is $(\langle g, t_i\rangle)$. □

6.2 Slepian's Inequality

In general, it is useful to consider a uniformly control for random process $(X_t)_{t\in[0,T]}$, that is, give a bound for

$$\mathbb{E}\left[\sup_{t\in[0,T]} X_t\right]$$

And for Brownian motion, we already have a good result called 'Reflection principle', which shows that

$$\mathbb{E}\left[\sup_{t\in[0,T]} B_t\right] = \sqrt{\frac{2T}{\pi}}$$

But for general random process, the similar results are not so trivial, however, Slepian's inequality gives another comparison result for Gaussian process, before introducing it, we need some preliminaries.

One basic technique is called *Gaussian interpolation*, in detail, suppose $T \subset \mathbb{R}^n$ is finite and $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ are two independent Gaussian random vectors, we can define

$$Z(u) := \sqrt{u}X_t + \sqrt{1-u}Y_t \quad \text{for } u \in [0,1]$$

which is also Gaussian. In fact, we have the following result.

$$\Sigma(Z(u)) = u\Sigma(X_t) + (1-u)\Sigma(Y_t)$$

Lemma 6.2.1 (Gaussian integration by parts)**.** *Let $X \sim \mathcal{N}(0,1)$. Then for any differentiable function $f : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}\left[f'\left(X\right)\right] = \mathbb{E}\left[Xf\left(X\right)\right] \tag{124}$$

Proof. Since

$$\mathbb{E}\left[f'\left(X\right)\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(x)e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xf(x)e^{-x^2/2}dx = \mathbb{E}\left[Xf\left(X\right)\right]$$

This completes our proof. □

In general, we have

$$\sigma^2\mathbb{E}\left[f'\left(X\right)\right] = \mathbb{E}\left[Xf\left(X\right)\right] \tag{125}$$

where $X \sim \mathcal{N}(0,\sigma^2)$, in addition, we have

Lemma 6.2.2 (Multivariate Gaussian integration by parts)**.** *Let $X \sim \mathcal{N}(0,\Sigma)$. Then for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ we have*

$$\Sigma\mathbb{E}\left[\bigtriangledown f\left(X\right)\right] = \mathbb{E}\left[Xf\left(X\right)\right] \tag{126}$$

Lemma 6.2.3 (Gaussian interpolation)**.** *Consider two independent Gaussian random vectors $X \sim \mathcal{N}(0,\Sigma^X)$ and $Y \sim \mathcal{N}(0,\Sigma^Y)$. Define the interpolation Gaussian vector*

$$Z(u) := \sqrt{u}X_t + \sqrt{1-u}Y_t \quad \text{for } u \in [0,1] \tag{127}$$

Then for any twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\frac{d}{du}\mathbb{E}\left[f\left(Z(u)\right)\right] = \frac{1}{2}\sum_{i,j=1}^{n}\left(\Sigma_{i,j}^X - \Sigma_{i,j}^Y\right)\mathbb{E}\left[\frac{\partial^2}{\partial x_i\partial x_j}f\left(Z(u)\right)\right] \tag{128}$$

Proof. First, it has

$$\frac{d}{du}\mathbb{E}\left[f\left(Z(u)\right)\right] = \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial}{\partial x_i}f\left(Z(u)\right)\frac{d}{du}Z_i(u)\right] = \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial}{\partial x_i}f\left(Z(u)\right)\left(\frac{X_t^{(i)}}{\sqrt{u}} - \frac{Y_t^{(i)}}{\sqrt{1-u}}\right)\right]$$

By Lemma 6.2.2, it concludes that

$$\mathbb{E}\left[\frac{\partial}{\partial x_i}f\left(\sqrt{u}X_t + \sqrt{1-u}Y_t\right)X_t^{(i)}\right] = \sum_{j=1}^{n}\Sigma_{i,j}^X\mathbb{E}\left[\sqrt{u}\bigtriangledown\frac{\partial^2}{\partial x_i\partial x_j}f(Z(u))\right]$$

Hence

$$\frac{d}{du}\mathbb{E}\left[f\left(Z(u)\right)\right] = \frac{1}{2}\sum_{i,j=1}^{n}\left(\Sigma_{i,j}^X - \Sigma_{i,j}^Y\right)\mathbb{E}\left[\frac{\partial^2}{\partial x_i\partial x_j}f\left(Z(u)\right)\right]$$

This completes our proof. □

Lemma 6.2.4 (Slepian's inequality, functional form)**.** *Consider two mean-zero Gaussian random vectors X and Y in $\mathbb{R}^n$. Assume that for all $i, j = 1, \cdots, n$, we have*

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[Y_i^2\right] \quad \text{and} \quad \mathbb{E}\left[(X_i - X_j)^2\right] \leq \mathbb{E}\left[(Y_i - Y_j)^2\right] \tag{129}$$

Consider a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\partial^2}{\partial x_i \partial x_j} f \geq 0 \quad \text{for } i \neq j$$

Then

$$\mathbb{E}\left[f(X)\right] \geq \mathbb{E}\left[f(Y)\right]$$

Proof. Define $Z(u)$ as in Lemma 6.2.3, since

$$\frac{d}{du}\mathbb{E}\left[f\left(Z(u)\right)\right] = \frac{1}{2}\sum_{i,j=1}^{n}\left(\Sigma_{i,j}^X - \Sigma_{i,j}^Y\right)\mathbb{E}\left[\frac{\partial^2}{\partial x_i \partial x_j} f\left(Z(u)\right)\right] \geq 0$$

The last inequality is deduced due to

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[Y_i^2\right] \quad \text{and} \quad \mathbb{E}\left[(X_i - X_j)^2\right] \leq \mathbb{E}\left[(Y_i - Y_j)^2\right]$$

which means $\Sigma_{i,j}^Y \geq \Sigma_{i,j}^Y$, hence this completes our proof. □

Theorem 6.2.1 (Slepian's inequality)**.** *Let X and Y be Gaussian random vectors as in Lemma 6.2.4. Then for every $\tau \in \mathbb{R}$ we have*

$$\mathbb{P}\left(\max_{i \leq n} X_i \geq \tau\right) \leq \mathbb{P}\left(\max_{i \leq n} Y_i \geq \tau\right) \tag{130}$$

Consequently,

$$\mathbb{E}\left[\max_{i \leq n} X_i \geq \tau\right] \leq \mathbb{E}\left[\max_{i \leq n} Y_i \geq \tau\right] \tag{131}$$

Proof. Let $\{h_n(x)\}_{n=1}^{\infty}$ be a sequence of smooth approximation of indicator function $1_{x \leq \tau}$, and define for $x \in \mathbb{R}^n$

$$f_n(x) := \prod_{i=1}^{n} h_i(x_i)$$

By Lemma 6.2.4, we have $\mathbb{E}\left[f_n(X)\right] \leq \mathbb{E}\left[f_n(Y)\right]$, where $\mathbb{E}\left[f(X)\right] = \mathbb{P}\left(\max_{i \leq n} X_i \geq \tau\right)$, which completes our proof from dominated convergence theorem. □

Theorem 6.2.2 (Sudakov–Fernique inequality)**.** *Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be two mean-zero Gaussian processes. Assume that, for all $t, s \in T$, we have*

$$\mathbb{E}\left[(X_t - X_s)^2\right] \leq \mathbb{E}\left[(Y_t - Y_s)^2\right]$$

Then it concludes that

$$\mathbb{E}\left[\sup_{t \in T} X_i\right] \leq \mathbb{E}\left[\sup_{t \in T} Y_i\right] \tag{132}$$

Proof. Define

$$f(x) := \frac{1}{\beta} \log \left(\sum_{i=1}^{n} e^{\beta x_i} \right)$$

It is easy to show that

$$\max_{i \leq n} x_i \leq f(x) \leq \max_{i \leq n} x_i + \frac{\log n}{\beta}$$

Hence $\lim_{\beta \to \infty} f(x) = \max_{i \leq n} x_i$. Next

$$\frac{d}{du} \mathbb{E}\left[f(Z(u)) \right] = \mathbb{E}\left[\sum_{i=1}^{n} \left(\frac{X_t^{(i)}}{\sqrt{u}} - \frac{Y_t^{(i)}}{\sqrt{1-u}} \right) p_i(Z(u)) \right]$$

where

$$p_i(x) := \frac{e^{\beta x_i}}{\sum_{j=1}^{n} e^{\beta x_j}}$$

In addition, we have

$$\mathbb{E}\left[X_t^{(i)} p_i(Z(u)) \right] = \sum_{j=1}^{n} \Sigma_{i,j}^{X_t} \mathbb{E}\left[\beta \sqrt{u} \frac{\partial}{\partial x_j} p_i(Z(u)) \right]$$

That is,

$$\begin{aligned}
\frac{d}{du} \mathbb{E}\left[f(Z(u)) \right] &= \beta \sum_{i,j=1}^{n} \left(\Sigma_{i,j}^{X_t} - \Sigma_{i,j}^{Y_t} \right) \mathbb{E}\left[\frac{\partial}{\partial x_j} p_i(Z(u)) \right] \\
&= \beta \sum_{i=1}^{n} \left(\Sigma_{i,i}^{X_t} - \Sigma_{i,i}^{Y_t} \right) \mathbb{E}\left[p_i(Z(u)) \left(1 - p_i(Z(u)) \right) \right] \\
&\quad - \beta \sum_{\substack{j,i=1 \\ j \neq i}}^{n} \left(\Sigma_{i,j}^{X_t} - \Sigma_{i,j}^{Y_t} \right) \mathbb{E}\left[p_i(Z(u)) p_j(Z(u)) \right]
\end{aligned}$$

Since

$$\Sigma_{i,j}^{X_t} = \mathbb{E}\left[X_t^{(i)} X_t^{(j)} \right] = \frac{1}{2} \left(\mathbb{E}\left[\left(X_t^{(i)} \right)^2 \right] + \mathbb{E}\left[\left(X_t^{(j)} \right)^2 \right] - \mathbb{E}\left[\left(X_t^{(i)} - X_t^{(j)} \right)^2 \right] \right)$$

As a result, we obtain

$$\begin{aligned}
\sum_{\substack{j,i=1 \\ j \neq i}}^{n} \Sigma_{i,j}^{X_t} \mathbb{E}\left[p_i(Z(u)) p_j(Z(u)) \right] &= \sum_{i=1}^{n} \mathbb{E}\left[\left(X_t^{(i)} \right)^2 \right] \left(\mathbb{E}\left[p_i(Z(u)) \sum_{j=1}^{n} p_j(Z(u)) \right] - \mathbb{E}\left[p_i^2(Z(u)) \right] \right) \\
&\quad - \sum_{j,i=1}^{n} \mathbb{E}\left[\left(X_t^{(i)} - X_t^{(j)} \right)^2 \right] \mathbb{E}\left[p_i(Z(u)) p_j(Z(u)) \right]
\end{aligned}$$

Since $\sum_{j=1}^{n} p_j(Z(u)) = 1$, we finally conclude that

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] = \frac{\beta}{2}\sum_{j,i=1}^{n}\left(\mathbb{E}\left[\left(X_t^{(i)} - X_t^{(j)}\right)^2\right] - \mathbb{E}\left[\left(Y_t^{(i)} - Y_t^{(j)}\right)^2\right]\right)\mathbb{E}\left[p_i(Z(u))p_j(Z(u))\right] \leq 0$$

which deduces that

$$\mathbb{E}\left[f(X)\right] \leq \mathbb{E}\left[f(Y)\right]$$

Take $\beta \to \infty$ and according to Lebesgue dominated convergence theorem, which completes our proof. □

Corollary 6.2.1 (Gaussian contraction inequality). *Consider a bounded subset $T \subset \mathbb{R}^n$, and let $g_1, \cdots, g_n$ be independent $\mathcal{N}(0,1)$ random variables. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be contractions, i.e., Lipschitz functions with $\|\phi_i\|_{\mathrm{Lip}} \leq 1$, then*

$$\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^{n} g_i\phi_i(t_i)\right] \leq \mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^{n} g_i t_i\right] \tag{133}$$

Proof. Define two Gaussian processes X_t and Y_t by

$$X_t := \sum_{i=1}^{n} g_i\phi_i(t_i) \quad \text{and} \quad Y_t := \sum_{i=1}^{n} g_i t_i$$

In addition, we have

$$\begin{aligned}\mathbb{E}\left[(X_t - X_s)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} g_i\left(\phi_i(t_i) - \phi_i(s_i)\right)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} g_i^2\left(\phi_i(t_i) - \phi_i(s_i)\right)^2\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{n} g_i^2\left(t_i - s_i\right)^2\right] = \mathbb{E}\left[(Y_t - Y_s)^2\right]\end{aligned}$$

Then according to Theorem 6.2.2, it completes our proof. □

Corollary 6.2.2 (Gordon's inequality). *Let $(X_{ut})_{u\in U, t\in T}$ and $(Y_{ut})_{u\in U, t\in T}$ be two mean-zero Gaussian processes indexed by pairs of points (u,t) in a product set $U \times T$. Assume that we have*

$$\mathbb{E}\left[X_{ut}^2\right] = \mathbb{E}\left[Y_{ut}^2\right] \quad \mathbb{E}\left[(X_{ut} - X_{us})^2\right] \leq \mathbb{E}\left[(Y_{ut} - Y_{us})^2\right] \quad \text{for all } u, t, s$$
$$\mathbb{E}\left[(X_{ut} - X_{vs})^2\right] \geq \mathbb{E}\left[(Y_{ut} - Y_{vs})^2\right] \quad \text{for all } u \neq v, t, s$$

Then for every $\tau \geq 0$ we have

$$\mathbb{P}\left(\inf_{u\in U}\sup_{t\in T} X_{ut} \geq \tau\right) \leq \mathbb{P}\left(\inf_{u\in U}\sup_{t\in T} Y_{ut} \geq \tau\right) \tag{134}$$

and consequently

$$\mathbb{E}\left[\inf_{u\in U}\sup_{t\in T} X_{ut} \geq \tau\right] \leq \mathbb{E}\left[\inf_{u\in U}\sup_{t\in T} Y_{ut} \geq \tau\right] \tag{135}$$

Proof. Define $h(\cdot)$ as in Lemma 6.2.1 and denote

$$f(x) := \prod_i \left(1 - \prod_j h(x_{ij})\right)$$

□

6.3 Sharp Bounds on Gaussian Matrices

Theorem 6.3.1 (Norms of Gaussian random matrices)**.** *Let A be an $m \times n$ matrix with independent $\mathcal{N}(0,1)$ entries. Then*

$$\mathbb{E}\left[\|A\|\right] \leq \sqrt{m} + \sqrt{n} \tag{136}$$

Proof. Define $X_{uv} := \langle Au, v\rangle$, where $u \in S_{n-1}$ and $v \in S_{m-1}$, it is easy to see that $X_{u,v} \sim \mathcal{N}(0,1)$. In addition,

$$\begin{aligned}\mathbb{E}\left[(X_{uv} - X_{wz})^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}(u_j v_i - w_j z_i)\right)^2\right] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}(u_j v_i - w_j z_i)^2 \leq \|u - w\|_2^2 + \|v - z\|_2^2\end{aligned}$$

For the last inequality, we only need to show that

$$1 + \sum_{i,j=1}^{n,m} u_i w_i v_j z_j \geq \sum_{i=1}^{n} u_i w_i + \sum_{j=1}^{m} v_j z_j$$

which is equivalent to

$$\left(1 - \sum_{i=1}^{n} u_i w_i\right)\left(1 - \sum_{j=1}^{m} v_j z_j\right) \geq 0$$

Next, define $Y_{uv} = \langle g, u\rangle + \langle h, v\rangle$, where $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and $h \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$, and $\mathbb{E}\left[(Y_{uv} - Y_{wz})\right] = \|u - w\|_2^2 + \|v - z\|_2^2$, then by Theorem 6.2.2, it concludes that

$$\begin{aligned}\mathbb{E}\left[\|A\|\right] &= \mathbb{E}\left[\max_{u \in S_{n-1}, v \in S_{m-1}} \langle Au, v\rangle\right] = \mathbb{E}\left[\max_{u \in S_{n-1}, v \in S_{m-1}} X_{uv}\right] \\ &\leq \mathbb{E}\left[\max_{u \in S_{n-1}, v \in S_{m-1}} Y_{uv}\right] \leq \mathbb{E}\left[\|g\|_2\right] + \mathbb{E}\left[\|h\|_2\right] \\ &\leq \mathbb{E}\left[\|g\|_2^2\right]^{1/2} + \mathbb{E}\left[\|h\|_2^2\right]^{1/2} = \sqrt{m} + \sqrt{n}\end{aligned}$$

which completes our proof. □

Corollary 6.3.1 (Norms of Gaussian random matrices: tails)**.** *Let A be an $m \times n$ matrix with independent $\mathcal{N}(0,1)$ entries. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\|A\| \geq \sqrt{m} + \sqrt{n} + t\right) \leq 2\exp\left(-ct^2\right) \tag{137}$$

where $c > 0$ is a constant.

Proof. Since we can regard A as $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{nm})$ and define $f(A) = \|A\|$, then by Theorem 4.2.2

$$\|f(A) - \mathbb{E}\left[f(A)\right]\|_{\psi_2} \leq C\|f\|_{\mathrm{Lip}}$$

In addition, since $f(A) \leq \|A\|_2$, it concludes that

$$\mathbb{P}\left(f(A) \geq \mathbb{E}\left[f(A)\right] + t\right) \leq \mathbb{P}\left(|f(A) - \mathbb{E}\left[f(A)\right]| \geq t\right) \leq 2\exp\left(-ct^2\right)$$

Finally, by Theorem 6.3.1, which completes our proof. □

Corollary 6.3.2 (Smallest singular values)**.** *The smallest singular value of an $m \times n$ random matrix A with independent $\mathcal{N}(0,1)$ entries has a lower bound of*

$$\mathbb{E}\left[s_n(A)\right] \geq \sqrt{m} - \sqrt{n} \tag{138}$$

Combine this result with concentration and then obtain

$$\mathbb{P}\left(\|A\| \leq \sqrt{m} - \sqrt{n} - t\right) \leq 2\exp\left(-ct^2\right) \tag{139}$$

Proof. According to Min-max principle for singular values, it concludes that

$$s_n(A) = \min_{u \in S_{n-1}} \max_{v \in S_{m-1}} \langle Au, v\rangle$$

Hence define $X_{uv} = \langle Au, v\rangle$ and $Y_{uv} = \langle g, u\rangle + \langle h, v\rangle$, then it has

$$\mathbb{E}\left[\left(X_{ut} - X_{us}\right)^2\right] = \mathbb{E}\left[\left(Y_{ut} - Y_{us}\right)^2\right] \quad \text{for all } u, t, s$$

$$\mathbb{E}\left[\left(X_{ut} - X_{us}\right)^2\right] = \mathbb{E}\left[\left(Y_{ut} - Y_{us}\right)^2\right] \quad \text{for all } u, t, s$$

□

6.4 Sudakov's Minoration Inequality

For a general Gaussian random process $(X_t)_{t\in T}$, define the canonical metric by

$$d(s,t) := \mathbb{E}\left[\left(X_t - X_s\right)^2\right]^{1/2} \tag{140}$$

which determines the covariance function $\Sigma(t,s)$ and in turn determines the distribution of the process $(X_t)_{t\in T}$. In this subsection, we will give a useful lower bound on $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ in terms of the metric entropy of the metric space (T,d), where $\log_2 \mathcal{N}\left(T,d,\epsilon\right)$ is called the metric entropy of T.

Theorem 6.4.1 (Sudakov's minoration inequality)**.** *Let $(X_t)_{t\in T}$ be a mean-zero Gaussian process. Then, for any $\epsilon \geq 0$, we have*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq c\epsilon\sqrt{\log_2 \mathcal{N}\left(T,d,\epsilon\right)} \tag{141}$$

where d is the canonical metric.

Proof. If (T, d) is not compact, i.e. $\mathcal{N}(T, d, \epsilon) = \infty$ for some $\epsilon > 0$, we can deduce there exists $\{t_i\}_{i=1}^{\infty} \in \mathbb{R}^n$ such that $\lim_{i\to\infty} t_i^{(k)} = \infty$, hence we can assume that this coordinate is 1 and others are zero, or consider a orthogonal rotation. Let $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and $\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq \mathbb{E}\left[\max_{i\leq n} t_i^{(1)} g_1\right]$, by Cheybshev's inequality, we conclude that

$$\mathbb{E}\left[\max_{i\leq n} t_i^{(1)} g_1\right] \geq t\mathbb{P}\left(\max_{i\leq n} t_i^{(1)} g_1 \geq t\right) \geq t\left(1 - \left(1 - \frac{1}{t_1^{(1)}}\right)^n\right)$$

Hence $\mathbb{E}\left[\sup_{t\in T} X_t\right] = \infty$. Now suppose $\mathcal{N}(T, d, \epsilon) := N < \infty$ and let $\mathcal{P}$ be ϵ-separated subset of T, then $|\mathcal{P}| \geq N$, hence it is enough to show that

$$\mathbb{E}\left[\sup_{t\in\mathcal{P}} X_t\right] \geq c\epsilon\sqrt{\log_2 N}$$

Define $Y_t := \epsilon g_t/\sqrt{2}$, where $t \in \mathcal{P}$ and $g_t \sim \mathcal{N}(0, 1)$ are independent, since

$$\mathbb{E}\left[(Y_t - Y_s)^2\right] = \epsilon^2 \geq \mathbb{E}\left[(X_t - X_s)^2\right]$$

By Theorem 6.2.2, it completes our proof. □

Corollary 6.4.1 (Sudakov's minoration inequality in $\mathbb{R}^n$)**.** *Let $T \subset \mathbb{R}^n$. Then, for any $\epsilon > 0$, we have*

$$\mathbb{E}\left[\sup_{t\in T}\langle g, t\rangle\right] \geq c\epsilon\sqrt{\log_2 \mathcal{N}(T, \epsilon)} \tag{142}$$

6.5 Gaussian Width

Definition 6.5.1. *The Gaussian width of a subset $T \subset \mathbb{R}^n$ is defined as*

$$w(T) := \mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle\right] \tag{143}$$

where $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$.

Proposition 6.5.1 (Gaussian width)**.** *According to above the definition, we have the following propositions.*

(i) The Gaussian width $w(T)$ is finite if and only if T is bounded.

(ii) The Gaussian width is invariant under affine unitary transformations. Thus, for every orthogonal matrix U and any vector y, we have

$$w(UT + y) = w(T)$$

(iii) The Gaussian width is invariant under the taking of convex hulls. Thus,

$$w(\mathrm{conv}(T)) = w(T)$$

(iv) The Gaussian width respects the Minkowski addition of sets and scaling. Thus, for $T, S \subset \mathbb{R}^n$ and $a \in \mathbb{R}$, we have

$$w(T+S) = w(T) + w(S), \quad w(aT) = |a|w(T)$$

(v) We have

$$w(T) = \frac{1}{2}w(T-T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g, x-y\rangle\right]$$

(vi) (Gaussian width and diameter). We have

$$\frac{1}{\sqrt{2\pi}}\text{diam}(T) \le w(T) \le \frac{\sqrt{n}}{2}\text{diam}(T)$$

Proof. For (i), suppose T is bounded, i.e. for $x \in T$, $\|x\|_2 \le M$, then it concludes that

$$w(T) := \mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle\right] \le M\mathbb{E}\left[\|g\|_2\right] \le M\sqrt{\mathbb{E}\left[\|g\|_2^2\right]} = M\sqrt{n}$$

On the other hand, suppose $w(T)$ is bounded, then by Theorem 6.4.1, it concludes that $\mathcal{N}(T, \epsilon)$ is finite for any $\epsilon > 0$, i.e. T is compact and hence bounded.

For (ii), since

$$\mathbb{E}\left[\sup_{z\in UT+y}\langle g, z\rangle\right] = \mathbb{E}\left[\sup_{z\in T}\langle g, Ux+y\rangle\right] = \mathbb{E}\left[\sup_{z\in T}\langle U^T g, x\rangle\right]$$

where $U^T g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, which concludes this proposition.

For (iii), since

$$w(\text{conv}(T)) = \mathbb{E}\left[\sup_{\substack{\alpha\in[0,1]\\ y,z\in T}} \alpha\langle g, y\rangle + (1-\alpha)\langle g, z\rangle\right] \le \sup_{\alpha\in[0,1]} \alpha\mathbb{E}\left[\sup_{y\in T}\langle g, y\rangle\right] + (1-\alpha)\mathbb{E}\left[\sup_{z\in T}\langle g, z\rangle\right]$$

Hence, $w(\text{conv}(T)) \le w(T)$. And the inverse is trivial due to $T \subset \text{conv}(T)$.

For (iv), since

$$\mathbb{E}\left[\sup_{x\in T+S}\langle g, x\rangle\right] = \mathbb{E}\left[\sup_{y\in T, z\in S}\langle g, y\rangle + \langle g, z\rangle\right] = \mathbb{E}\left[\sup_{y\in T}\langle g, y\rangle\right] + \mathbb{E}\left[\sup_{z\in T}\langle g, z\rangle\right]$$

and

$$\mathbb{E}\left[\sup_{x\in aT}\langle g, x\rangle\right] = \mathbb{E}\left[\sup_{x\in T}\langle g, ax\rangle\right] = \mathbb{E}\left[\sup_{x\in T}\langle ag, x\rangle\right] = \mathbb{E}\left[\sup_{x\in T}\langle |a|g, x\rangle\right]$$

The last inequality is valid due to the symmetry of g.

For (v), since

$$\mathbb{E}\left[\sup_{x\in T-T}\langle g,x\rangle\right]=\mathbb{E}\left[\sup_{y,z\in T}\langle g,y-z\rangle\right]=\mathbb{E}\left[\sup_{y\in T}\langle g,y\rangle+\sup_{z\in T}\langle g,z\rangle\right]$$

The last inequality is valid due to (iv).

For (vi), the lower bound can be deduced by

$$\begin{aligned}w(T)&=\frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g,x-y\rangle\right]\geq\frac{1}{2}\mathbb{E}\left[\max\left\{\langle g,x-y\rangle,\langle g,y-x\rangle\right\}\right]\\&=\frac{1}{2}\mathbb{E}\left[|\langle g,x-y\rangle|\right]=\sqrt{\frac{1}{2\pi}}\|x-y\|_2\end{aligned}$$

The last equality is valid since $\langle g,x-y\rangle\sim\mathcal{N}(0,\|x-y\|_2^2)$. Finally, for the upper bound, we have

$$\begin{aligned}w(T)&=\frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g,x-y\rangle\right]\leq\frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\|g\|_2\|x-y\|_2\right]\\&\leq\frac{\operatorname{diam}(T)}{2}\mathbb{E}\left[\|g\|_2\right]\leq\frac{\operatorname{diam}(T)}{2}\mathbb{E}\left[\|g\|_2^2\right]^{1/2}=\frac{\sqrt{n}}{2}\operatorname{diam}(T)\end{aligned}$$

which completes our proof. □

Corollary 6.5.1 (Gaussian width under linear transformations)**.** *For any $m\times n$ matrix A, it concludes that*

$$w(AT)\leq\|A\|w(T)\tag{144}$$

Proof. Denote $X_t:=\langle g,At\rangle$ and $Y_t:=\|A\|\langle h,t\rangle$ where $t\in T$ and $g\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_m)$ and $h\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_n)$, then

$$\mathbb{E}\left[(X_t-X_s)^2\right]=\sum_{i=1}^m\left(\sum_{j=1}^n a_{i,j}(t_j-s_j)\right)^2=\|A(t-s)\|_2^2\leq\|A\|^2\|t-s\|_2^2=\mathbb{E}\left[(Y_t-Y_s)^2\right]$$

By Theorem 6.2.2, we obtain

$$w(AT)=\mathbb{E}\left[\sup_{t\in T}X_t\right]\leq\mathbb{E}\left[\sup_{t\in T}Y_t\right]=\|A\|w(T)$$

which completes our proof. □

Definition 6.5.2 (Spherical width)**.** *The spherical width of a subset $T\subset\mathbb{R}^n$ is defined as*

$$w_s(T):=\mathbb{E}\left[\sup_{x\in T}\langle\theta,x\rangle\right]\quad\text{where }\theta\sim\operatorname{Unif}(S_{n-1})\tag{145}$$

Lemma 6.5.1 (Gaussian vs. spherical width)**.** *We have*

$$\left(\sqrt{n}-C\right)w_s(T) \le w(T) \le \left(\sqrt{n}+C\right)w_s(T) \tag{146}$$

Proof. Since

$$g = \|g\|_2 \frac{g}{\|g\|_2} := r\theta$$

where $\theta \sim \text{Unif}(S_{n-1})$ and r, θ are independent. Hence

$$w(T) = \mathbb{E}\left[\sup_{x\in T}\langle r\theta, x\rangle\right] = \mathbb{E}\left[r\sup_{x\in T}\langle \theta, x\rangle\right] = \mathbb{E}\left[r\right]\mathbb{E}\left[\sup_{x\in T}\langle \theta, x\rangle\right]$$

In addition, by Corollary 2.1.1, $|\mathbb{E}\left[r\right] - \sqrt{n}| \le C$, which completes our proof. □

6.6 Stable Dimension, Stable Rank, and Gaussian Complexity

In this subsection, we define the square form of Gaussian width by

$$h(T)^2 := \mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle^2\right] \tag{147}$$

where $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. It is not difficult to see that the squared and usual versions of the Gaussian width are equivalent up to a constant factor.

Theorem 6.6.1 (Equivalence)**.** *The following statement is valid.*

$$w(T-T) \le h(T-T) \le w(T-T) + C_1\text{diam}(T) \le Cw(T-T) \tag{148}$$

In particular, we have

$$2w(T) \le h(T-T) \le 2cW(T) \tag{149}$$

Proof. First, by Jensen's inequality, we have

$$h(T) = \mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle^2\right]^{1/2} \ge \mathbb{E}\left[\sqrt{\sup_{x\in T}\langle g, x\rangle^2}\right] = \mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle\right] = w(T)$$

In addition, denote $F : g \to \sup_{x,y\in T}\langle g, x-y\rangle$ and

$$F(g) = \sup_{x,y\in T}\langle g, x-y\rangle = \sup_{x,y\in T}\langle g-g', x-y\rangle + \sup_{x,y\in T}\langle g, x-y\rangle \le \text{diam}(T)\|g-g'\|_2 + F(g')$$

Hence, F is Lipschitz with $\|F\|_{\text{Lip}} = \text{diam}(T)$, by Theorem 4.2.2, it concludes

$$\|F(g) - \mathbb{E}\left[F(g)\right]\|_{\psi_2} \le C\text{diam}(T)$$

In addition, since

$$\text{Var}\left(F(g)\right) = \int_0^\infty \mathbb{P}\left(|F(g) - \mathbb{E}\left[F(g)\right]| \ge \sqrt{t}\right)dt \le 2\int_0^\infty \exp\left(-\frac{t}{\left(C\text{diam}(T)\right)^2}\right)dt = K\text{diam}(T)^2$$

As a result, we have

$$h(T-T)=\sqrt{\mathbb{E}\left[F(g)^2\right]}=\sqrt{\mathbb{E}\left[F(g)\right]^2+\mathrm{Var}\left(F(g)\right)}\le w(T-T)+C_1\mathrm{diam}(T)$$

Finally, due to the last property in Proposition 6.5.1, we conclude that

$$w(T-T)+C_1\mathrm{diam}(T)\le\left(1+\sqrt{\frac{\pi}{2}}\right)w(T-T)$$

which completes our proof. □

Definition 6.6.1 (Stable dimension)**.** *For a bounded set $T\subset\mathbb{R}^n$, the stable dimension of T is defined as*

$$d(T):=\frac{h(T-T)^2}{\mathrm{diam}(T)^2}\sim\frac{w(T)^2}{\mathrm{diam}(T)^2}\tag{150}$$

Lemma 6.6.1 (Algebraic bound of stable dimension)**.** *For any sets $T\subset\mathbb{R}^n$, we have*

$$d(T)\le\dim(T)\tag{151}$$

Proof. Suppose $\dim(T)=K$ and by rotation invariance we can assume that E is the coordinate subspace, i.e., $E=\mathbb{R}^k$. We have

$$\mathbb{E}\left[\sup_{x,y\in T}\langle g,x-y\rangle^2\right]\le\mathrm{diam}(T)^2\mathbb{E}\left[\sup_{z\in\mathcal{B}_2^K}\langle g,z\rangle^2\right]=\mathrm{diam}(T)^2K$$

The first inequality is valid due to Cauchy inequality and $\langle g,z\rangle\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_K)$, which completes our proof. □

As a consequence of above Lemma, if T is a Euclidean ball in any subspace of $\mathbb{R}^n$, then $d(T)=\mathrm{diam}(T)$. Let T be a finite set of points in $\mathbb{R}^n$. Then $d(T)\le C\log|T|$.

Proposition 6.6.1 (Ellipsoids)**.** *Let A be an $m\times n$ matrix, and let $\mathcal{B}_2^n$ denote the unit Euclidean ball. Then the squared mean width of the ellipsoid $A\mathcal{B}_2^n$ is the Frobenius norm of A, i.e.,*

$$h(A\mathcal{B}_2^n)=\|A\|_F\tag{152}$$

Proof. Since we have

$$h(A\mathcal{B}_2^n)^2=\mathbb{E}\left[\sup_{x\in\mathcal{B}_2^n}\langle g,Ax\rangle^2\right]=\mathbb{E}\left[\sup_{x\in\mathcal{B}_2^n}\langle A^Tg,x\rangle^2\right]$$

where $A^Tg\sim\mathcal{N}(\mathbf{0},A^TA)$ and $\langle A^Tg,x\rangle\sim\mathcal{N}(0,x^TA^TAx)$, which completes our proof since $\|Ax\|_2\le\|A\|_F\|x\|_2$. □

Definition 6.6.2 (Stable rank)**.** *The stable rank of an $m \times n$ matrix A is defined as*

$$r(A) := \frac{\|A\|_F^2}{\|A\|} \tag{153}$$

The robustness of the stable rank makes it a useful quantity in numerical linear algebra. The usual, algebraic, rank is the algebraic dimension of the image of A; in particular

$$\operatorname{rank}(A) = \dim(A\mathcal{B}_2^n)$$

Similarly, as a consequence of Proposition 6.6.1, we obtain

$$d(A\mathcal{B}_2^n) = \frac{\|A\|_F^2}{\operatorname{diam}\left(A\mathcal{B}_2^n\right)^2}$$

which shows that the stable rank is the statistical dimension of the image: Finally, note that the stable rank is always bounded by the usual rank:

Definition 6.6.3. *The Gaussian complexity of a subset $T \subset \mathbb{R}^n$ is defined as*

$$\gamma(T) := \mathbb{E}\left[\sup_{x \in T} |\langle g, x \rangle|\right] \tag{154}$$

where $g \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$.

Proposition 6.6.2 (Gaussian width vs. Gaussian complexity)**.** *Consider a set $T \subset \mathbb{R}^n$ and a point $y \in T$, then*

$$\frac{1}{3}\left[w(T) + \|y\|_2\right] \le \gamma(T) \le 2\left[w(T) + \|y\|_2\right] \tag{155}$$

Proof. First, since $T - T$ is original-symmetric, we have

$$\mathbb{E}\left[\sup_{x,y \in T} |\langle g, x - y \rangle|\right] = \mathbb{E}\left[\sup_{x,y \in T} \langle g, x - y \rangle\right]$$

Hence

$$\mathbb{E}\left[\sup_{x \in T} |\langle g, x \rangle|\right] \le \mathbb{E}\left[\sup_{x \in T} |\langle g, x - y \rangle| + |\langle g, y \rangle|\right] = 2w(T) + \|y\|_2 \mathbb{E}\left[|Z|\right]$$

where $Z \sim \mathcal{N}(0,1)$ since $\langle g, y \rangle \sim \mathcal{N}(0, \|y\|_2^2)$ and $\mathbb{E}\left[|Z|\right] = \sqrt{2/\pi}$. For the lower bound, we have

$$3\mathbb{E}\left[\sup_{x \in T} |\langle g, x \rangle|\right] \ge \mathbb{E}\left[\sup_{x \in T} |\langle g, x \rangle|\right] + 2\mathbb{E}\left[|\langle g, y \rangle|\right] \ge w(T) + \sqrt{2/\pi}\|y\|_2$$

which completes our proof. □

6.7 Random Projections of Sets

Lemma 6.7.1. *Let P be a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \mathrm{Unif}(G_{n,m})$. Let Q be an $m \times n$ matrix obtained by choosing the first m rows of a random $n \times n$ matrix $U \sim \mathrm{Unif}(O(n))$ drawn uniformly from the orthogonal group. Then*

(i) for any fixed point $x \in \mathbb{R}^n$, $\|Px\|_2$ and $\|Qx\|_2$ have the same distribution.

(ii) for any fixed point $z \in S_{m1}$, then

$$Q^T z \sim \mathrm{Unif}\left(S_{n-1}\right) \tag{156}$$

In other words, the map Q^T acts as a random isometric embedding of $\mathbb{R}^m$ into $\mathbb{R}^n$.

Proof. For (i), consider the singular decomposition of $P = \sum_{i=1}^n s_i(P) u_i v_i^T$ □

Theorem 6.7.1 (Sizes of random projections of sets)**.** *Consider a bounded set $T \subset \mathbb{R}^n$. Let P be a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \mathrm{Unif}(G_{n,m})$. Then, with probability at least $1 - 2e^{-m}$, we have*

$$\mathrm{diam}(PT) \le C\left(w_s(T) + \frac{m}{n}\mathrm{diam}(T)\right) \tag{157}$$

where $G_{n,m}$ is the Grassmann manifold which contains all m-dimensional subspaces of $\mathbb{R}^n$.

7 Chaining

This chapter presents some central concepts and methods for bounding random processes. Chaining is a powerful and general technique that can be used to prove uniform bounds on a random process $(X_t)_{t\in T}$.

7.1 Dudley's Inequality

Definition 7.1.1 (Sub-Gaussian increments)**.** *Consider a random process* $(X_t)_{t\in T}$ *on a metric space* (T, d)*. We say that the process has sub-Gaussian increments if there exists* $K \geq 0$ *such that*

$$\|X_t - X_s\|_{\psi_2} \leq K d(t, s) \tag{158}$$