

INVARIANCE PRINCIPLE OF RANDOM MATRIX

A Dissertation
Presented to
The Academic Faculty

By

JunTao Duan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology

August 2022

© JunTao Duan 2022

INVARIANCE PRINCIPLE OF RANDOM MATRIX

Thesis committee:

Dr. Heinrich Matzinger
Department of Mathematics
Georgia Institute of Technology

Dr. Vladimir Koltchinskii
Department of Mathematics
Georgia Institute of Technology

Dr. Ionel Popescu
Department of Mathematics and
Computer Science
University of Bucharest

Dr. Brani Vidakovic
Department of Statistics
Texas A & M University

Dr. Wenjing Liao
Department of Mathematics
Georgia Institute of Technology

Date approved: May 12, 2022

For my family, for their constant support.

ACKNOWLEDGMENTS

I would like to thank the members of my thesis committee for their help and suggestions in preparation of this work. Particularly, my advisors Heinrich and Ionel, without whom I would have been doomed to never complete it. Special thanks are due to the friends and colleagues who made this work possible.

TABLE OF CONTENTS

Acknowledgments	iv
List of Figures	vii
Chapter 1: Introduction	1
1.1 Random projection of deterministic vectors	1
1.2 Random projection and embedding of random vectors	3
Chapter 2: Product central limit theorem	6
2.1 Brief review of CLT for dependent random variables	6
2.2 Main theorems	7
2.3 Proof of Theorem 2: product central limit theorem	10
2.3.1 A proof based on Lindeberg swapping	10
2.3.2 Alternative assumptions	12
2.4 Proof of Theorem 3: rate of convergence	15
2.4.1 Proof of two Lemmas	17
2.4.2 Discussion on the assumptions	20
Chapter 3: Invariance of random matrix for the inner product	25
3.1 Main theorems	27

3.2	Proof of Theorem 4: CLT - random matrix preserves inner product	28
3.3	Proof of Theorem 5: rate of convergence	35
3.4	Simulation and open questions	37
Chapter 4: Invariance of of random matrix for the norm		41
4.1	Concentration of projected or embedded norm for sub-Gaussian variables	42
4.2	Random projection preserves distribution of norm	49
4.2.1	CLT for Random projection of norm	51
4.2.2	Simulation	65
4.2.3	Possibility of extending CLT to random embedding $m \geq O(n)$	67
4.3	Conjecture on rate of convergence	68
Chapter 5: Conclusion		71
References		73

LIST OF FIGURES

3.1	Random projected inner product ($m=10, n=100$)	37
3.2	Random projected inner product ($m=500, n=5000$)	37
3.3	Random embedded inner product ($m=500, n=50$)	38
3.4	Random embedded inner product ($m=5000, n=500$)	38
4.1	Random projected norm ($m=10, n=100$)	65
4.2	Random projected norm ($m=500, n=5000$)	65
4.3	Random embedded norm ($m=200, n=20$)	66
4.4	Random embedded norm ($m=2000, n=200$)	66

SUMMARY

Random matrix has been found useful in many real world applications. The celebrated Johnson-Lindenstrauss lemma states that certain geometric structure of deterministic vectors is preserved when projecting high dimensional space \mathbb{R}^n to a lower dimensional space \mathbb{R}^m . However, when random vectors are concerned, it is still unclear how the distribution of the geometry is affected by random matrices. Since random projection or embedding introduces dependence to independent random vectors, does it imply random matrices are inferior for transforming random vectors?

We will start with establishing a new central limit theorem for random variables with certain product dependence structure. At the same time, we obtained its Berry-Esseen type rate of convergence. Then we apply this general central limit theorem to random projection and embedding of two independent random vectors X, Z . In particular, we show the distribution of inner product structure is preserved by random matrices. Roughly speaking, two orthogonal random vectors remain orthogonal in the randomly projected lower dimensional space or randomly embedded high dimensional space. More importantly, we also quantitatively characterize the distortion of distribution introduced by random matrices. The error term has a bound at most $O(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})$.

Then we also establish the fact that random matrices have low distortion on the norm of a random vector. It is first justified by establishing concentration of the projected or embedded norm under sub-Gaussian assumptions. A central limit theorem for the randomly projected norm is established as well similar to the CLT for inner product.

CHAPTER 1

INTRODUCTION

In the first chapter, we will use two sections to introduce the problem of random projections. In the first section we introduce the historical development for random projection of deterministic vectors and discuss its applications. In the second section, we introduce the random projection and embedding of random vectors which is our focus of this thesis (based on the author's work [1, 2]).

1.1 Random projection of deterministic vectors

Due to the internet boom and computer technology advancement in the last few decades, data collection and storage have been growing exponentially. With 'gold' mining demand on the enormous amount of data reaches to a new level, we are facing many technical challenges in understanding the information we have collected. In many different cases, including text and images, data can be represented as points or vectors in high dimensional space. On one hand, it is very easy to collect more and more information about the object so that the dimensionality of the represented vectors grows quickly. On the other hand it is very difficult to analyze and create useful models for high dimensional data due to several reasons including computational difficulty as a result of curse of dimensionality [3] and high noise to signal ratio in random matrix setting [4, 5]. It is therefore necessary to reduce the dimensionality of the data while preserving the relevant structures.

The celebrated Johnson-Lindenstrauss lemma [6] states that random projections can be used as a general dimension reduction technique to embed topological structures in high dimensional Euclidean space into a low dimensional space without distorting its topology too much. Since then random projection has been found very useful in many applications such as signal processing and machine learning. For example fast Johnson-Lindenstrauss

random projection is used to approximate K-nearest neighbors to speed up computation [7, 8]. Random sketching uses random projection to reduce sample sizes in regression model and low rank matrix approximation [9]. Random projected features can be used to create low dimensional base classifiers which are combined as robust ensemble model [10]. Practitioners found applications of random projection in privacy and security [11] as well. Before we begin to state our problem, let us state the Johnson-Lindenstrauss lemma [12].

Lemma 1.1.1 (Johnson and Lindenstrauss). *Given a set of vectors $\{u_1, \dots, u_p\}$ in \mathbb{R}^n , for any $m \geq 8\epsilon^{-2} \log p$, there exists a linear map $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that*

$$(1 - \epsilon)\|u_i - u_j\| \leq \|Au_i - Au_j\| \leq (1 + \epsilon)\|u_i - u_j\|$$

Given two fixed vectors $X, Z \in \mathbb{R}^n$, by Johnson-Lindenstrauss lemma, we can find a random projection $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that the projected distance $\|AX - AZ\|$ has only a small distortion of the original distance $\|X - Z\|$. More precisely, take $\epsilon = \frac{3}{\sqrt{m}}$, we have (remarkably independent with the original dimension n)

$$\left[1 - O\left(\frac{1}{\sqrt{m}}\right)\right] \|X - Z\|^2 \leq \|A(X - Z)\|^2 \leq \left[1 + O\left(\frac{1}{\sqrt{m}}\right)\right] \|X - Z\|^2 \quad (1.1)$$

Equivalently, this property can be reformulated as random projection preserves the inner product of two vectors (equivalence can be obtained by elementary computation which we will present afterwards). Namely given X, Z two vectors in the unit ball of \mathbb{R}^n ($\|X\| \leq 1, \|Z\| \leq 1$), then there is a random projection $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$|\langle AX, AZ \rangle - \langle X, Z \rangle| \leq O\left(\frac{1}{\sqrt{m}}\right)$$

For general vectors not in the unit ball, the bound on the right hand side has the norms as a

factor

$$|\langle AX, AZ \rangle - \langle X, Z \rangle| \leq O\left(\frac{1}{\sqrt{m}}\right)\|X\|\|Z\| \quad (1.2)$$

To go from Equation 1.1 to Equation 1.2. We note by Johnson-Lindenstrauss Lemma 1.1.1, we can select a matrix A that has the property that X , Z and $X - Z$ satisfy norm preservation with error $O(\frac{1}{\sqrt{m}})$ similar to Equation 1.1. That is $|\|AX\|^2 - \|X\|^2|$, $|\|AZ\|^2 - \|Z\|^2|$ and $|\|A(X - Z)\|^2 - \|X - Z\|^2|$ are small and at most $O(\frac{1}{\sqrt{m}})$. Then linearly combining these small terms, we will obtain Equation 1.2.

$$\begin{aligned} & |\langle AX, AZ \rangle - \langle X, Z \rangle| \\ &= \frac{1}{2} |(\|A(X - Z)\|^2 - \|X - Z\|^2) - (\|AX\|^2 - \|X\|^2) - (\|AZ\|^2 - \|Z\|^2)| \\ &\leq \frac{3}{2} O\left(\frac{1}{\sqrt{m}}\right) \end{aligned}$$

To go from Equation 1.2 to Equation 1.1, we replace the vectors X and Z in Equation 1.2 by $(X - Z)$.

By looking at the inner product form Equation 1.2, Johnson-Lindenstrauss Lemma 1.1.1 essentially can be viewed as a random matrix preserving inner product of a set of fixed vectors with errors depend on the projected dimension m and number of vectors p .

1.2 Random projection and embedding of random vectors

The natural extension is to consider random vectors X , Z and ask the question that how random projections affect random vectors. Suppose X and Z are independent. After applying a random projection or embedding, independent random vectors become strongly dependent. Does this mean random projection is inferior to be used for dimension reduction of random vectors? Is there an invariance phenomenon in the distribution sense? Closeness of distribution is often characterized by the difference of cumulative distribution function

(cdf). If we look at inner product, then we will be interested in comparing cdf of inner products,

$$\sup_t |\mathbb{P}(\langle AX, AZ \rangle < t) - \mathbb{P}(\langle X, Z \rangle < t)|$$

We will try to address this question under the constraint that X and Z are independent random vectors in chapter 2 and chapter 3 by deriving central limit theorems and rate of convergence.

In chapter 2 we develop a general central limit theorem involving dependent random variables which will serve as a tool for chapter 3. We first briefly review the historical achievements on central limit theorem (CLT) involving dependent random variables in section 2.1. After that we prove product-CLT theorems and obtain the rate of convergence in section 2.2 and section 2.3. Along the way, we will also discuss some alternative assumptions for product-CLT theorems.

In chapter 3, using product-CLT theorems proved in chapter 2, we derive an invariance principle of random projections and random embeddings for independent random vectors which is similar to the inner product form of Johnson-Lindenstrauss lemma but extended to the distribution sense.

Furthermore, one particular important question is whether the distribution of norm $\langle X, X \rangle$ is preserved, so we need to understand

$$\sup_t |\mathbb{P}(\langle AX, AX \rangle < t) - \mathbb{P}(\langle X, X \rangle < t)|$$

which will be partially addressed in chapter 4. We first derive a concentration of the projected norm around original norm for any m, n (projections and embeddings) but with sub-Gaussian assumption on the random variables in section 4.1. For the case $m/n \rightarrow 0$ (random projections), we prove a CLT for the projected norm in section 4.2 without sub-Gaussian assumption.

Along the way, we use some simulations to validate our theoretical results. The random

distributions used to generate random vectors and random projection matrices range from common continuous distributions to discrete distributions.

In summary, our contributions include:

1. We prove random matrices preserve distribution of inner product of independent random vectors. Roughly speaking, two orthogonal random vectors remain orthogonal in the randomly projected lower dimensional space and randomly embedded higher dimensional space.
2. More importantly, we also quantitatively characterize the distortion of distribution introduced by random matrices. The error term has a bound at most $O(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})$. For $m \leq n$, this shows the error term is the same order as in Johnson-Lindenstrauss lemma.
3. We also establish the fact that random matrices have low distortion on the norm of a random vector by establishing concentration of the projected norm. For the case $m/n \rightarrow 0$, we obtain a stronger theorem which is a CLT for the randomly projected norm.
4. Lastly, a new central limit theorem is established for random variables with certain product dependence structure. At the same time, we obtained its Berry-Esseen type rate of convergence. This alone can be of great interests in many applications involving dependent random variables.

CHAPTER 2

PRODUCT CENTRAL LIMIT THEOREM

This chapter is focused on discussing central limit theorems (CLT) involving dependent random variables. It will be used as a tool to study random projections and embeddings. We will start with review of CLT for dependent random variables in section 2.1. Then we will state our results, product-CLT theorem and the rate of convergence, in section 2.2. After that, we will give proofs in section 2.3 and section 2.4

2.1 Brief review of CLT for dependent random variables

Central limit theorem plays an important role in probability and has many real world applications. One pitfall in the classical theory is that we can only deal with independent random variables. There are many attempts to extend the theory to handle dependent random variables. Hoeffding and Robbins [13] formulated one of the early result which shows CLT still holds for locally dependent sequence. One of the most interesting development is the martingale difference central limit theorem in [14]. In a nutshell, if the conditional variance converges in probability, then a Lindeberg condition implies CLT for the sequence.

Theorem 1 (Martingale CLT). *Let $\{x_k\}$ be a sequence of martingale differences, $\{\mathcal{F}_k\}$ be the natural filtration, Let $\mathbb{E} x_k^2 = 1$, denote $\mathbb{E}[x_k^2|\mathcal{F}_{k-1}] := \sigma_k^2$. If the following two conditions hold*

1. $\frac{1}{n} \sum_k \sigma_k^2 \xrightarrow{p} 1$
2. *Lindeberg condition: $\frac{1}{n} \sum_k \mathbb{E} x_k^2 I(|x_k| > \varepsilon \sqrt{n}) \rightarrow 0$ for all $\varepsilon > 0$.*

Then

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n x_k \rightarrow \mathcal{N}(0, 1)$$

The exact rate of convergence is obtained by [15]: with uniformly boundedness condition, the rate of convergence is shown as $O(\frac{\log n}{\sqrt{n}})$. Slightly more general results can be found in [16] and [17]. There is another line of research considers mixing weak dependence which is extensively discussed in [18] and [19]. A mixing condition requires dependence between random variables in the sequence decays as their positions are further apart. Essentially, far apart random variables become almost independent.

2.2 Main theorems

Theorem 2 (product-CLT). *Given random variables $\{x_k\}$ such that $\mathbb{E} x_k = 0$ and $\mathbb{E} x_k^2 = 1$. Given another sequence of random variables $\{y_k\}$. Assume $\{y_k\}$ are independent with $\{x_k\}$ (y_k and $y_{k'}$ could be dependent). Assume all third moments exist and bounded, namely there is fixed large number A*

$$\mathbb{E}[|x_k|^3] < A < \infty, \quad \mathbb{E}[|y_k|^3] < A < \infty, \quad \forall k \quad (2.1)$$

Further assume

$$\mathbb{E}[x_k | \mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[x_k^2 | \mathcal{F}_{k-1}] = 1 \quad (2.2)$$

where \mathcal{F}_k is the filtration generated by the (martingale difference) sequence $\{x_k\}$.

Assume $\{y_k\}$ satisfies

$$\frac{1}{n} \sum_{k=1}^n y_k^2 \xrightarrow{p} 1 \quad (2.3)$$

Then we have the following CLT

$$\frac{1}{\sqrt{n}} \sum_{k=1}^n x_k y_k \rightarrow \mathcal{N}(0, 1)$$

Remark. *The product-CLT can be viewed as an extension of Martingale-CLT Theorem 1. If X is vector of martingale difference sequence which has CLT by Theorem 1. Our product-CLT asserts that if there is another Y vector with complicated unknown depen-*

dence but satisfies a law of large number condition, then the dot product $X^T Y$ has a CLT. This extension is useful because no other CLT can deal with a sequence $\{X_i Y_i\}$ with unknown dependence. As we will see in the proof of invariance principle of random projections, there is no way to apply martingale-CLT. Moreover, the assumptions on conditional variance in martingale-CLT are very hard to verify in practice. In real world applications it is almost impossible to compute a conditional variance. Instead, we can decouple the dependence, for example extract a sequence of independent random variables X , and a sequence of Y that has complicated dependence controlled by law of large number on the squares so that we can apply our product-CLT.

In principle, one can replace the third order moment condition (Equation 2.1) by Lindeberg type condition. But we prefer third moments in this work since third moments is necessary in the control of rate of convergence which are more useful for our study on random projections. After proving the theorem in section 2.3, we will also give a few practical conditions that guarantees the weak law of large number (LLN) assumption (Equation 2.3) which will be useful for practitioners.

Moreover, we are interested in the rate of convergence. In developing a Berry-Esseen type rate of convergence theorem, we will also need assumptions on how fast the average of $\{y_k^2\}$ converges. We state our result as follows,

Theorem 3 (rate of convergence product-CLT). *Assume all conditions in Theorem 2 holds. Further assume if rate of convergence for LLN of y_k^2 is controlled by the following condition*

$$\mathbb{E} \left[1 \wedge \left| \sqrt{\frac{1}{n} \sum_{k=1}^n y_k^2} - 1 \right| \right] < O(\varepsilon_n) \quad (2.4)$$

where ε_n converges to zero. Then we have

$$\left| \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{k=1}^n x_k y_k < t\right) - \mathbb{P}(G < t) \right| \leq O\left(\frac{1}{\sqrt{n}} \vee \varepsilon_n\right) \quad \forall t \in \mathbb{R}$$

where G is the standard normal random variable.

Those two product-CLT theorems will be used to obtain invariance principle of random matrices in chapter 3. We shall present the proof in the following sections. Some of the techniques used in the proofs are very general, which play an important role in justifying some claims in chapter 3.

2.3 Proof of Theorem 2: product central limit theorem

2.3.1 A proof based on Lindeberg swapping

Proof Let us begin with the Lindeberg argument.

Take any function f from $C_c^\infty(\mathbb{R})$ smooth function with bounded support on the real line. Let $S_n = \sum_1^n x_i y_i$. Let $z, \{z_i\}_{1 \leq i \leq n}$ be independent standard normal random variables. It is sufficient to show

$$\mathbb{E}[f(\frac{1}{\sqrt{n}}S_n)] - \mathbb{E}[f(z)] \rightarrow 0$$

Our strategy is to split the difference into two parts

$$\mathbb{E}[f(\frac{1}{\sqrt{n}}S_n)] - \mathbb{E}[f(\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i y_i)], \quad \text{and} \quad \mathbb{E}[f(\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i y_i)] - \mathbb{E}[f(z)]$$

then show both are small.

First step, let us try to show

$$\mathbb{E}[f(\frac{1}{\sqrt{n}}S_n)] - \mathbb{E}[f(\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i y_i)] \rightarrow 0$$

We write the difference as a telescopic sum,

$$\Delta_n := f(\frac{1}{\sqrt{n}}S_n) - f(\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i y_i) = \sum_{k=1}^n f(T_k) - f(T_{k-1})$$

where

$$T_k := \frac{1}{\sqrt{n}} \left[\sum_{i=1}^k x_i y_i + \sum_{i=k+1}^n z_i y_i \right]$$

To make notation easier to read, denote

$$U_k := \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{k-1} x_i y_i + \sum_{i=k+1}^n z_i y_i \right]$$

It is easy to see $U_k = T_k - \frac{1}{\sqrt{n}}x_k y_k = T_{k-1} - \frac{1}{\sqrt{n}}z_k y_k$. Now let us take a Taylor expansion on $f(T_k), f(T_{k-1})$ around U_k ,

$$f(T_k) = f(U_k) + f'(U_k) \frac{1}{\sqrt{n}}x_k y_k + \frac{1}{2}f''(U_k) \frac{1}{n}x_k^2 y_k^2 + O(n^{-\frac{3}{2}}x_k^3 y_k^3 \sup_x f'''(x))$$

$$f(T_{k-1}) = f(U_k) + f'(U_k) \frac{1}{\sqrt{n}}z_k y_k + \frac{1}{2}f''(U_k) \frac{1}{n}z_k^2 y_k^2 + O(n^{-\frac{3}{2}}z_k^3 y_k^3 \sup_x f'''(x))$$

Since y_k is independent with x_k, z_k , by conditioning on \mathcal{F}_{k-1} we compute

$$\begin{aligned}\mathbb{E}[f'(U_k) \frac{1}{\sqrt{n}}x_k y_k] &= \mathbb{E}[f'(U_k) \frac{1}{\sqrt{n}}y_k \mathbb{E}[x_k | \mathcal{F}_{k-1}]] = 0 \\ \mathbb{E}[f'(U_k) \frac{1}{\sqrt{n}}z_k y_k] &= \mathbb{E}[f'(U_k) \frac{1}{\sqrt{n}}y_k \mathbb{E}[z_k]] = 0\end{aligned}$$

and found the first order terms match. Similar argument shows second order terms match,

$$\begin{aligned}\mathbb{E}[f''(U_k) \frac{1}{n}x_k^2 y_k^2] &= \mathbb{E}[\mathbb{E}[f''(U_k) \frac{1}{n}x_k^2 y_k^2 | \mathcal{F}_{k-1}]] \\ &= \mathbb{E}[\frac{1}{n}f''(U_k)y_k^2 \mathbb{E}[x_k^2 | \mathcal{F}_{k-1}]] \\ &= \mathbb{E}[\frac{1}{n}f''(U_k)y_k^2] \\ \mathbb{E}[f''(U_k) \frac{1}{n}z_k^2 y_k^2] &= \mathbb{E}[f''(U_k) \frac{1}{n}y_k^2] \mathbb{E}[z_k^2] \\ &= \mathbb{E}[\frac{1}{n}f''(U_k)y_k^2]\end{aligned}$$

Therefore, we obtain

$$\mathbb{E} f(T_k) - f(T_{k-1}) = O(n^{-\frac{3}{2}} \mathbb{E} x_k^3 y_k^3 \sup_x f'''(x))$$

Sum up the n terms,

$$\mathbb{E} \Delta_n = O(\frac{1}{\sqrt{n}} \mathbb{E}(x_k^3 + z_k^3) y_k^3 \sup_x f'''(x))$$

In the case x_k, y_k have finite third moments, we conclude replacing x_i by Gaussian random variables z_i will only introduce error of the order $n^{-1/2}$

$$\mathbb{E} \Delta_n = O\left(\frac{1}{\sqrt{n}}\right)$$

Now it suffices to show

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i y_i \rightarrow \mathcal{N}(0, 1)$$

Notice by computing the moment generating function, we can verify, for all n

$$\frac{1}{\sqrt{\sum_{i=1}^n y_i^2}} \sum_{i=1}^n z_i y_i \sim \mathcal{N}(0, 1)$$

Then by Slutsky's theorem and condition $\frac{1}{n} \sum_{i=1}^n y_i^2 \rightarrow 1$, we conclude our desired result. ■

2.3.2 Alternative assumptions

Here we discuss a variation of the law of large number condition of y_i^2 (Equation 2.3) in product-CLT. This version has the advantage that the assumptions are easier to verify in practice. We only impose the mixed second moments conditions which can be approximately computed with empirical data.

Proposition 2.3.1. *In Theorem 2, if y_k satisfies,*

$$\mathbb{E} y_k^2 \rightarrow 1, \quad \mathbb{E}[y_k^4] < C < \infty, \quad \forall 1 \leq k \leq n$$

Further assume the mixed second moments satisfy

$$\frac{1}{n^2} \sum_{i \neq j} \mathbb{E}[y_i^2 y_j^2] \xrightarrow{n \rightarrow \infty} 1 \tag{2.5}$$

Then the following LLN holds

$$\frac{1}{n} \sum_k y_k^2 \xrightarrow{p} 1$$

Proof By Chebyshev's inequality,

$$\mathbb{P}\left(\frac{1}{n} \left| \sum_i (y_i^2 - 1) \right| > \varepsilon\right) \leq \frac{\frac{1}{n^2} \mathbb{E}[|\sum_i (y_i^2 - 1)|^2]}{\varepsilon^2}$$

Notice by the assumptions on finite fourth moments and Equation 2.5,

$$\begin{aligned} \frac{1}{n^2} \mathbb{E}[|\sum_i (y_i^2 - 1)|^2] &= \frac{1}{n^2} \left[\sum_i \mathbb{E} y_i^4 + \sum_{i \neq j} \mathbb{E} y_i^2 y_j^2 - 2n \sum_i \mathbb{E} y_i^2 + n^2 \right] \\ &\rightarrow \frac{1}{n^2} \left[\sum_i \mathbb{E} y_i^4 + \sum_{i \neq j} \mathbb{E} y_i^2 y_j^2 \right] - 1 \\ &\rightarrow 0 \end{aligned}$$

Therefore we see for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n} \left| \sum_i (y_i^2 - 1) \right| > \varepsilon\right) \rightarrow 0$$

This implies there is a weak law of large number for the sequence $\{y_k^2\}$, namely

$$\frac{1}{n} \sum_k y_k^2 \xrightarrow{p} 1$$

■

In practice, Equation 2.5 can be verified by showing the average $\frac{1}{n^2} \sum_{i,j} y_i^2 y_j^2$ is close to 1. This is very convenient for practitioners. It turns out that the mixed second moments condition is equivalent to a fourth moment matching to fourth moment of standard normal which is a necessary condition for a CLT. And these results presented here (Propo-

sition 2.3.1 and Proposition 2.3.2) show it is also sufficient.

Proposition 2.3.2. *Assume $\mathbb{E} x_i^4 < C < \infty, \forall k$. The condition Equation 2.5 in Proposition 2.3.1 is equivalent to*

$$\mathbb{E} [P_n^4] \rightarrow 3, \quad \text{where } P_n = \frac{1}{\sqrt{n}} \sum_i x_i y_i \quad (2.6)$$

Proof

$$\mathbb{E}[P_n^4] = n^{-2} \sum_{1 \leq i_1, \dots, i_4 \leq n} \mathbb{E} x_{i_1} y_{i_1} \dots x_{i_4} y_{i_4}$$

Now we want to analyze the indices $I = \{i_1, i_2, i_3, i_4\}$. If one of index i_k is different from the other three, then $\mathbb{E}[x_{i_k} | \mathcal{F}_{k-1}] = 0$ implies the whole product vanishes. Therefore the only surviving terms must be either all indices the same or indices appear as pairs. Namely

$$\mathbb{E}[P_n^4] = n^{-2} \left[\sum_{1 \leq i \leq n} \mathbb{E} x_i^4 y_i^4 + 3 \sum_{1 \leq i \neq j \leq n} \mathbb{E} x_i^2 y_i^2 x_j^2 y_j^2 \right]$$

where the factor 3 is because there are three cases for pairs $(i_1 = i_2, i_3 = i_4)$, $(i_1 = i_3, i_2 = i_4)$ and $(i_1 = i_4, i_2 = i_3)$.

Then notice

$$\mathbb{E} x_i^2 y_i^2 x_j^2 y_j^2 = \mathbb{E}(x_i^2 x_j^2) \mathbb{E} y_i^2 y_j^2 = \mathbb{E} y_i^2 y_j^2$$

since $\mathbb{E}(x_i^2 x_j^2) = \mathbb{E}[\mathbb{E}[x_i^2 x_j^2 | \mathcal{F}_{\min(i,j)}]] = 1$. Combining the assumption that fourth moment is bounded we see

$$\mathbb{E}[P_n^4] \rightarrow 3 \iff \frac{1}{n^2} \sum_{i \neq j} \mathbb{E} y_i^2 y_j^2 \rightarrow 1$$

■

2.4 Proof of Theorem 3: rate of convergence

The proof will be several steps. First we record a variation formula in Lemma 2.4.1. Then we use the variation formula to rewrite the error term by introducing a standard normal variable in Lemma 2.4.2. Then we use Lindeberg type argument to reduce the control of error to the control of two separate terms. One term is a telescopic sum which we will control in Lemma 2.4.3 with the moments information. The other term is the difference of two cumulative distribution functions (cdfs) that are close to normal cdf which we will control in Lemma 2.4.4 with the LLN property of y_i^2 (condition Equation 2.4).

Lemma 2.4.1. *Let X and ξ be two independent random variables. Let $\sigma = \sqrt{\mathbb{E}\xi^2}$. Let Φ be the cumulative distribution of standard normal. Denote*

$$\delta = \sup_t |\mathbb{P}(X \leq t) - \Phi(t)| \quad \delta^* = \sup_t |\mathbb{P}(X + \xi \leq t) - \Phi(t)|$$

Then

$$\delta \leq 2\delta^* + \frac{5}{\sqrt{2\pi}}\sigma, \quad \delta^* \leq 2\delta + \frac{3}{2\sqrt{\pi}}\sigma$$

Proof See for example [15] ■

Lemma 2.4.2. *Denote*

$$\delta := \sup_t \left| \mathbb{P} \left(\frac{\sum x_i y_i}{\sqrt{n}} \leq t \right) - \Phi(t) \right|, \quad \delta_\xi := \sup_t \left| \mathbb{P} \left(\frac{\xi + \sum x_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P} \left(\frac{\xi}{\sqrt{n}} + G \leq t \right) \right|$$

Given the same setting in Theorem 3, and let G, ξ be independent standard normal random variable. Then

$$\delta \leq 2\delta_\xi + \frac{3}{\sqrt{n}}$$

Proof

By Lemma 2.4.1 we have

$$\begin{aligned}
\eta &:= \sup_t \left| \mathbb{P}(G \leq t) - \mathbb{P}\left(\frac{\xi}{\sqrt{n}} + G \leq t\right) \right| \\
&\leq 2 \sup_t |\mathbb{P}(G \leq t) - \mathbb{P}(G \leq t)| + \frac{3}{2\sqrt{\pi}} \sqrt{\frac{1}{n}} \\
&= \frac{3}{2\sqrt{\pi}} \frac{1}{\sqrt{n}}
\end{aligned}$$

Again by Lemma 2.4.1, we see

$$\begin{aligned}
\delta &\leq 2 \sup_t \left| \mathbb{P}\left(\frac{\xi + \sum x_i y_i}{\sqrt{n}} \leq t\right) - \Phi(t) \right| + \frac{3}{2\sqrt{\pi}} \frac{1}{\sqrt{n}} \\
&\leq 2(\delta_\xi + \eta) + \frac{3}{2\sqrt{\pi}} \frac{1}{\sqrt{n}} \\
&< 2\delta_\xi + \frac{3}{\sqrt{n}}
\end{aligned}$$

■

Now we are ready to prove the rate of convergence in Theorem 3.

Proof Let $\{z_i\}$ be a sequence of independent standard normal random variables which is independent from $\{x_i, y_i\}$. By conditioning, we can rewrite δ_ξ of Lemma 2.4.2

$$\begin{aligned}
\delta_\xi &= \sup_t \left| \mathbb{P}\left(\frac{\xi + \sum x_i y_i}{\sqrt{n}} \leq t\right) - \mathbb{P}\left(\frac{\xi}{\sqrt{n}} + G \leq t\right) \right| \\
&= \sup_t \left| \mathbb{P}\left(\frac{\xi + \sum x_i y_i}{\sqrt{n}} \leq t\right) - \mathbb{P}\left(\frac{\xi + \sum z_i y_i}{\sqrt{n}} \leq t\right) + \Delta_t \right| \\
&= \sup_t \left| \mathbb{E} \left[\sum_{m=1}^n \Phi(T_m) - \Phi(T_{m-1}) \right] + \Delta_t \right|
\end{aligned}$$

where

$$\Delta_t = \mathbb{P}\left(\frac{\xi + \sum z_i y_i}{\sqrt{n}} \leq t\right) - \mathbb{P}\left(\frac{\xi}{\sqrt{n}} + G \leq t\right)$$

$$T_m = t\sqrt{n} - \sum_{i=1}^m x_i y_i - \sum_{i=m+1}^n z_i y_i$$

Therefore with Lemma 2.4.3 controlling the part of telescopic sum and Lemma 2.4.4 controlling $\sup_t |\Delta_t|$ (which we will prove later in subsection 2.4.1), we see,

$$\sup_t \left| \mathbb{P} \left(\frac{\xi + \sum x_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P} \left(\frac{\xi}{\sqrt{n}} + G \leq t \right) \right| \leq O(\varepsilon_n \vee \frac{1}{\sqrt{n}})$$

Then by Lemma 2.4.2, we conclude the desired result

$$\sup_t \left| \mathbb{P} \left(\frac{\sum x_i y_i}{\sqrt{n}} \leq t \right) - \Phi(t) \right| \leq O(\varepsilon_n \vee \frac{1}{\sqrt{n}})$$

■

Remark. If we let $y_k = 1$ for all k , then we recover the rate of convergence $O(\frac{1}{\sqrt{n}})$ for a martingale difference sequence $\{x_k\}$. This is not contradicting the Martingale difference CLT which has a rate $O(\frac{\log n}{\sqrt{n}})$, see [15]. Martingale CLT is derived under a slightly weaker condition on variance, which only requires $\frac{1}{n} \sum_k \mathbb{E}[x_k^2 | \mathcal{F}_{k-1}] \rightarrow 1$ instead of our condition that $\mathbb{E}[x_k^2 | \mathcal{F}_{k-1}]$ to be constant 1 for all k .

2.4.1 Proof of two Lemmas

Lemma 2.4.3. If $\mathbb{E} x_k^3 < A < \infty, \mathbb{E} y_k^3 < A < \infty, \forall k$ then there is a constant c

$$\sup_t \left| \mathbb{E} \left[\sum_{m=1}^n \Phi(T_m) - \Phi(T_{m-1}) \right] \right| \leq \frac{c}{\sqrt{n}} \quad (2.7)$$

Proof Let $U_k = T_k - x_k y_k = T_{k-1} - z_k y_k$, then

$$\Phi(T_k) - \Phi(U_k) = \Phi'(U_k) \frac{1}{\sqrt{n}} x_k y_k + \frac{1}{2} \Phi''(U_k) \frac{1}{n} x_k^2 y_k^2 + O(n^{-\frac{3}{2}} |x_k^3 y_k^3| \sup_x \Phi'''(x))$$

$$\Phi(T_{k-1}) - \Phi(U_k) = \Phi'(U_k) \frac{1}{\sqrt{n}} z_k y_k + \frac{1}{2} \Phi''(U_k) \frac{1}{n} z_k^2 y_k^2 + O(n^{-\frac{3}{2}} |z_k^3 y_k^3| \sup_x \Phi'''(x))$$

Similar arguments from the CLT proof shows the first two terms match. Therefore

$$\begin{aligned} \left| \mathbb{E} \left[\sum_{m=1}^n \Phi(T_m) - \Phi(T_{m-1}) \right] \right| &\leq \mathbb{E} \left[\sum_{m=1}^n O(n^{-\frac{3}{2}} (|x_k^3| + |z_k^3|) |y_k^3| \sup_x \Phi'''(x)) \right] \\ &\leq \frac{c}{\sqrt{n}} \end{aligned}$$

Note $\Phi'''(x) = \frac{x^2-1}{\sqrt{2\pi}} e^{-x^2/2}$ and $|\sup_x \Phi'''(x)| < \frac{2}{5}$. ■

Lemma 2.4.4. *Suppose condition Equation 2.4*

$$\mathbb{E} \left[1 \wedge \left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| \right] \leq O(\varepsilon_n)$$

is satisfied. Then

$$\sup_t |\Delta_t| \leq O(\varepsilon_n \vee \frac{1}{\sqrt{n}})$$

Proof With similar computation in Lemma 2.4.2, we can remove the same variation term, the normal random variable $\frac{\xi}{\sqrt{n}}$ in Δ_t .

$$\begin{aligned} \sup_t |\Delta_t| &\leq \sup_t \left| \mathbb{P} \left(\frac{\xi + \sum z_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P}(G \leq t) \right| + \sup_t \left| \mathbb{P}(G \leq t) - \mathbb{P} \left(\frac{\xi}{\sqrt{n}} + G \leq t \right) \right| \\ &\leq 2 \sup_t \left| \mathbb{P} \left(\frac{\sum z_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P}(G \leq t) \right| + \frac{3}{2\sqrt{\pi}} \frac{1}{\sqrt{n}} + \frac{3}{2\sqrt{\pi}} \frac{1}{\sqrt{n}} \end{aligned}$$

Therefore,

$$\sup_t |\Delta_t| \leq 2 \sup_t \left| \mathbb{P} \left(\frac{\sum z_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P}(G \leq t) \right| + \frac{2}{\sqrt{n}}$$

$$\begin{aligned}
\sup_t \left| \mathbb{P} \left(\frac{\sum z_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P}(G \leq t) \right| &= \sup_t \mathbb{E} \left| \mathbb{P}(G \leq t \sqrt{\frac{n}{\sum y_i^2}}) - \mathbb{P}(G \leq t) \right| \\
&= \sup_t \mathbb{E} \left| \int_t^{t \sqrt{\frac{n}{\sum y_i^2}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right| \\
&:= \sup_t \mathbb{E} h(t) \\
&\leq \mathbb{E} \sup_t h(t)
\end{aligned}$$

where we denote $S_n = \sqrt{\frac{\sum y_i^2}{n}}$, $h(t) = \left| \int_t^{t/S_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right|$.

Notice $0 < h(t) < 1$ and $h(t) < |t - t/S_n| \frac{1}{\sqrt{2\pi}} e^{-\frac{\min(t^2, t^2/S_n^2)}{2}}$. So

$$\sup_t |h(t)| \leq 1 \wedge \sup_t \left[|t/S_n - t| \frac{1}{\sqrt{2\pi}} e^{-\frac{\min(t^2, t^2/S_n^2)}{2}} \right]$$

Notice the fact $\sup_x \frac{1}{\sqrt{2\pi}} |x e^{-\frac{x^2}{2}}| < \frac{1}{2}$. When $S_n > 1$, $\min(t^2, t^2/S_n^2) = t^2/S_n^2$, we conclude

$$\sup_t |h(t)| \leq 1 \wedge \frac{1}{2} |1 - S_n| \leq 1 \wedge |1 - S_n|$$

When $\frac{1}{2} < S_n < 1$, we have $4S_n^2 > 1$. Then $\min(t^2, t^2/S_n^2) \geq \frac{t^2}{4S_n^2}$. We see

$$\begin{aligned}
\sup_t |h(t)| &\leq 1 \wedge \left[|2 - 2S_n| \frac{t}{2S_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{4S_n^2}} \right] \\
&\leq 1 \wedge \frac{1}{2} |2 - 2S_n| \\
&= 1 \wedge |1 - S_n|
\end{aligned}$$

When $S_n < \frac{1}{2}$, we use the bound $\sup_t |h(t)| < 1$. And

$$\mathbb{P}(S_n < \frac{1}{2}) \leq 2 \mathbb{E}[1 \wedge |1 - S_n|, S_n < \frac{1}{2}]$$

Combining condition Equation 2.4, we conclude

$$\begin{aligned}
\mathbb{E}[\sup_t h(t)] &\leq \mathbb{E}\left[1 \wedge |S_n - 1|, S_n > \frac{1}{2}\right] + \mathbb{E}\left[1, S_n < \frac{1}{2}\right] \\
&\leq O(\varepsilon_n) + \mathbb{P}\left(S_n < \frac{1}{2}\right) \\
&\leq O(\varepsilon_n) + 2\mathbb{E}[1 \wedge |1 - S_n|, S_n < \frac{1}{2}] \\
&\leq O(\varepsilon_n)
\end{aligned}$$

Then we conclude

$$\sup_t |\Delta_t| \leq O(\varepsilon_n \vee \frac{1}{\sqrt{n}})$$

■

2.4.2 Discussion on the assumptions

A natural question is whether the condition Equation 2.4 is necessary for Theorem 3. We will first show the condition Equation 2.4 for Lemma 2.4.4 is sharp when x_i are replaced with independent standard normal random variables z_i . This is done by obtaining a lower bound for $\sup_t \mathbb{E} h(t)$. Since our theorems are essentially normal approximations. Then for general x_i , it is unlikely the condition Equation 2.4 is not sharp. This also implies the rate of convergence in Theorem 3 is optimal for this specific case.

Proposition 2.4.1.

$$\sup_t \mathbb{E} h(t) := \sup_t \left| \mathbb{P}\left(\frac{\sum z_i y_i}{\sqrt{n}} \leq t\right) - \mathbb{P}(G \leq t) \right| \geq O\left(\mathbb{E}\left[1 \wedge \left|\sqrt{\frac{\sum y_i^2}{n}} - 1\right|\right]\right) \quad (2.8)$$

Proof We follow the same notation as in the proof of Theorem 3. $S_n = \sqrt{\frac{\sum y_i^2}{n}}$. Let's take

$t = 1$ we find

$$\begin{aligned}
\mathbb{E} h(1) &= \mathbb{E} \left| \int_1^{1/S_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right| \\
&\geq c \mathbb{E} \left[\int_1^{1/S_n} dx, \frac{1}{S_n} \leq 2 \right] + \mathbb{E} \left[\int_1^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx, \frac{1}{S_n} > 2 \right] \\
&\geq c \mathbb{E} \left[\left| 1 - \frac{1}{S_n} \right|, S_n \geq \frac{1}{2} \right] + c \mathbb{E} \left[1, S_n < \frac{1}{2} \right]
\end{aligned}$$

where $c = \frac{1}{\sqrt{2\pi}} e^{-\frac{2^2}{2}} \geq \frac{1}{20}$.

We will further separate $S_n \geq \frac{1}{2}$ into three events.

$$\begin{aligned}
\frac{1}{2} \leq S_n \leq 1: \quad & \frac{1}{S_n} - 1 \geq 1 - S_n \geq 0 \\
1 < S_n < 2: \quad & 1 - \frac{1}{S_n} \geq \frac{1}{2}(S_n - 1) \geq 0 \\
S_n \geq 2: \quad & 1 - \frac{1}{S_n} \geq \frac{1}{2}
\end{aligned}$$

So overall on the event $S_n \geq \frac{1}{2}$, we have

$$\left| \frac{1}{S_n} - 1 \right| \geq \frac{1}{2} [1 \wedge |S_n - 1|]$$

Combining all together,

$$\begin{aligned}
\sup_t \mathbb{E} h(t) &\geq \mathbb{E} h(1) \geq \frac{1}{40} \mathbb{E} \left[1 \wedge |S_n - 1|, S_n \geq \frac{1}{2} \right] + \frac{1}{20} \mathbb{E} \left[1, S_n < \frac{1}{2} \right] \\
&\geq \frac{1}{40} \mathbb{E} [1 \wedge |S_n - 1|]
\end{aligned}$$

Therefore we conclude

$$\sup_t \left| \mathbb{P} \left(\frac{\sum z_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P}(G \leq t) \right| \geq \frac{1}{40} \mathbb{E} \left[1 \wedge \left| \sqrt{\frac{\sum y_i^2}{n}} - 1 \right| \right]$$

■

Let $\{x_i\}, \{y_i\}$ be i.i.d. random variables with mean zero, variance one (e.g. standard normal). Then by the classical CLT and Berry-Esseen we know

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i y_i < t\right) = \mathbb{P}(G < t) + O\left(\frac{1}{\sqrt{n}}\right) \quad \forall t \in \mathbb{R}$$

We will show for i.i.d. x_i, y_i , the rate of convergence obtained from our Theorem 3 is the same as Berry-Esseen in classical CLT. This implies the condition Equation 2.4 is essentially sharp for this specific example. In general, any nontrivial improvement will require more restrictive assumptions on the random variables.

Proposition 2.4.2. *In Theorem 3, if x_i, y_i are i.i.d. sequences with mean zero and variance one. Then condition Equation 2.4 is satisfied with $\varepsilon_n = \frac{1}{\sqrt{n}}$.*

Proof

For i.i.d. y_i mean zero and variance one, we have by CLT

$$\sqrt{n}\left(\sum y_i^2/n - 1\right) \rightarrow \mathcal{N}(0, 1)$$

and

$$\left(\sqrt{\sum y_i^2/n} + 1\right) \xrightarrow{p} 2$$

which we used the fact that (by LLN) for any $0 < \epsilon < 0.1$,

$$\begin{aligned} \mathbb{P}\left(\left|\sqrt{\sum y_i^2/n} - 1\right| > \epsilon\right) &= \mathbb{P}\left(\sum y_i^2/n > (1 + \epsilon)^2\right) + \mathbb{P}\left(\sum y_i^2/n < (1 - \epsilon)^2\right) \\ &\leq \mathbb{P}\left(\left|\sum y_i^2/n - 1\right| > \epsilon\right) \rightarrow 0 \end{aligned}$$

Now we can apply Slutsky's theorem,

$$\sqrt{n} \left(\sqrt{\frac{\sum y_i^2}{n}} - 1 \right) = \sqrt{n} \frac{(\sum y_i^2/n - 1)}{(\sqrt{\sum y_i^2/n} + 1)} \rightarrow \mathcal{N}(0, \frac{1}{4})$$

Therefore condition Equation 2.4 is satisfied with

$$\mathbb{E} \left(1 \wedge \left| \sqrt{\frac{\sum y_i^2}{n}} - 1 \right| \right) = O\left(\frac{1}{\sqrt{n}}\right)$$

Then Theorem 3 gives the same conclusion as Berry-Esseen. ■

The condition Equation 2.4 may not be easy to verify. Here we provide an alternative condition for our theorem. A more intuitive control of the LLN of y_k^2 would be controlling the tail probability directly, which will not be sharp.

Proposition 2.4.3. *In Theorem 3, condition Equation 2.4 can be replaced by*

$$\mathbb{P} \left(\left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| > O(\varepsilon_n) \right) \leq O(\varepsilon_n) \tag{2.9}$$

where $\varepsilon_n \rightarrow 0$. This condition is stronger than Equation 2.4. In other words, it is sufficient but not necessary for Theorem 3.

Proof Let $S_n = \sqrt{\frac{\sum y_k^2}{n}}$. Assume Equation 2.9 holds.

$$\begin{aligned} \mathbb{E} [1 \wedge |S_n - 1|] &= \mathbb{E} [1 \wedge |S_n - 1|, |S_n - 1| > O(\varepsilon_n)] + \mathbb{E} [1 \wedge |S_n - 1|, |S_n - 1| \leq O(\varepsilon_n)] \\ &\leq \mathbb{P}(|S_n - 1| > O(\varepsilon_n)) + O(\varepsilon_n) \\ &\leq O(\varepsilon_n) \end{aligned}$$

To show condition Equation 2.9 is stronger than condition Equation 2.4, we look at the example of i.i.d. $\{y_i\}$ sequence.

For i.i.d. y_i mean zero and variance one, we have $\sqrt{n}(\sum y_i^2/n - 1) \rightarrow \mathcal{N}(0, 1)$ and $(\sqrt{\sum y_i^2/n} + 1) \rightarrow 2$ in probability and we can apply Slutsky's theorem.

$$\sqrt{n} \left(\sqrt{\frac{\sum y_i^2}{n}} - 1 \right) = \sqrt{n} \frac{(\sum y_i^2/n - 1)}{(\sqrt{\sum y_i^2/n} + 1)} \rightarrow \mathcal{N}(0, \frac{1}{4})$$

Therefore condition Equation 2.4 is satisfied

$$\mathbb{E} \left(1 \wedge \left| \sqrt{\frac{\sum y_i^2}{n}} - 1 \right| \right) = O\left(\frac{1}{\sqrt{n}}\right)$$

However, condition Equation 2.9 is not satisfied since

$$\begin{aligned} \mathbb{P} \left(\left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| > O\left(\frac{1}{\sqrt{n}}\right) \right) &= \mathbb{P} \left(\sqrt{n} \left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| > O(1) \right) \\ &\approx \mathbb{P} \left(|\mathcal{N}(0, \frac{1}{4})| > O(1) \right) \\ &= O(1) \end{aligned}$$

■

CHAPTER 3

INVARIANCE OF RANDOM MATRIX FOR THE INNER PRODUCT

Suppose we have X, Z two independent random vectors. In this section, we will investigate how much the independence structure is preserved in the projected space. Let S be a random projection, the resulting projected random vectors SX, SZ will be dependent. We will see the distribution of inner product is preserved under certain conditions.

Given two independent random vectors in \mathbb{R}^n :

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

with i.i.d. entries and $\mathbb{E} x_i = \mathbb{E} z_i = 0$, $\mathbb{E} x_i^2 = \mathbb{E} z_i^2 = 1$, $\mathbb{E} |x_i|^3 \vee \mathbb{E} |z_i|^3 < C < \infty$.

Then it is clear the following CLT holds:

$$\frac{1}{\sqrt{n}} X^T Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i z_i \xrightarrow{d} \mathcal{N}(0, 1)$$

And the classical Berry-Esseen theorem [20, 21, 22] tells us

$$\sup_t \left| \mathbb{P} \left(\frac{1}{\sqrt{n}} X^T Z < t \right) - \mathbb{P} (\mathcal{N}(0, 1) < t) \right| \leq O \left(\frac{1}{\sqrt{n}} \right) \quad (3.1)$$

Consider a random matrix $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ whose entries have mean 0 and variance 1. Then the natural question is whether CLT holds for product of the randomly projected vectors SX and SZ . Namely

$$\frac{1}{\sqrt{n} a_{n,m}} X^T S^T S Z \xrightarrow{?} \mathcal{N}(0, 1)$$

where $a_{n,m}$ is a scaling parameter depending on both m and n . Moreover, if there is a CLT, we will need to derive the rate of convergence in the spirit of Berry-Esseen theorem, namely find

$$\sup_t \left| \mathbb{P} \left(\frac{1}{\sqrt{n} a_{n,m}} X^T S^T S Z < t \right) - \mathbb{P}(\mathcal{N}(0, 1) < t) \right| \leq ?$$

If we try to use existing CLT that dealing with dependent random variables, for example martingale CLT, it will not be applicable. The major difficulty is that there is no natural filtration since the terms in the sum will be very dependent so the conditional variance in martingale CLT is not computable. It turns out our product-CLT is the right tool to use. We decouple the dependence into the sequence of independent random variables X and another sequence $S^T S Z$ with complicated dependence.

Now what are the necessary conditions required to apply our product-CLT? Since $\{x_i\}$ is a sequence with independent random variables, it satisfies all conditions in Theorem 2 and Theorem 3. So we need to show the assumptions on the second dependent sequence

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} := \frac{1}{a_{n,m}} S^T S Z$$

are also satisfied. Denote i -th column of S as S_i , then $y_i = \frac{1}{a_{n,m}} S_i^T S Z$. Moreover, $\{y_i\}$ are identically distributed even though they are dependent random variables. The Lindeberg swap idea in Theorem 2 requires the variables y_i have finite third moments and a weak law of large number of y_i^2 . We shall prove the weak law of large number in Lemma 3.2.1. In the proof we will follow Proposition 2.3.1 using Chebyshev's inequality to show the weak law of large number statement. To find the rate of convergence, we will need to compute the exact order of Equation 2.4.

3.1 Main theorems

Theorem 4 (Randomly mapped inner product CLT). *Given two independent random vectors in \mathbb{R}^n :*

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

with i.i.d. entries. And assume $\mathbb{E} x_i = \mathbb{E} z_i = 0$, $\mathbb{E} x_i^2 = \mathbb{E} z_i^2 = 1$, $\mathbb{E} |x_i|^3 \vee \mathbb{E} |z_i|^3 < C < \infty$. Consider a random matrix $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with independent entries and $\mathbb{E} S_{i,j} = 0$ and $\mathbb{E} S_{i,j}^2 = 1$. Further assume S, X, Z are all independent and $\mathbb{E} S_{1,1}^8 \vee \mathbb{E} z_1^4 < C < \infty$, then we have

$$\frac{1}{\sqrt{m^2n + mn^2}} X^T S^T S Z \rightarrow \mathcal{N}(0, 1) \quad \text{as } m, n \rightarrow \infty$$

Theorem 5 (Invariance of randomly mapped inner product). *Given the same moment assumptions as in Theorem 4, the following bounds hold,*

$$\sup_t \left| \mathbb{P}\left(\frac{1}{\sqrt{m^2n + mn^2}} X^T S^T S Z < t\right) - \mathbb{P}(G < t) \right| \leq O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \quad (3.2)$$

$$\sup_t \left| \mathbb{P}\left(\frac{1}{\sqrt{m^2n + mn^2}} X^T S^T S Z < t\right) - \mathbb{P}\left(\frac{1}{\sqrt{n}} X^T Z < t\right) \right| \leq O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \quad (3.3)$$

where G is a standard normal random variable.

3.2 Proof of Theorem 4: CLT - random matrix preserves inner product

Lemma 3.2.1. *Given $m, n \rightarrow \infty$, $a_{n,m} = \sqrt{m^2 + mn}$, and $\mathbb{E} S_{i,j}^4 \vee \mathbb{E} S_{i,j}^8 \vee \mathbb{E} z_i^4 < C < \infty$. If we let*

$$y_i = \frac{1}{a_{n,m}} S_i^T S Z$$

then we have

$$\mathbb{E} y_i^2 \rightarrow 1, \forall i$$

and

$$\frac{1}{n} \sum_{i=1}^n y_i^2 \rightarrow 1$$

Proof First, we note y_i are identically distributed. By Proposition 2.3.1, it suffices to prove $\mathbb{E} y_i^2 \rightarrow 1$, $\mathbb{E} y_i^4 < C < \infty$, and $\mathbb{E} y_i^2 y_j^2 \rightarrow 1$.

For the second moment,

$$\begin{aligned} \mathbb{E}[y_1^2] &= \frac{1}{a_{n,m}^2} \mathbb{E}[(S_1^T S Z)^2] \\ &= \frac{1}{a_{n,m}^2} \sum_{1 \leq i, j \leq m, 1 \leq p, q \leq n} \mathbb{E} S_{i,1} S_{i,p} z_p S_{j,1} S_{j,q} z_q \\ &= \frac{1}{a_{n,m}^2} \sum_{1 \leq i, j \leq m, 1 \leq p, q \leq n} \mathbb{E} S_{i,1} S_{i,p} S_{j,1} S_{j,q} \mathbb{E} z_p z_q \end{aligned}$$

Notice the random matrix S and random vector Z are centered, $\mathbb{E} S = 0$, $\mathbb{E} Z = 0$. The surviving terms have to be even powers, which are $\{p = q \neq 1, i = j\}$, $\{p = q = 1, i, j\}$.

Therefore

$$\begin{aligned} \mathbb{E}[y_1^2] &= \frac{1}{a_{n,m}^2} \left[\sum_{1 \leq i \leq m, 2 \leq p \leq n} \mathbb{E} S_{i,1}^2 S_{i,p}^2 \mathbb{E} z_p^2 + \sum_{1 \leq i, j \leq m} \mathbb{E} S_{i,1}^2 S_{j,1}^2 \mathbb{E} z_1^2 \right] \\ &= \frac{1}{(m^2 + mn)} [m(n-1) + (m \mathbb{E} S_{1,1}^4 + m^2 - m)] \\ &= 1 + \frac{\mathbb{E} S_{1,1}^4 - 2}{m + n} \end{aligned}$$

$$= 1 + O\left(\frac{1}{m+n}\right) \rightarrow 1 \quad (3.4)$$

Now we will show $\mathbb{E} y_i^2 y_j^2 \rightarrow 1$ for all $i \neq j$.

$$\begin{aligned} \mathbb{E}[y_1^2 y_2^2] &= \frac{1}{a_{n,m}^4} \mathbb{E}[(S_1^T SZ)^2 (S_2^T SZ)^2] \\ &= \frac{1}{a_{n,m}^4} \sum \mathbb{E}(S_{i_1,1} S_{i_1,p_1} S_{j_1,1} S_{j_1,q_1} z_{p_1} z_{q_1}) (S_{i_2,2} S_{i_2,p_2} S_{j_2,2} S_{j_2,q_2} z_{p_2} z_{q_2}) \end{aligned}$$

First, there are eight indices in the summation. And $1 \leq i_1, i_2, j_1, j_2 \leq m, 1 \leq p_1, q_1, p_2, q_2 \leq n$. Since $\mathbb{E} S_{i,j} = 0, \mathbb{E} z_i = 0$, the surviving terms in the summation must have higher powers for $S_{i,j}$ and z_i . We will count the total number of possible such terms.

Surviving terms will satisfy the following condition

$$z_{p_1} z_{q_1} z_{p_2} z_{q_2} = z_p^2 z_q^2, \quad 1 \leq p, q \leq n$$

We will analyze and count in two different cases:

$$\{p, q\} \cap \{1, 2\} \neq \emptyset, \quad \{p, q\} \cap \{1, 2\} = \emptyset$$

There are still many sub-cases, we need to treat differently.

- Case 1: $\{p, q\} \cap \{1, 2\} \neq \emptyset$

- Case 1-1: $\{p, q\} \subseteq \{1, 2\}$.

- * Case 1-1-1: $p = q = 1$. Then each term is $\mathbb{E} S_{i_1,1}^2 S_{j_1,1}^2 S_{i_2,2} S_{i_2,1} S_{j_2,2} S_{j_2,1} z_1^4$.

Then $i_2 = j_2$ in order to have squares. So the total is

$$m^3 \mathbb{E} z_1^4 + O(m^2)$$

- * Case 1-1-2: $p = q = 2$. Same as the computation in case 1-1-1, we have

total

$$m^3 \mathbb{E} z_1^4 + O(m^2)$$

* Case 1-1-3: $p = 1, q = 2$. This will give us $\binom{4}{2} = 6$ separate cases.

p_1	q_1	p_2	q_2
1	1	2	2
1	2	1	2
1	2	2	1
2	1	1	2
2	1	2	1
2	2	1	1

Only the first case (1, 1, 2, 2) produces terms $\mathbb{E} S_{i_1,1}^2 S_{j_1,1}^2 S_{i_2,2}^2 S_{j_2,2}^2 z_1^2 z_2^2$.

In total it is $m^4 + O(m^3)$. All other five cases admit similar analysis with

same number of terms, we only show the second case (1, 2, 1, 2), which is

$\mathbb{E} S_{i_1,1}^2 S_{j_1,1} S_{j_1,2} S_{i_2,2} S_{i_2,1} S_{j_2,2}^2 z_1^2 z_2^2$. Then $j_1 = i_2$ must hold for the surviving

terms, which in total is $m^3 + O(m^2)$. Combining all together, we have

in total

$$m^4 + O(m^3)$$

– Case 1-2: $p = 1, q \notin \{1, 2\}$. Same as 1-1 there are $\binom{4}{2} = 6$ separate cases.

p_1	q_1	p_2	q_2
1	1	q	q
1	q	1	q
1	q	q	1
q	1	1	q
q	1	q	1
q	q	1	1

Only the first case (1, 1, q, q) produces terms $\mathbb{E} S_{i_1,1}^2 S_{j_1,1}^2 S_{i_2,2} S_{i_2,q} S_{j_2,2} S_{j_2,q} z_1^2 z_q^2$.

In this case $i_2 = j_2$ must hold. In total, there are $m^3(n - 2) + O(m^2n)$ terms.

All other five cases have similar analysis with same number of terms, we only show the the second case $(1, q, 1, q)$, $\mathbb{E} S_{i_1,1}^2 S_{j_1,1} S_{j_1,q} S_{i_2,2} S_{i_2,1} S_{j_2,2} S_{j_2,q} z_1^2 z_q^2$. In this case $j_1 = i_2 = j_2$ must hold. In total, there are $m^2(n - 2) + O(mn)$ terms. Combining all together, we have in total

$$m^3n + O(m^3 + m^2n)$$

– Case 1-3: $p = 2, q \notin \{1, 2\}$. Again there are $\binom{4}{2} = 6$ separate cases.

p_1	q_1	p_2	q_2
2	2	q	q
2	q	2	q
2	q	q	2
q	2	2	q
q	2	q	2
q	q	2	2

Only the last case $(q, q, 2, 2)$ produces terms $\mathbb{E} S_{i_1,1} S_{i_1,q} S_{j_1,1} S_{j_1,q} S_{i_2,2}^2 S_{j_2,2}^2 z_q^2 z_2^2$.

Then $i_1 = j_1$ must hold. In total it is $m^3(n - 2) + O(m^2n)$.

All other five cases have similar analysis, we only show the first $(2, 2, q, q)$ here.

$\mathbb{E} S_{i_1,1} S_{i_1,2} S_{j_1,1} S_{j_1,2} S_{i_2,2} S_{i_2,q} S_{j_2,2} S_{j_2,q} z_q^2 z_2^2$. Then $i_1 = j_1, i_2 = j_2$ must hold for the surviving terms, which in total is $m^2(n - 2) + O(mn)$. Combining all together, we have in total

$$m^3n + O(m^3 + m^2n)$$

- Case 2: $\{p, q\} \cap \{1, 2\} = \emptyset$.

To have squares for variables from matrix S , we must have squares produced for

$S_{i_1,1}S_{j_1,1}S_{i_2,2}S_{j_2,2}$ and $S_{i_1,p_1}S_{j_1,q_1}S_{i_2,p_2}S_{j_2,q_2}$ separately. Therefore $i_1 = j_1, i_2 = j_2$. Denote $i_1 := i, j_1 := j$. Then we can further split into two cases, $i = j$ and $i \neq j$.

- Case 2-1: $\{p, q\} \cap \{1, 2\} = \emptyset$ and $i = j$. Then each term involving S is $\mathbb{E} S_{i,1}^2 S_{i,2}^2 S_{i,p_1} S_{i,p_2} S_{i,q_1} S_{i,q_2}$. This will produce 3 possible matches for $\{p_1, p_2, q_1, q_2\} = \{p, q\}$. which counting all indices will yields total $3m(n-2)^2$ terms. Some of those terms will have $p = q$, which will produce $m(n-2) \mathbb{E} S_{1,1}^4 \mathbb{E} z_1^4$ which is of a smaller order. So total will be

$$3mn^2 + O(mn)$$

- Case 2-2: $\{p, q\} \cap \{1, 2\} = \emptyset$ and $i \neq j$. In this case $\{p_1 = q_1, p_2 = q_2\}$ must be true. That in total will produce $(m^2 - m)[(n-2)^2 - (n-2)]$ terms which we excluded the cases when $p = q$. Then the cases of $p = q$ in total are $(m^2 - m)(n-2)$ of $\mathbb{E} z_1^4$. In total

$$\begin{aligned} & (m^2 - m)[(n-2)^2 - (n-2)] + (m^2 - m)(n-2) \mathbb{E} z_1^4 \\ & = m^2 n^2 - mn^2 + (\mathbb{E} z_1^4 - 5)m^2 n + O(mn + m^2) \end{aligned}$$

Adding all the cases together we obtain

$$\mathbb{E}[y_1^2 y_2^2] = \frac{1}{(m^2 + mn)^2} [m^4 + 2m^3 n + m^2 n^2 + O(m^3 + m^2 n + mn^2)] \quad (3.5)$$

$$= 1 + \frac{O(m^3 + m^2 n + mn^2)}{(m^2 + mn)^2} \quad (3.6)$$

$$= 1 + O\left(\frac{1}{m} + \frac{1}{m+n}\right) \rightarrow 1 \quad (3.7)$$

Therefore,

$$\frac{1}{n^2} \sum_{i \neq j} \mathbb{E}[(y_i^2 - 1)(y_j^2 - 1)] = \frac{n^2 - n}{n^2} \mathbb{E}[y_1^2 y_2^2 - y_1^2 - y_2^2 + 1] \rightarrow 0$$

For the fourth moment, we will show $\mathbb{E} y_i^4 \leq C < \infty$.

$$\begin{aligned} \mathbb{E}[y_1^4] &= \frac{1}{a_{n,m}^4} \mathbb{E}[(S_1^T S Z)^4] \\ &= \frac{1}{a_{n,m}^4} \sum \mathbb{E} S_{i_1,1} S_{i_1,p_1} S_{j_1,1} S_{j_1,q_1} z_{p_1} z_{q_1} \\ &\quad S_{i_2,1} S_{i_2,p_2} S_{j_2,1} S_{j_2,q_2} z_{p_2} z_{q_2} \end{aligned}$$

Similarly, the surviving terms are $\{p_1 = q_1, p_2 = q_2, i_1 = j_1, i_2 = j_2\}$, $\{p_1 = p_2, q_1 = q_2, i_1 = i_2, j_1 = j_2\}$ and $\{p_1 = q_2, q_1 = p_2, i_1 = j_2, j_1 = i_2\}$ which in total will give $3m^2n^2 + m^4 + 6m^3n + O(m^3 + m^2n + mn^2)$ where m^4 comes from counting terms of the form $\{p_1 = q_1 = p_2 = q_2 = 1\}$, and m^3n comes from

$$\{p_1, q_1, p_2, q_2\} = \{1, q\}$$

Therefore

$$\mathbb{E} y_i^4 = 3 \frac{n}{m+n} + \frac{m^2}{(m+n)^2} + \frac{O(m^3 + m^2n + mn^2)}{(m^2 + mn)^2} \leq 4 + O\left(\frac{1}{m} + \frac{1}{m+n}\right)$$

Lastly we shall apply Chebyshev's inequality.

$$\mathbb{P}\left(\left|\frac{1}{n} \sum y_i^2 - 1\right| > \varepsilon\right) \leq \frac{1}{n^2 \varepsilon^2} \left[\sum \mathbb{E}(y_i^2 - 1)^2 + \sum_{i \neq j} \mathbb{E}(y_i^2 - 1)(y_j^2 - 1) \right] \rightarrow 0$$

■

Proof (Theorem 4)

Combing Lemma 3.2.1 with the product-CLT Theorem 2, we conclude random projection preserves distribution of inner product of X and Z , namely given conditions in

Theorem 4 we have

$$\frac{1}{\sqrt{m^2n + mn^2}} X^T S^T S Z \rightarrow \mathcal{N}(0, 1) \quad \text{as } m, n \rightarrow \infty$$

■

As in every computation, we do not specify whether $m > n$ or $n > m$. Then the results hold for both random embeddings and projections.

3.3 Proof of Theorem 5: rate of convergence

Now we shall discuss the rate of convergence. Obtaining the exact rate is usually very hard since both lower bound and upper bound need to be obtained. In our case, one has to compute the exact rate of convergence for the law of large number statement on y_i^2 (namely condition Equation 2.4) for which the quantity is not practically computable if no further information given. However it is possible to carry out an argument (for example using relaxations or proposition Proposition 2.4.3) to obtain an upper bound.

Proof (Theorem 5)

We will start with relaxation. Since $\left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| \leq \left| \frac{\sum y_k^2}{n} - 1 \right|$

$$\begin{aligned} \mathbb{E} \left(1 \wedge \left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| \right) &\leq \mathbb{E} \left(1 \wedge \left| \frac{\sum y_k^2}{n} - 1 \right| \right) \\ &\leq \mathbb{E} \left[\left| \frac{\sum y_k^2}{n} - 1 \right| \right] \\ &\leq \sqrt{\mathbb{E} \left[\left(\frac{\sum y_k^2}{n} - 1 \right)^2 \right]} \end{aligned}$$

The last step uses Jensen's inequality and $f(x) = \sqrt{x}$ is concave. Notice,

$$\begin{aligned} \mathbb{E} \left[\left(\frac{\sum y_k^2}{n} - 1 \right)^2 \right] &= \frac{1}{n^2} \left[\sum_{k=1}^n \mathbb{E}(y_k^2 - 1)^2 + \sum_{i \neq j} \mathbb{E}(y_i^2 - 1)(y_j^2 - 1) \right] \\ &= \frac{1}{n} \mathbb{E}(y_1^2 - 1)^2 + \frac{n^2 - n}{n^2} \mathbb{E}(y_1^2 - 1)(y_2^2 - 1) \end{aligned}$$

Therefore computing a bound for the rate of convergence boils down to compute the order of $\mathbb{E}(y_1^2 - 1)^2$ and $\mathbb{E}(y_1^2 - 1)(y_2^2 - 1)$ explicitly which have already been computed in the proof of Lemma 3.2.1.

$$\mathbb{E}[y_2^2] = \mathbb{E}[y_1^2] = 1 + O\left(\frac{1}{m+n}\right)$$

$$\begin{aligned}\mathbb{E}[y_1^4] &\leq 4 + O\left(\frac{1}{m} + \frac{1}{m+n}\right) \leq 4 + O\left(\frac{1}{m}\right) \\ \mathbb{E} y_1^2 y_2^2 &= 1 + O\left(\frac{1}{m} + \frac{1}{m+n}\right) = 1 + O\left(\frac{1}{m}\right)\end{aligned}$$

This implies

$$\mathbb{E}(y_1^2 - 1)^2 = O\left(\frac{1}{m}\right), \quad \mathbb{E}(y_1^2 - 1)(y_2^2 - 1) = O\left(\frac{1}{m}\right)$$

and So we conclude

$$\begin{aligned}\mathbb{E}\left(1 \wedge \left|\sqrt{\frac{\sum y_k^2}{n}} - 1\right|\right) &\leq \sqrt{\mathbb{E}\left[\left(\frac{\sum y_k^2}{n} - 1\right)^2\right]} \\ &= O\left(\frac{1}{\sqrt{m}}\right)\end{aligned}$$

Applying Theorem 3, we conclude Equation 3.2. Then combining Berry-Esseen inequality Equation 3.1 and triangle inequality, we obtain Equation 3.3. ■

3.4 Simulation and open questions

We first give some simulations to show the random embedded or projected inner product converges to normal distribution.

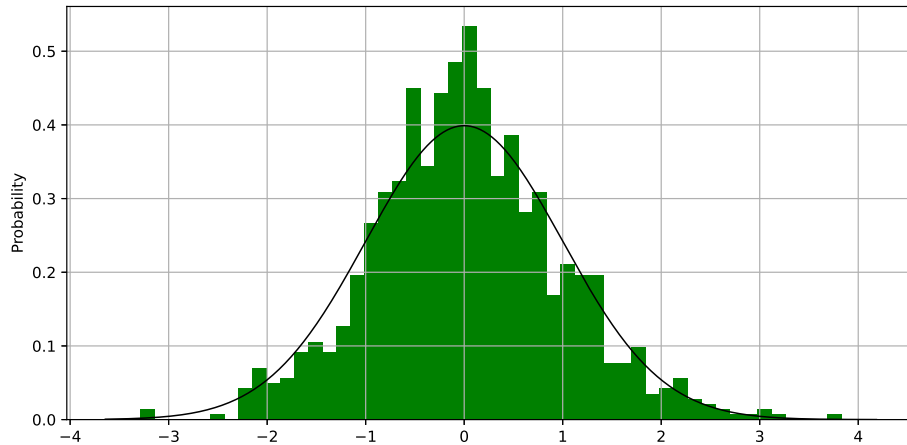


Figure 3.1: Random projected inner product (m=10, n=100)

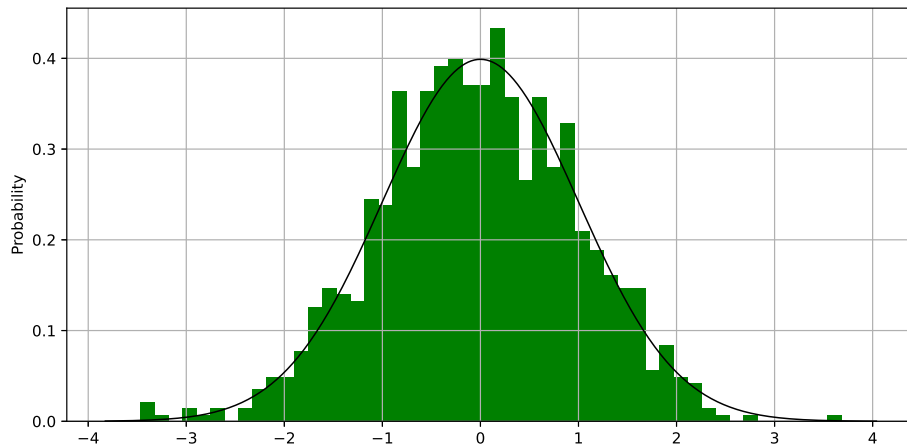


Figure 3.2: Random projected inner product (m=500, n=5000)

Figure 3.1 and Figure 3.2 plotted histograms of 1000 samples of the projected inner product $\frac{1}{\sqrt{m^2n+mn^2}} X^T S^T S Z$ with different dimension settings. The random variables we

used for X, S, Z are standard normal random variables. As dimension m, n increases, the convergence improves.

Next we give simulations for random embedded inner product where $m > n$.

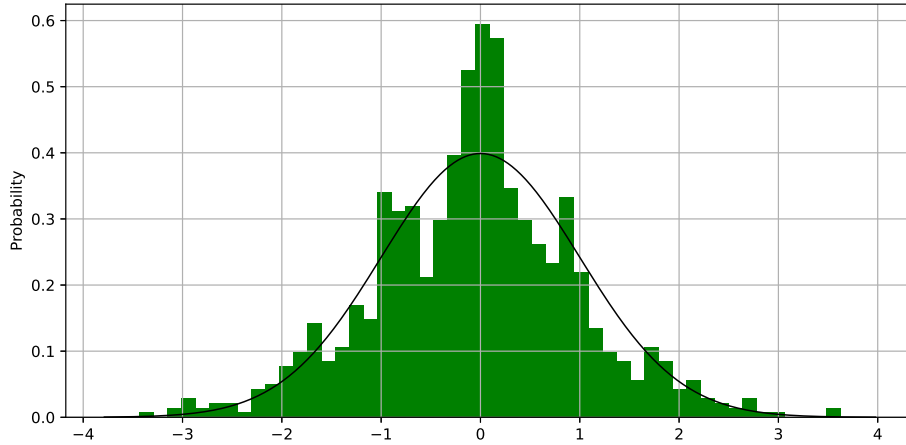


Figure 3.3: Random embedded inner product ($m=500, n=50$)

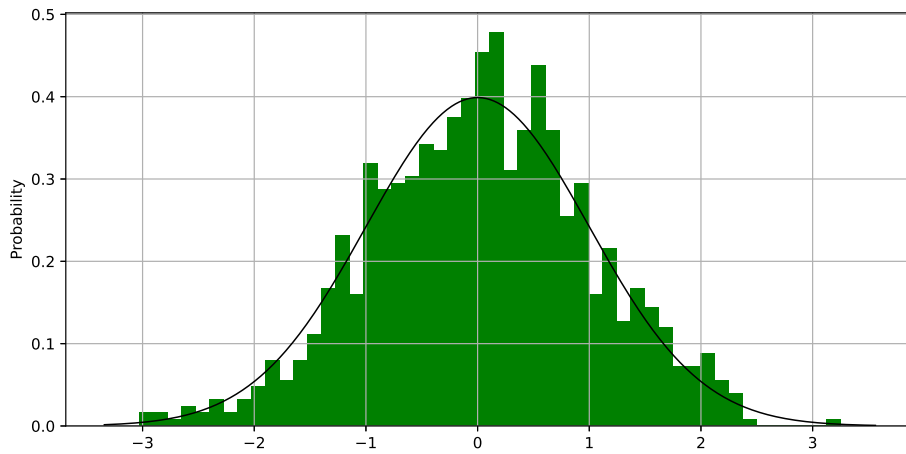


Figure 3.4: Random embedded inner product ($m=5000, n=500$)

Figure 3.3 and Figure 3.4 plotted histograms of 1000 samples of the embedded inner product $\frac{1}{\sqrt{m^2n+mn^2}}X^T S^T S Z$. The random variables we used for X, S, Z take discrete values $\{-2.5, 0, 2.5\}$ with probability $\{0.08, 0.84, 0.08\}$. The kurtosis of such random

variable is 6.25 which is much larger than standard normal random variable. Again as dimensions m, n increase, the histogram converges to a standard normal shape.

To have CLT result in Theorem 4, it is essential the dimension of the projected space m diverges. Fixed m will not lead to a CLT.

For example, we let $m = 1, n \rightarrow \infty$. Then let $X \in \mathbb{R}^n$ be Gaussian vector, $Z \in \mathbb{R}^n$ be Rademacher vector and let $S : \mathbb{R}^n \rightarrow \mathbb{R}$ has Rademacher entries. Suppose all random variables are independent, then

$$\frac{1}{\sqrt{n}}X^T Z = \mathcal{N}(0, 1)$$

which holds exactly without error. On the other hand

$$\frac{1}{\sqrt{1^2n + 1n^2}}X^T S^T S Z \sim \mathcal{N}(0, 1) \times \mathcal{N}'(0, 1) + O\left(\frac{1}{\sqrt{n}}\right)$$

that is the product of two independent standard Gaussian random variable. To see this is the case, note first $\frac{1}{\sqrt{n}}X^T S^T$ is exactly standard Gaussian $\mathcal{N}(0, 1)$. $\frac{1}{\sqrt{n+1}}S Z$ converges to another $\mathcal{N}'(0, 1)$ with error $O\left(\frac{1}{\sqrt{n}}\right)$. The independence is due to the fact that Rademacher in S can be absorbed into X and Z so that we may replace all entries of S by constant 1's. Therefore the cdf of $\frac{1}{\sqrt{n}}X^T Z$ and $\frac{1}{\sqrt{1^2n+1n^2}}X^T S^T S Z$ differ by $O(1) = O\left(\frac{1}{\sqrt{m}}\right)$.

The bound Equation 3.2 in general can not be improved if there is no additional assumption. $O\left(\frac{1}{\sqrt{n}}\right)$ is necessary as it is in Berry-Esseen. $O\left(\frac{1}{\sqrt{m}}\right)$ is also very likely to be necessary as the above example achieves the error rate when $m = 1$. For general m we do not pursue a precise proof here but we give some heuristics. Let $X \in \mathbb{R}^n$ be standard Gaussian vector, $Z \in \mathbb{R}^n$ be standard Rademacher vector and let $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ has Rademacher entries as well. Suppose all random variables are independent. Denote $Y = \frac{1}{\sqrt{m^2+mn}}S^T S Z$. Notice $\frac{1}{\sqrt{n}}X^T Z$ is a standard Gaussian variable. By the proof in

Lemma 2.4.4 and Proposition 2.4.1, we have the lower bound.

$$\sup_t \left| \mathbb{P} \left(\frac{\sum x_i y_i}{\sqrt{n}} \leq t \right) - \mathbb{P} \left(\frac{1}{\sqrt{n}} X^T Z \leq t \right) \right| \geq O \left(\mathbb{E} \left[1 \wedge \left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| \right] \right)$$

Now it is very likely $\mathbb{E} \left[1 \wedge \left| \sqrt{\frac{\sum y_k^2}{n}} - 1 \right| \right] = 1 + O \left(\frac{1}{\sqrt{m}} \right)$ since $\mathbb{E} \left[\left(\frac{\sum y_k^2}{n} - 1 \right)^2 \right] = O \left(\frac{1}{m} \right)$. Therefore a lower bound of $O \left(\frac{1}{\sqrt{m}} \right)$ is obtained.

On the other hand, it is not clear whether Equation 3.3 can be improved. In some cases, $O \left(\frac{1}{\sqrt{n}} \right)$ is not necessary. For example, if we let $m \rightarrow \infty, n = 1$, then

$$\frac{1}{\sqrt{m^2 1 + m 1^2}} X^T S^T S Z \approx \left(\frac{1}{m} \sum_{i=1}^m S_i^2 \right) X Z \rightarrow X Z$$

In the original Johnson-Lindenstrauss lemma, the number of vectors p can be arbitrary ($p \geq 2$) and the error has a factor $\log p$. So far, we only discussed the case $p = 2$ in chapter 3. It would be interesting to see if there is a $\log p$ factor for invariance of p random vectors.

Moreover, we only discussed invariance of independence for random projection. To stretch the understanding to another level, we need to characterize invariance of dependent random vectors. A special case one can consider is when $X = Z$, so that we will have a quadratic form $X^T S^T S X$, which will be addressed in next chapter.

CHAPTER 4

INVARIANCE OF OF RANDOM MATRIX FOR THE NORM

Suppose there are p independent random vectors X_1, \dots, X_p with i.i.d. entries, and a random matrix S . In this chapter, we will investigate how much the full structure is preserved in the projected or embedded space. As in the dot product form of Johnson Lindenstrauss lemma, we are interested in $2p^2$ quantities:

$$\langle X_i, X_j \rangle, \langle SX_i, SX_j \rangle \quad \forall 1 \leq i, j \leq p$$

In the previous chapter, we have addressed the cases $i \neq j$. So we have a good understanding of the off-diagonal terms. We will focus on the remaining diagonal terms. Namely, we will try to understand how random projection affects the distribution of the norm of a random vector X , and try to find relations between the norm and projected norm

$$\langle X, X \rangle, \langle SX, SX \rangle$$

Before getting into technicality, we will first obtain a loose concentration result. Such concentration properties allow one to analyze the problem from an error control perspective. In section 4.1, concentration result in our setting (random matrix and random vectors) will be obtained by carrying out a similar argument of proving Bernstein's inequality.

In section 4.2, we would go one step further to deal with the distribution directly and show the distribution of the norm is invariant under random projections. In particular, if $m/n \rightarrow 0$ we show the random projected norm converges to normal distribution after properly centered and scaled.

4.1 Concentration of projected or embedded norm for sub-Gaussian variables

The purpose of this section is to show the randomly projected norm is concentrated around both the original random norm and expectation of the norm. Concentration inequalities concern the tails of a random quantity deviates from its mean, which are very powerful tools in many applications ([23, 24, 25]). Johnson-Lindenstrauss Lemma 1.1.1 itself is a result of concentration inequality for sub-Gaussian random matrix over fixed vectors (see [24, 25]). Concentration properties of random quadratic forms involving either deterministic vectors or deterministic matrix have been well-studied in the literature (see [26, 27, 28, 25]). Most of the existing results control the tail probability of the distortion by the matrix or expected distortion quantitatively. We shall use similar techniques to prove the concentration of randomly projected norm of sub-Gaussian random vectors. First let us recall some properties of sub-Gaussian random variables.

Definition 4.1. *We say X is a sub-Gaussian random variable if there is $v > 0$ such that*

$$\mathbb{E} e^{\lambda X} \leq e^{\frac{\lambda^2 v}{2}}$$

Using Markov inequality, we can easily see sub-Gaussian random variable admits the tail probability

$$\mathbb{P}(|X| > t) \leq 2e^{-\frac{t^2}{2v}}, \quad \forall t \geq 0$$

For now let us assume sub-Gaussian random variable X is centered and standardized, namely $\mathbb{E} X = 0, v = 1$. It is not hard to verify sub-Gaussian tail property implies moments bounds (see section 2.3 of [24]).

$$\begin{aligned} \mathbb{E} X^{2q} &= \int_0^\infty \mathbb{P}(X^{2q} > s) ds \\ &= \int_0^\infty qt^{q-1} \mathbb{P}(X^2 > t) dt \end{aligned}$$

$$\begin{aligned} &\leq \int_0^\infty qt^{q-1}2e^{-t/2}dt \\ &\leq 2^{q+1}q! \end{aligned}$$

This will allow us to compute moment generating function of X^2 , and some useful bounds to be used later. Firstly,

$$\begin{aligned} \mathbb{E} e^{\lambda X^2} &= 1 + \sum_{q=1}^{\infty} \frac{\lambda^q \mathbb{E} X^{2q}}{q!} \\ &\leq 1 + \sum_{q=1}^{\infty} \lambda^q 2^{q+1} \\ &= \frac{1 + 2\lambda}{1 - 2\lambda}, \quad \forall \lambda < \frac{1}{2} \\ &\leq e^{5\lambda}, \quad \forall \lambda < \frac{1}{5} \end{aligned} \tag{4.1}$$

Secondly, let X' be an independent copy of X .

$$\mathbb{E} e^{\lambda(X^2-1)} \mathbb{E} e^{-\lambda(X'^2-1)} = \mathbb{E} e^{\lambda[(X^2-1)-(X'^2-1)]} = 1 + \sum_{q=1}^{\infty} \frac{\lambda^{2q} \mathbb{E}(X^2 - X'^2)^{2q}}{(2q)!}$$

Notice by Minkowski's inequality we have $\mathbb{E}(X^2 - X'^2)^{2q} \leq 2^{2q} \mathbb{E} X^{4q}$, and by Jensen's inequality $\mathbb{E} e^{-\lambda(X'^2-1)} \geq e^{-(\mathbb{E} X'^2-1)} = 1$. Therefore

$$\begin{aligned} \mathbb{E} e^{\lambda(X^2-1)} &\leq 1 + \sum_{q=1}^{\infty} \frac{\lambda^{2q} 2^{2q} 2^{2q+1} (2q)!}{(2q)!} \\ &\leq \frac{1 + 16\lambda^2}{1 - 16\lambda^2}, \quad \forall \lambda < \frac{1}{4} \\ &\leq e^{40\lambda^2}, \quad \forall \lambda < \frac{1}{5} \end{aligned} \tag{4.2}$$

From a high level, these properties (Equation 4.1, Equation 4.2) of X^2 are expected since it is a sub-exponential random variable. For a centered sub-Gaussian X with variance 1, it is easy to see X^2 has sub-exponential tail decay. Because $\mathbb{P}(X^2 > t) = \mathbb{P}(|X| >$

$\sqrt{t}) \leq 2e^{-\frac{t}{2}}$. The centered version $X^2 - 1$ has tails shifted by a constant 1 thus again admits sub-exponential decay. For extensive detailed discussions and proofs of the properties, we refer to [24, 25].

Theorem 6. *Given $X = [x_1, \dots, x_n]^T$. Let $A = \frac{1}{\sqrt{m}}S$ be a $m \times n$ random matrix. Let all random variables $x_i, S_{i,j}$ are independent identically distributed with mean zero and variance one. Suppose all random variables are sub-Gaussian, then we have*

1.

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \left| \|X\|^2 - n \right| > t\right) < 2 \exp\left\{-\min\left(\frac{t^2}{160}, \frac{t\sqrt{n}}{10}\right)\right\} \quad (4.3)$$

2.

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \left| \|AX\|^2 - \|X\|^2 \right| > \frac{\|X\|^2}{n} t\right) < 2 \exp\left\{-\min\left(\frac{t^2 C}{160}, \frac{t C \sqrt{n}}{10}\right)\right\} \quad (4.4)$$

where $C = \frac{m}{n}$.

Proof

1. Notice x_i are i.i.d. sub-Gaussian with variance $v = 1$, then $(x_i^2 - 1)$ are i.i.d. sub-exponential random variables. We can use Chernoff type argument (or apply Bernstein's concentration inequality directly) to calculate

$$\begin{aligned} \mathbb{P}\left(\frac{1}{\sqrt{n}} (\|X\|^2 - n) > t\right) &= \mathbb{P}\left(e^{\frac{\lambda}{\sqrt{n}}(\|X\|^2 - n)} > e^{\lambda t}\right) \\ &\leq e^{-\lambda t} \mathbb{E}\left[e^{\frac{\lambda}{\sqrt{n}}(\|X\|^2 - n)}\right] \\ &= e^{-\lambda t} \prod_{i=1}^n \left[\mathbb{E} e^{\frac{\lambda}{\sqrt{n}}(x_i^2 - 1)}\right] \end{aligned}$$

which holds for any $\lambda \geq 0$. We know by Equation 4.2, for any $\lambda/\sqrt{n} \leq \frac{1}{5}$, we have

$$\mathbb{E} e^{\frac{\lambda}{\sqrt{n}}(x_i^2 - 1)} \leq e^{40\lambda^2/n}$$

then we obtain

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}(\|X\|^2 - n) > t\right) \leq e^{-\lambda t + 40\lambda^2}, \quad \lambda < \frac{\sqrt{n}}{5}$$

Minimizing the right hand under the constraint $\lambda \leq \frac{\sqrt{n}}{5}$, we find optimal $\lambda^* = \min(\frac{\sqrt{n}}{5}, \frac{t}{80})$. Therefore we obtain

$$\begin{aligned} \mathbb{P}\left(\frac{1}{\sqrt{n}}(\|X\|^2 - n) > t\right) &\leq \begin{cases} \exp\{-\frac{t^2}{160}\}, & \text{if } t < 16\sqrt{n} \\ \exp\{\frac{8}{5}n - \frac{t\sqrt{n}}{5}\} \leq \exp\{-\frac{t\sqrt{n}}{10}\}, & \text{if } t \geq 16\sqrt{n} \end{cases} \\ &= \exp\left\{-\min\left(\frac{t^2}{160}, \frac{t\sqrt{n}}{10}\right)\right\} \end{aligned} \quad (4.5)$$

Repeat the same argument for $\frac{1}{\sqrt{n}}(n - \|X\|^2)$, we find the other half admits the same tail bound, thus we obtain Equation 4.4

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\| \|X\|^2 - n \| > t\right) < 2 \exp\left\{-\min\left(\frac{t^2}{160}, \frac{t\sqrt{n}}{10}\right)\right\}$$

We see as $n \rightarrow \infty$, the tail has a Gaussian behavior namely e^{-ct^2} which coincide with the CLT of $\frac{1}{\sqrt{n}}(n - \|X\|^2) \rightarrow \mathcal{N}(0, \mathbb{E}(x_1^2 - 1)^2)$.

2. Now we want to quantify how much $\|AX\|^2$ deviates from $\|X\|^2$.

$$\|AX\|^2 = \sum_{i=1}^m (A_{i,\cdot} X)^2 = \sum_{i=1}^m \left(\sum_{j=1}^n A_{i,j} x_j\right)^2 = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^n A_{i,j} A_{i,k} x_j x_k$$

We can use conditioning on X and only deal with the conditional probability.

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\| \|AX\|^2 - \|X\|^2 \| > t\right) = \mathbb{E}\left[\mathbb{P}\left(\frac{1}{\sqrt{n}}\| \|AX\|^2 - \|X\|^2 \| > t \mid X\right)\right] \quad (4.6)$$

In which case, we can think of X being fixed when computing conditional probabil-

ity. To simplify notation, define

$$y_i := \left[\left(\sum_{j=1}^n A_{i,j} x_j \right) \middle| X \right], \quad 1 \leq i \leq m$$

Therefore $\|AX\|^2|X = \sum_{i=1}^m y_i^2$. Notice $\mathbb{E} y_i = 0$ since $\mathbb{E} A_{i,j} = 0$. Moreover $A_{i,j}, A_{i,k}$ are independent if $j \neq k$, we find

$$\begin{aligned} \mathbb{E} y_i^2 &= \mathbb{E} \left[\left(\sum_{j=1}^n A_{i,j} x_j \right)^2 \middle| X \right] \\ &= \mathbb{E} \left[\sum_{j=1}^n \sum_{k=1}^n A_{i,j} A_{i,k} x_j x_k \middle| X \right] \\ &= \sum_{j=1}^n x_j^2 \mathbb{E} A_{i,j}^2 \\ &= \sum_{j=1}^n x_j^2 \frac{1}{m} \mathbb{E} S_{i,j}^2 \\ &= \frac{1}{m} \|X\|^2 \end{aligned}$$

And this also shows conditional expectation of projected norm is the original norm

$$\mathbb{E} [\|AX\|^2|X] = \sum_{i=1}^m \mathbb{E}[y_i^2|X] = \|X\|^2$$

We may rewrite the tail of norms (Equation 4.6) in terms of random variable y_i ,

$$\mathbb{P} \left(\frac{1}{\sqrt{n}} \left| \|AX\|^2 - \|X\|^2 \right| > t \right) = \mathbb{E} \left[\mathbb{P} \left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 - \mathbb{E} y_i^2) \right| > t \right) \right]$$

In fact, linear combination of sub-Gaussian is still sub-Gaussian. We shall prove y_i is sub-Gaussian random variable so that we can obtain Bernstein's type inequality

again by applying Equation 4.2.

$$\begin{aligned}
\mathbb{E}[e^{y_i t}] &= \mathbb{E} e^{t \sum_{j=1}^n A_{i,j} x_j} = \prod_{j=1}^n \mathbb{E} e^{x_j t A_{i,j}} = \prod_{j=1}^n \mathbb{E} e^{x_j t \frac{S_{i,j}}{\sqrt{m}}} \\
&\leq \prod_{j=1}^n e^{x_j^2 t^2 / 2m} \\
&= e^{(\sum_{j=1}^n x_j^2 t^2 / 2m)} \\
&= e^{(\|X\|^2 / m) \frac{t^2}{2}}
\end{aligned}$$

Then the tail probability

$$\begin{aligned}
\mathbb{P}(y_i > t) &= \mathbb{P}(e^{y_i \lambda} > e^{t \lambda}) \\
&\leq e^{-t \lambda} \mathbb{E} e^{y_i \lambda} \\
&\leq \exp\left(-t \lambda + (\|X\|^2 / m) \frac{\lambda^2}{2}\right) \quad \forall \lambda > 0
\end{aligned}$$

If we minimize on the right side over $\lambda > 0$, we should take $\lambda = tm / \|X\|^2$. Therefore we obtain

$$\mathbb{P}(y_i > t) \leq \exp\left(-\frac{t^2}{2(\|X\|^2 / m)}\right)$$

Repeat the same argument for $\mathbb{P}(y_i < -s)$, we will obtain two-sided sub-Gaussian tail bound. This shows y_i is sub-Gaussian with variance $v = \|X\|^2 / m$. Therefore $y_i^2 - \mathbb{E} y_i^2$ is sub-exponential. Then with Chernoff's method, we calculate

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 - \mathbb{E} y_i^2) \right| > t\right) &= \mathbb{P}\left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 / v - \mathbb{E} y_i^2 / v) \right| > t / v\right) \\
&\leq 2e^{-\lambda t / v} \prod_{i=1}^m \mathbb{E} e^{\frac{\lambda}{\sqrt{n}} (y_i^2 / v - \mathbb{E} y_i^2 / v)}
\end{aligned}$$

which holds for any $\lambda \geq 0$. Notice y_i / v are centered and standardized independent

sub-Gaussian random variables. By Equation 4.2, we know for any $\lambda \leq \frac{\sqrt{n}}{5}$, we have

$$\mathbb{E} e^{\frac{\lambda}{\sqrt{n}}(y_i^2 - \mathbb{E} y_i^2)} \leq e^{40\lambda^2/n}$$

then we obtain

$$\mathbb{P} \left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 - \mathbb{E} y_i^2) \right| > t \right) \leq 2 \exp \left\{ -\lambda t \frac{m}{\|X\|^2} + 40\lambda^2 \frac{m}{n} \right\} \quad \forall \lambda \leq \frac{\sqrt{n}}{5}$$

Optimize the right hand side under the constraint $\lambda \leq \frac{\sqrt{n}}{5}$, we find optimal $\lambda^* = \min(\frac{\sqrt{n}}{5}, \frac{tn}{80\|X\|^2})$. Therefore we obtain

$$\begin{aligned} \mathbb{P} \left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 - \mathbb{E} y_i^2) \right| > t \right) &\leq \begin{cases} 2 \exp\{-\frac{t^2 mn}{160(\|X\|^2)^2}\}, & \text{if } t < 16\|X\|^2/\sqrt{n} \\ 2 \exp\{\frac{8m}{5} - \frac{t\sqrt{n} m}{5\|X\|^2}\}, & \text{if } t \geq 16\|X\|^2/\sqrt{n} \end{cases} \\ &\leq \begin{cases} 2 \exp\{-\frac{t^2 Cn^2}{160(\|X\|^2)^2}\}, & \text{if } t < 16\|X\|^2/\sqrt{n} \\ 2 \exp\{-\frac{t\sqrt{n} Cn}{10\|X\|^2}\}, & \text{if } t \geq 16\|X\|^2/\sqrt{n} \end{cases} \end{aligned}$$

where $C := \frac{m}{n} \geq 0$ is a constant. Therefore

$$\mathbb{P} \left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 - \mathbb{E} y_i^2) \right| > t \right) \leq 2 \exp \left\{ -\min \left(\frac{t^2 Cn^2}{160(\|X\|^2)^2}, \frac{t\sqrt{n} Cn}{10\|X\|^2} \right) \right\}$$

Replacing t by $\|X\|^2 t/n$, we obtain

$$\mathbb{P} \left(\frac{1}{\sqrt{n}} \left| \sum_{i=1}^m (y_i^2 - \mathbb{E} y_i^2) \right| > \frac{\|X\|^2}{n} t \right) \leq 2 \exp \left\{ -\min \left(\frac{t^2 C}{160}, \frac{t\sqrt{n} C}{10} \right) \right\}$$

Taking expectation with respect to X we obtain Equation 4.4. ■

Remark. The first tail bound implies $\|X\|^2$ is close to n . The second tail bound implies

$\|AX\|^2$ is close to $\|X\|^2$ and thus also close to n . Later next section we will obtain a CLT type result for

$$\|AX\|^2 = \frac{1}{m} X^T S S X$$

This concentration result actually suggests the centered and rescaled projected norm has a tail that decays at a sub-Gaussian rate when $t \leq \sqrt{n}$ provided the random variables are originally sub-Gaussian. Thus a Gaussian limit (though without assuming random variables are sub-Gaussian) which we will prove in section 4.2 is not surprising.

4.2 Random projection preserves distribution of norm

Notice by central limit theorem, we have

$$\frac{\|X\|^2 - n}{\sqrt{n(\mathbb{E} x_1^4 - 1)}} \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$$

To understand if the random projected norm $\|SX\|^2$ has any type of convergence, it is necessary to find proper center and scale which are corresponding to first and second moments. Let us compute the mean first.

$$\begin{aligned} \mathbb{E} \|SX\|^2 - mn &= \mathbb{E} X^T S^T S X - mn \\ &= \mathbb{E} \operatorname{tr}(X X^T S^T S) - mn \\ &= \operatorname{tr}(\mathbb{E} X X^T \mathbb{E} S^T S) - mn \\ &= \operatorname{tr}(I_n m I_n) - mn \\ &= 0 \end{aligned}$$

For the variance,

$$\begin{aligned} &\mathbb{E}(\|SX\|^2 - mn)^2 \\ &= \mathbb{E}(X^T S^T S X)^2 - m^2 n^2 \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}(X^T S^T S X X^T S^T S X) - m^2 n^2 \\
&= \mathbb{E} \left(\sum x_{i_1} S_{i_2, i_1} S_{i_2, i_3} x_{i_3} x_{j_1} S_{j_2, j_1} S_{j_2, j_3} x_{j_3} \right) - m^2 n^2
\end{aligned}$$

The surviving terms must have even powers since first moments of the random variables are all 0. Therefore we only need to count four cases $\{i_1 = i_3 = j_1 = j_3 := i\}$, $\{i_1 = i_3 := i \neq j_1 = j_3 := j\}$, $\{i_1 = j_1 := i \neq i_3 = j_3 := j\}$, $\{i_1 = j_3 := i \neq i_3 = j_1 := j\}$.

$$\begin{aligned}
&\mathbb{E} \left(\sum x_{i_1} S_{i_2, i_1} S_{i_2, i_3} x_{i_3} x_{j_1} S_{j_2, j_1} S_{j_2, j_3} x_{j_3} \right) \\
&= \mathbb{E} \sum_{i_2, j_2=1}^m \sum_{i=1}^n x_i S_{i_2, i} S_{i_2, i} x_i x_i S_{j_2, i} S_{j_2, i} x_i + \mathbb{E} \sum_{i_2, j_2=1}^m \sum_{\substack{i, j=1 \\ i \neq j}}^n x_i S_{i_2, i} S_{i_2, i} x_i x_j S_{j_2, j} S_{j_2, j} x_j \\
&\quad + \mathbb{E} \sum_{i_2, j_2=1}^m \sum_{\substack{i, j=1 \\ i \neq j}}^n x_i S_{i_2, i} S_{i_2, j} x_j x_i S_{j_2, i} S_{j_2, j} x_j + \mathbb{E} \sum_{i_2, j_2=1}^m \sum_{\substack{i, j=1 \\ i \neq j}}^n x_i S_{i_2, i} S_{i_2, j} x_j x_j S_{j_2, j} S_{j_2, i} x_i \\
&= \mathbb{E} \sum_{i_2, j_2=1}^m \sum_{i=1}^n x_i^4 S_{i_2, i}^2 S_{j_2, i}^2 + \mathbb{E} \sum_{i_2, j_2=1}^m \sum_{\substack{i, j=1 \\ i \neq j}}^n x_i^2 x_j^2 S_{i_2, i}^2 S_{j_2, j}^2 \\
&\quad + \mathbb{E} \sum_{k=1}^m \sum_{\substack{i, j=1 \\ i \neq j}}^n x_i^2 x_j^2 S_{k, i}^2 S_{k, j}^2 + \mathbb{E} \sum_{k=1}^m \sum_{\substack{i, j=1 \\ i \neq j}}^n x_i^2 x_j^2 S_{k, i}^2 S_{k, j}^2
\end{aligned}$$

where the last two terms dropped the zero terms $i_2 \neq j_2$,

$$\begin{aligned}
&= (m^2 n - mn + mn \mathbb{E} S_{11}^4) \mathbb{E} x_1^4 + m^2 (n^2 - n) + 2m(n^2 - n) \\
&= m^2 n^2 + (\mathbb{E} x_1^4 - 1) m^2 n + 2mn^2 + mn[(\mathbb{E} S_{11}^4 - 1) \mathbb{E} x_1^4 - 2] \\
&= m^2 n^2 + \sigma^2 m^2 n + 2mn^2 + \xi mn
\end{aligned}$$

where $\sigma^2 := \mathbb{E} S_{11}^4 - 1$, $\xi := [(\mathbb{E} S_{11}^4 - 1) \mathbb{E} x_1^4 - 2]$. Therefore

$$\mathbb{E} \left[\left(\frac{X^T S^T S X - mn}{\sqrt{\sigma^2 m^2 n + 2mn^2 + \xi mn}} \right)^2 \right] = 1 \tag{4.7}$$

Next we will show the centered and scaled projected norm actually also converges to a

normal. That means distribution of the norm of a vector (with independent entries) is also invariant under random projection.

4.2.1 CLT for Random projection of norm

Theorem 7. *Given a random vector X in \mathbb{R}^n with i.i.d. entries*

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Let $\mathbb{E} x_1 = 0, \mathbb{E} x_1^2 = 1, \mathbb{E} x_1^4 = 1 + \sigma^2 (0 \leq \sigma < \infty)$. Consider a random matrix $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with independent identically distributed entries $S_{i,j}$ with $\mathbb{E} S_{i,j} = 0, \mathbb{E} S_{i,j}^2 = 1$ and $\mathbb{E} S_{1,1}^4 < c < \infty$. Further assume S, X are all independent. Define

$$A(m, n) := \frac{\|SX\|^2 - mn}{\sqrt{\sigma^2 m^2 n + 2mn^2 + \xi mn}}$$

where $\xi = [(\mathbb{E} S_{1,1}^4 - 1) \mathbb{E} x_1^4 - 2]$. If $\frac{m}{n} \rightarrow 0$, then

$$A(m, n) \xrightarrow{m, n \rightarrow \infty} \mathcal{N}(0, 1) \tag{4.8}$$

Remark. *Before we proceed with the proof, it is worth mentioning that the random norm is a complicated sum of mn^2 correlated terms.*

$$X^T S^T S X = \sum_{k=1}^m \sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j$$

Therefore most analytical methods and tools fail to treat the quantity properly. For example, characteristic function need independent property, Lindeberg swapping needs martingale property, and Stein's method needs exchangeable structure and precise control of first and second moments of conditional perturbed differences. So we are constrained to use a robust

and universal approach, the moment method, which can be used to prove convergence to a limit law with known moments.

In the moment method we present below, we need to control the order of m by $m \leq o(n)$ because the counting procedure would be impossible to carry out if this is not the case. The scaling in this case is dominated by $\sqrt{2mn^2}$ which allows us to limit the significant terms in the moment calculation. However in simulations, we will see convergence to normal even when $m > n$. But we could not find a proof for the general case yet due to too many correlated terms.

Proof We will first note that a truncation argument will show it is sufficient to prove the same CLT result for bounded random variables. Details can be found in many standard moment method proof for CLT in many standard textbook (see for example [26] 2.2).

From now on, we assume all random variables are bounded, so that they have finite moments of all order which is very important in moment method. We will compute all moments of $A(m, n)$ in the limit and we expect all odd moments vanish and all even moments match with standard normal random variable.

The key idea is to separate the random norm into m identically distributed but dependent random variables. Let $S_{k\cdot}$ be k -th row of S . Then $\|SX\|^2 = \sum_{k=1}^m (S_{k\cdot}, X)^2$. Define

$$\begin{aligned} L_k &:= \frac{(S_{k\cdot}, X)^2 - n}{\sqrt{\sigma^2 mn + 2n^2 + \xi n}} = \frac{(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j - n)}{\sqrt{\sigma^2 mn + 2n^2 + \xi n}} \\ \implies A(m, n) &= \frac{1}{\sqrt{m}} \sum_{k=1}^m L_k \\ \implies \lim_{m,n \rightarrow \infty} \mathbb{E} A(m, n)^t &= \lim_{m,n \rightarrow \infty} \frac{1}{m^{t/2}} \mathbb{E} \left(\sum_{k=1}^m L_k \right)^t \end{aligned}$$

Let us first record some moments properties of these identically distributed L_k .

(1) $\mathbb{E} L_k = 0$.

(2) $\mathbb{E} L_k^2 = 1 - O(\frac{m}{n})$.

(3) For any fixed $t \in \mathbb{R}$, there is a constant C_t independent of m and n so that

$$|\mathbb{E} L_k^t| \leq C_t < \infty \quad (4.9)$$

(4) The order of the expectation of a polynomial is determined by the number of singletons. Given integer $q_1, \dots, q_r \geq 0$,

$$\mathbb{E}[L_1^{q_1} \dots L_r^{q_r}] \leq O\left(\frac{1}{(\sigma^2 m + 2n)^{d/2}}\right), \quad \text{where } d = \sum_{i=1}^r 1_{(q_i=1)} \quad (4.10)$$

Here d is the total number of variables L_i with multiplicity 1.

(5) For any fixed $r \in \mathbb{R}$,

$$\mathbb{E}(L_1^2 \dots L_r^2) = \left(\frac{2n}{\sigma^2 m + 2n}\right)^r + O\left(\frac{1}{n}\right)$$

which converges to 1 if $\frac{m}{n} \rightarrow 0$.

We will prove one by one.

(1) Obviously, $\mathbb{E} L_k = 0$.

(2) The variance $\mathbb{E} L_k^2 = 1 - O(\frac{mn}{\sigma^2 mn + 2n^2 + \xi n})$ since

$$\begin{aligned} \mathbb{E}\left(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j - n\right)^2 &= \mathbb{E}\left(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j\right)^2 - n^2 \\ &= \mathbb{E}\left(\sum_{i=1}^n x_i^2 S_{k,i}^2\right)^2 + 2 \mathbb{E}\left[\left(\sum_{i=1}^n x_i^2 S_{k,i}^2\right)\left(\sum_{\substack{i,j=1 \\ i \neq j}}^n x_i S_{k,i} S_{k,j} x_j\right)\right] \\ &\quad + \mathbb{E}\left(\sum_{\substack{i,j=1 \\ i \neq j}}^n x_i S_{k,i} S_{k,j} x_j\right)^2 - n^2 \end{aligned}$$

$$\begin{aligned}
&= [n \mathbb{E} x_1^4 S_{1,1}^4 + n(n-1)] + 0 + 2n(n-1) - n^2 \\
&= 2n^2 + O(n)
\end{aligned}$$

Therefore $\mathbb{E} L_k^2 = 1 - O(\frac{1}{n})$ when $m = o(n)$.

(3) It is also true that L_k has finite moments of all order which also hold in the limit. We will use a careful counting procedure. First we notice

$$\begin{aligned}
|\mathbb{E} L_k^t| &= \left| \mathbb{E} \left[\frac{(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j - n)}{\sqrt{\sigma^2 m n + 2n^2 + \xi n}} \right]^t \right| \\
&\leq n^{-t} \left| \mathbb{E} \left[\left(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j - n \right)^t \right] \right| \tag{4.11}
\end{aligned}$$

Since $n = \sum_{i,j} \delta_{i,j}$ where $\delta_{i,j} = 1$ when $i = j$ and 0 otherwise. Then we expand the t -th moment on the right as polynomials.

$$\begin{aligned}
&\mathbb{E} \left[\left(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j - n \right)^t \right] \\
&= \mathbb{E} \left[\left(\sum_{i_1, j_1=1}^n (x_{i_1} S_{k,i_1} S_{k,j_1} x_{j_1} - \delta_{i_1, j_1}) \right) \cdots \left(\sum_{i_t, j_t=1}^n (x_{i_t} S_{k,i_t} S_{k,j_t} x_{j_t} - \delta_{i_t, j_t}) \right) \right] \\
&= \sum_{i_1, j_1=1}^n \cdots \sum_{i_t, j_t=1}^n \mathbb{E} [(x_{i_1} S_{k,i_1} S_{k,j_1} x_{j_1} - \delta_{i_1, j_1}) \cdots (x_{i_t} S_{k,i_t} S_{k,j_t} x_{j_t} - \delta_{i_t, j_t})]
\end{aligned}$$

Notice $\mathbb{E}[(x_{i_1} S_{k,i_1} S_{k,j_1} x_{j_1} - \delta_{i_1, j_1}) \cdots (x_{i_t} S_{k,i_t} S_{k,j_t} x_{j_t} - \delta_{i_t, j_t})]$ vanishes if cardinality of $\{i_1, j_1, \dots, i_t, j_t\}$ is greater than t since that will have at least one singleton factor $\mathbb{E} S_{k,i} = 0$. So the index set must collapse to a set of at most t distinct indices. This means total number of nonzero terms is exactly n^t . As we are assume all variables are truncated to have bounded moments, we find

$$\left| \mathbb{E} \left[\left(\sum_{i,j=1}^n x_i S_{k,i} S_{k,j} x_j - n \right)^t \right] \right| \leq O(n^t)$$

Therefore plugging into Equation 4.11 we find $|\mathbb{E} L_k^t| \leq O(1)$. This proves Equation 4.9.

- (4) Equation 4.10 is an important property concerns the product of L_k 's. The order of the expectation of a polynomial $\mathbb{E}[L_1^{q_1} \cdots L_r^{q_r}]$ is determined by the number of singletons $d = \sum_{i=1}^r 1_{(q_i=1)}$. More precisely, for any product of L_k involving d term of power 1, then the expected value is of order $(m+n)^{-d/2}$. In other words, each power 1 term contribute a factor of $(m+n)^{-1/2}$. we use an argument by conditioning and careful counting. First, noticing L_k conditioning on X are independent,

$$\begin{aligned}
\mathbb{E}[L_1^{q_1} L_2^{q_2} \cdots L_r^{q_r}] &= \mathbb{E}[(L_2^{q_2} \cdots L_r^{q_r}) \mathbb{E}(L_1^{q_1} | X, S_{2,..}, \cdots S_{r,..})] \\
&= \mathbb{E}[(L_2^{q_2} \cdots L_r^{q_r}) \mathbb{E}(L_1^{q_1} | X)] \\
&= \mathbb{E}[\mathbb{E}((L_2^{q_2} \cdots L_r^{q_r}) | X, S_{3,..}, \cdots S_{r,..}) \mathbb{E}(L_1^{q_1} | X)] \\
&= \mathbb{E}[(L_3^{q_3} \cdots L_r^{q_r}) \mathbb{E}(L_2^{q_2} | X) \mathbb{E}(L_1^{q_1} | X)] \\
&\dots \\
&= \mathbb{E}[\mathbb{E}(L_1^{q_1} | X) \mathbb{E}(L_2^{q_2} | X) \cdots \mathbb{E}(L_r^{q_r} | X)]
\end{aligned}$$

Without loss of generality, assume the first d variables are of multiplicity 1, namely

$$q_1 = q_2 = \cdots = q_d = 1$$

To simplify notation, we denote $\mathbb{E}(L_i | X) := \mu_i, 1 \leq i \leq d$. We would also only need the above conditioning argument up to d -th variable L_d . That is

$$\mathbb{E}[L_1 \cdots L_d L_{d+1}^{q_{d+1}} \cdots L_r^{q_r}] = \mathbb{E}[\mu_1 \cdots \mu_d L_{d+1}^{q_{d+1}} \cdots L_r^{q_r}]$$

Then we apply Cauchy-Schwarz inequality, we find

$$\mathbb{E}[L_1 \cdots L_d L_{d+1}^{q_{d+1}} \cdots L_r^{q_r}] \leq (\mathbb{E}[\mu_1^2 \cdots \mu_d^2])^{\frac{1}{2}} \cdot (\mathbb{E}[L_{d+1}^{2q_{d+1}} \cdots L_r^{2q_r}])^{\frac{1}{2}} \quad (4.12)$$

Then we notice the random variables μ_1, \dots, μ_d (conditional expectation is also a random variable) are identical random variables (not just identically distributed).

$$\begin{aligned} \mu_1 = \mathbb{E}(L_1|X) &= \mathbb{E}\left(\frac{(\sum_{i,j=1}^n x_i S_{1,i} S_{1,j} x_j - n)}{\sqrt{\sigma^2 mn + 2n^2 + \xi n}} \middle| X\right) \\ &= \frac{1}{\sqrt{\sigma^2 mn + 2n^2 + \xi n}} \left(\sum_{i=1}^n x_i^2 \mathbb{E} S_{1,i}^2 + \sum_{\substack{i,j=1 \\ i \neq j}}^n x_i x_j \mathbb{E}(S_{1,i} S_{1,j}) - n \right) \\ &= \frac{\sum_{i=1}^n x_i^2 - n}{\sqrt{\sigma^2 mn + 2n^2 + \xi n}} \xrightarrow{a.s.} 0 \end{aligned}$$

This shows μ_1 does not depend on the index 1 and indeed μ_1, \dots, μ_d are identical. As a side note, μ_1 converges to 0 almost surely due to the strong law of large number. We would not need this fact though. Now we can analyze first term on the right hand side of Equation 4.12,

$$\begin{aligned} \mathbb{E}[\mu_1^2 \cdots \mu_d^2] &= \mathbb{E}[\mu_1^{2d}] \\ &= \frac{\mathbb{E}(\sum_{i=1}^n x_i^2 - n)^{2d}}{(\sigma^2 mn + 2n^2 + \xi n)^d} \\ &= \frac{1}{(\sigma^2 m + 2n + \xi)^d} \mathbb{E}\left(\frac{\sum_{i=1}^n x_i^2 - n}{\sqrt{n}}\right)^{2d} \\ &= O\left(\frac{1}{(\sigma^2 m + 2n)^d}\right) \end{aligned} \quad (4.13)$$

The last step, we used the the fact

$$\mathbb{E}\left(\frac{\sum_{i=1}^n x_i^2 - n}{\sqrt{n}}\right)^t \leq C_t < \infty$$

that is due to independent sums $(\sum_{i=1}^n x_i^2 - n)$ in CLT has t -th moments of $O(n^{t/2})$ if x_i has finite moments of all order (see [29, 30]), which holds true due to its boundedness by truncation argument.

On the other hand, applying Cauchy-Schwarz inequality repeatedly for the second term on the right hand side of Equation 4.12, we can bound it by a constant c that does not depend on m and n . Namely using the fact L_k has bounded moments of all order (Equation 4.9)

$$\begin{aligned} \mathbb{E}[L_{d+1}^{2q_{d+1}} \cdots L_r^{2q_r}] &\leq \left[\mathbb{E} L_{d+1}^{4q_{d+1}} \mathbb{E} \left[L_{d+2}^{4q_{d+2}} \cdots L_r^{4q_r} \right] \right]^{1/2} \\ &\leq \left[\mathbb{E} L_{d+1}^{4q_{d+1}} \right]^{\frac{1}{2}} \left[\mathbb{E} L_{d+2}^{8q_{d+2}} \right]^{\frac{1}{4}} \cdots \left[\mathbb{E} L_r^{2^{r-d+1}q_r} \right]^{\frac{1}{2^{r-d}}} \\ &\leq c < \infty \end{aligned} \tag{4.14}$$

Therefore Equation 4.14 combined with Equation 4.12 and Equation 4.13, we obtain our desired result Equation 4.10.

(5) To compute $\mathbb{E}(L_1^2 \cdots L_r^2)$ we will use conditioning argument again.

$$\mathbb{E}(L_1^2 \cdots L_r^2) = \mathbb{E} \left[\mathbb{E}(L_1^2 | X) \cdots \mathbb{E}(L_r^2 | X) \right]$$

Same as before, $\mathbb{E}(L_1^2 | X), \dots, \mathbb{E}(L_r^2 | X)$ does not depend on the indices $1, \dots, r$, and they are all identical random variables not just with same distribution. By definition,

$$\mathbb{E}(L_1^2 | X) = \frac{\mathbb{E} \left(\left(\sum_{i,j=1}^n x_i S_{1,i} S_{1,j} x_j - n \right)^2 \middle| X \right)}{\sigma^2 mn + 2n^2 + \xi n}$$

We shall simplify the numerator (denoted as Q),

$$Q := \mathbb{E} \left(\left(\sum_{i,j=1}^n x_i S_{1,i} S_{1,j} x_j - n \right)^2 \middle| X \right)$$

$$= \mathbb{E} \left(\left(\left(\sum_{i=1}^n (x_i^2 S_{1,i}^2 - 1) \right) + \sum_{\substack{i,j=1 \\ i \neq j}}^n x_i x_j S_{1,i} S_{1,j} \right)^2 \middle| X \right)$$

Expand the quadratic we find the cross terms vanish

$$\begin{aligned} & 2 \mathbb{E} \left(\left(\sum_{k=1}^n (x_k^2 S_{1,k}^2 - 1) \right) \left(\sum_{\substack{i,j=1 \\ i \neq j}}^n x_i x_j S_{1,i} S_{1,j} \right) \middle| X \right) \\ &= 2 \sum_{k=1}^n \sum_{\substack{i,j=1 \\ i \neq j}}^n \mathbb{E} [(x_k^2 S_{1,k}^2 - 1) x_i x_j S_{1,i} S_{1,j} | X] \\ &= 2 \sum_{k=1}^n \sum_{\substack{i,j=1 \\ i \neq j}}^n (x_k^2 x_i x_j \mathbb{E}[S_{1,k}^2 S_{1,i} S_{1,j}] - x_i x_j \mathbb{E}[S_{1,i} S_{1,j}]) \end{aligned}$$

since each term $\mathbb{E}[S_{1,k}^2 S_{1,i} S_{1,j}] = \mathbb{E}[S_{1,i} S_{1,j}] = 0$ as $i \neq j$. So we are left with

$$\begin{aligned} Q &= \mathbb{E} \left(\left(\sum_{i=1}^n x_i^2 S_{1,i}^2 - n \right)^2 \middle| X \right) + \mathbb{E} \left(\left(\sum_{\substack{i,j=1 \\ i \neq j}}^n x_i x_j S_{1,i} S_{1,j} \right)^2 \middle| X \right) \\ &= \left(\sum_{i=1}^n \sum_{j=1}^n x_i^2 x_j^2 \mathbb{E} S_{1,i}^2 S_{1,j}^2 - 2n \sum_{i=1}^n x_i^2 \mathbb{E} S_{1,i}^2 + n^2 \right) \\ &\quad + \left(\sum_{\substack{i,j=1 \\ i \neq j}}^n \sum_{\substack{i',j'=1 \\ i' \neq j'}}^n x_i x_j x_{i'} x_{j'} \mathbb{E} S_{1,i} S_{1,j} S_{1,i'} S_{1,j'} \right) \end{aligned}$$

Notice the surviving terms in the second half are $\{i = i' \neq j = j'\}$ and $\{i = j' \neq j = i'\}$. Therefore,

$$Q = \left(\sum_{\substack{i,j=1 \\ i \neq j}}^n x_i^2 x_j^2 + \sum_{i=1}^n x_i^4 \mathbb{E} S_{1,1}^4 - 2n \sum_{i=1}^n x_i^2 + n^2 \right) + \left(2 \sum_{\substack{i,j=1 \\ i \neq j}}^n x_i^2 x_j^2 \right)$$

$$= 3 \sum_{\substack{i,j=1 \\ i \neq j}}^n x_i^2 x_j^2 - 2n \sum_{i=1}^n x_i^2 + n^2 + \sum_{i=1}^n x_i^4 \mathbb{E} S_{1,1}^4$$

Since $\mathbb{E}(L_k^2|X) = Q/(\sigma^2 mn + 2n^2 + \xi n)$ for all k , we simplifies

$$\mathbb{E}(L_1^2 \cdots L_r^2) = \mathbb{E}[Q^r] / (\sigma^2 mn + 2n^2 + \xi n)^r$$

To prove $\mathbb{E}(L_1^2 \cdots L_r^2) \rightarrow 1$ it suffices to prove $\mathbb{E}[Q/(2n^2)]^r \rightarrow 1$ since $(\sigma^2 mn + 2n^2 + \xi n)^r$ is dominated by $(2n^2)^r$ as $m/n \rightarrow 0$.

$$\begin{aligned} \frac{Q}{2n^2} &= \frac{1}{2n^2} \left(3 \sum_{\substack{i,j=1 \\ i \neq j}}^n x_i^2 x_j^2 - 2n \sum_{i=1}^n x_i^2 + n^2 + \sum_{i=1}^n x_i^4 \mathbb{E} S_{1,1}^4 \right) \\ &= \frac{1}{2} P + \frac{1}{2} + P_0 \end{aligned}$$

where (to simplify notation) we denoted

$$P := \left(\frac{1}{n^2} \sum_{i=1}^n x_i^2 \left(\sum_{j=1, j \neq i}^n 3x_j^2 - 2n \right) \right), \quad P_0 = \frac{1}{2n^2} \sum_{i=1}^n x_i^4 \mathbb{E} S_{1,1}^4$$

First of all it is fairly easy to see $\mathbb{E} \left(\frac{Q}{2n^2} \right)^r < \infty$ for all $r \in \mathbb{R}$. This is because the total number of nonzero polynomial terms $(x_i^2, x_i^4, x_i^2 x_j^2)$ in Q is $O(n^2)$ which will produce $O(n^{2r})$ polynomial terms for Q^r , and x_i has finite moment of all order. Now let us use induction to prove the moments are actually constant 1 in the limit. Suppose we have proved for all $k \leq r$,

$$\mathbb{E} \left(\frac{Q}{2n^2} \right)^k \rightarrow 1, \quad \text{and} \quad \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^{k-1} P \right] \rightarrow 1$$

It is easy to see this holds for the initial steps. Then we try to prove the next induction step namely

$$\mathbb{E} \left(\frac{Q}{2n^2} \right)^{r+1} \rightarrow 1, \quad \text{and} \quad \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r P \right] \rightarrow 1$$

Notice by linearity of expectation and Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r P_0 \right] &= \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r \left(\frac{1}{2n^2} \sum_{i=1}^n x_i^4 \mathbb{E} S_{1,1}^4 \right) \right] \\ &= \frac{1}{2n} \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r x_1^4 \mathbb{E} S_{1,1}^4 \right] \\ &\leq \frac{1}{2n} \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^{2r} \right]^{1/2} \mathbb{E} [x_1^8 \mathbb{E} S_{1,1}^8]^{1/2} \rightarrow 0\end{aligned}$$

We find

$$\begin{aligned}\mathbb{E} \left(\frac{Q}{2n^2} \right)^{r+1} &= \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r \left[\frac{1}{2} P + \frac{1}{2} + P_0 \right] \right] \\ &\rightarrow \frac{1}{2} \lim_{n \rightarrow \infty} \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r P \right] + \frac{1}{2} + 0\end{aligned}$$

Now it suffices to show $\mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r P \right] \rightarrow 1$. We expand again

$$\mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^r P \right] = \mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^{r-1} \left(\frac{1}{2} P^2 + \frac{1}{2} P + P P_0 \right) \right]$$

By induction hypothesis $\mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^{r-1} \frac{1}{2} P \right] = \frac{1}{2}$. And by Cauchy-Schwarz we know $\mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^{r-1} P P_0 \right] \leq O\left(\frac{1}{2n}\right) \rightarrow 0$. So it suffices to show $\mathbb{E} \left[\left(\frac{Q}{2n^2} \right)^{r-1} P^2 \right] \rightarrow 1$.

Repeat this argument r times, we find it suffices to prove $\mathbb{E} P^{r+1} \rightarrow 1$. And since this has to be true for every induction step, we indeed need to show

$$\mathbb{E} P^k \rightarrow 1, \quad \forall k \in \mathbb{R}$$

Now we calculate

$$\mathbb{E} P^k = \mathbb{E} \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n x_i^2 (3x_j^2 - \frac{2n}{n-1}) \right)^k$$

$$= \frac{1}{n^{2k}} \sum_{i_1, \dots, i_k=1}^n \sum_{j_1, \dots, j_k=1, j_t \neq i_t}^n \mathbb{E} \left(\prod_{t=1}^k x_{i_t}^2 \left(3x_{j_t}^2 - \frac{2n}{n-1} \right) \right)$$

The summations produce total $[n(n-1)]^k$ terms. Among them, there are total

$$n(n-1) \cdots (n-2k+1) = \prod_{s=0}^{2k-1} (n-s) = n^{2k} - O(n^{2k-1})$$

terms that all indices are distinct, that is the cardinality $|\{i_1, \dots, i_k, j_1, \dots, j_k\}| = 2k$. So that we can evaluate the expectation directly by independence. For $i_t \neq j_t, i_t \neq i'_t, j_t \neq j'_t$

$$\begin{aligned} \mathbb{E} \left(\prod_{t=1}^k x_{i_t}^2 \left(3x_{j_t}^2 - \frac{2n}{n-1} \right) \right) &= \prod_{t=1}^k \mathbb{E} x_{i_t}^2 \mathbb{E} \left(3x_{j_t}^2 - \frac{2n}{n-1} \right) \\ &= \left(1 - \frac{2}{n-1} \right)^k = 1 - O\left(\frac{1}{n}\right) \rightarrow 1 \end{aligned}$$

For the remaining $n^k(n-1)^k - \prod_{s=0}^{2k} (n-s) = O(n^{2k-1})$ terms, the expectations of correlated variables are still bounded by some constant c_k . So we find

$$\begin{aligned} \mathbb{E} P^k &= \frac{1}{n^{2k}} \prod_{s=0}^{2k} (n-s) + \frac{1}{n^{2k}} c_k \left(n^k(n-1)^k - \prod_{s=0}^{2k} (n-s) \right) \\ &= 1 - O\left(\frac{1}{n}\right) + \frac{1}{n^{2k}} c_k O(n^{2k-1}) \\ &= 1 + O\left(\frac{1}{n}\right) \rightarrow 1 \end{aligned}$$

After establishing these properties, we can start the standard moments method for a CLT.

By our definition of L_k ,

$$\lim_{m, n \rightarrow \infty} \mathbb{E} A(m, n)^t = \lim_{m, n \rightarrow \infty} \frac{1}{m^{t/2}} \mathbb{E} \left(\sum_{k=1}^m L_k \right)^t$$

We first expand $\mathbb{E} \left(\sum_{k=1}^m L_k \right)^t$. Any term will have form $L_{p_1}^{q_1} \cdots L_{p_r}^{q_r}$ with the number of distinct indices $r : 1 \leq r \leq t$ and positive integer powers satisfy $q_1 + \cdots + q_r = t$. If we group the terms by total number of distinct indices

$$\mathbb{E} \left(\sum_{k=1}^m L_k \right)^t = \sum_{r=1}^t \sum_{\substack{q_1 + \cdots + q_r = t \\ \{p_1, \dots, p_r\} \subset \{1, \dots, m\}}} c_{t, q_1, \dots, q_r} \mathbb{E}(L_{p_1}^{q_1} \cdots L_{p_r}^{q_r})$$

where c_{t, q_1, \dots, q_r} is the total number of orderings when we order $\{q_1$ number of index p_1, \dots, q_r number of index $p_r, q_1 + \cdots + q_r = t\}$ all together, which is

$$c_{t, q_1, \dots, q_r} = \frac{t!}{q_1! \cdots q_r!}$$

This constant only depend on t and q_1, \dots, q_r , and it may be upper bounded by t^t .

Now we are going to analyze how much the terms contribute for each fixed r . In particular, we will show the only significant terms which will survive after scaling are when $r = \frac{t}{2}$ (if t is odd, that means no surviving terms).

For any term $L_1^{q_1} \cdots L_r^{q_r}$ with $r > \frac{t}{2}$, there are at least

$$2 \binom{r - \frac{t}{2}}{1} = 2r - t \tag{4.15}$$

variables L_i have multiplicity 1. It is true because increasing the length r by 1 will create at least two singleton terms. For example if we want to increase length of $L_1^2 L_2^2$ from 2 to 3, we would end up breaking a square term into two singletons so that we have $L_1^2 L_2 L_3$ or $L_1 L_2^2 L_3$. Formally, suppose that's not the case. Namely suppose there are only $s \leq 2r - t - 1$ variables L_i of multiplicity 1. Then adding all the multiplicity we get

$$t = q_1 + \cdots + q_r \geq s + 2(r - s) = 2r - s \geq t + 1$$

This is a contradiction.

Combining Equation 4.15 and Equation 4.10, each of the term when $r > \frac{t}{2}$ contribute at most

$$O\left(\frac{1}{(\sigma^2 m + 2n)^{d/2}}\right), \quad \text{where } d \geq 2r - t$$

For each fixed r , the total number of possible choices of $\{p_1, \dots, p_r\} \subset \{1, \dots, m\}$ is $\binom{m}{r} \leq O(m^r)$. Then we also need to count total number of ways to generate $q_1 + \dots + q_r = t$. That is we are looking at separating the integer t into r nonzero integers q_1, \dots, q_r . Total number will be $\binom{t-1}{r-1} \leq O(t^r)$ which only depend on t (we can model it as separating t stones into r piles. $\binom{t-1}{r-1}$ is due to the fact we can select $r - 1$ separating positions out of $t - 1$ spaces.).

Therefore the total contribution for each fixed $r > \frac{t}{2}$ is

$$\begin{aligned} & \frac{1}{m^{t/2}} \sum_{\substack{q_1 + \dots + q_r = t \\ \{p_1, \dots, p_r\} \subset \{1, \dots, m\}}} c_{t, q_1, \dots, q_r} \mathbb{E}(L_{p_1}^{q_1} \dots L_{p_r}^{q_r}) \\ & \leq m^{-\frac{t}{2}} O(m^r) O(t^r) c_{t, q_1, \dots, q_r} O\left(\frac{1}{(\sigma^2 m + 2n)^{d/2}}\right) \\ & \leq O\left(m^{r-\frac{t}{2}} (\sigma^2 m + 2n)^{-d/2}\right) \\ & \leq O\left(m^{r-\frac{t}{2}} (\sigma^2 m + 2n)^{-r+\frac{t}{2}}\right) \end{aligned}$$

Last step we used the fact $d \geq 2r - t$. Since we assumed $\frac{m}{n} \rightarrow 0$, we find the total contribution is bounded by

$$O\left(\frac{m}{\sigma^2 m + 2n}\right)^{r-\frac{t}{2}} \rightarrow 0, \quad \forall r > \frac{t}{2}$$

Therefore we can safely drop all cases of $r > \frac{t}{2}$.

For all cases $r < \frac{t}{2}$, each term $\mathbb{E}(L_{p_1}^{q_1} \dots L_{p_r}^{q_r}) = O(1)$ depend only on t by a repeated Cauchy-Schwarz argument similar to Equation 4.14. We find total contribution is

$$m^{-\frac{t}{2}} O(m^r) O(t^r) c_{t, q_1, \dots, q_r} O(1) \rightarrow 0, \quad \forall r < \frac{t}{2}$$

This implies for t odd

$$\lim_{m,n \rightarrow \infty} \frac{1}{m^{t/2}} \mathbb{E} \left(\sum_{k=1}^m L_k \right)^t \rightarrow 0$$

Now for even moments, the above analysis shows we only need to count $r = \frac{t}{2}$ (since contributions from $r < \frac{t}{2}$ and $r > \frac{t}{2}$ are both negligible). In this case, we are looking at separate the integer t into $r = \frac{t}{2}$ positive integers q_1, \dots, q_r . The total number will be $\binom{t-1}{r-1}$ which only depends on t ($\binom{t-1}{r-1}$ is due to the fact we can select $r - 1$ separating positions out of $t - 1$ spaces). There are still many terms not significant. By Equation 4.10, any term has a L_k of multiplicity 1 will contribute at most $O(n^{-\frac{1}{2}})$, thus we may drop these terms. Among these $\binom{t-1}{r-1}$ terms, there is only one term that every L_k has multiplicity at least 2, which will survive the scaling namely

$$q_1 = \dots = q_r = 2$$

In other words,

$$\begin{aligned} \lim_{m,n \rightarrow \infty} \frac{1}{m^{t/2}} \mathbb{E} \left(\sum_{k=1}^m L_k \right)^t &= \lim_{m,n \rightarrow \infty} \frac{1}{m^{t/2}} \sum_{\substack{r=t/2 \\ \{p_1, \dots, p_r\} \subset \{1, \dots, m\}}} c_{t,2,\dots,2} \mathbb{E}(L_{p_1}^2 \dots L_{p_r}^2) \\ &= \lim_{m,n \rightarrow \infty} \frac{1}{m^{t/2}} \binom{m}{t/2} c_{t,2,\dots,2} \mathbb{E}(L_1^2 \dots L_{t/2}^2) \\ &= \lim_{m \rightarrow \infty} \frac{1}{m^{t/2}} \binom{m}{t/2} \frac{t!}{2^{t/2}} \\ &= \frac{t!}{2^{t/2}} \end{aligned}$$

Which is exactly the even moments of standard normal random variable $\mathcal{N}(0, 1)$. One can see this by finding a recurrent relation between moments using moment generating function.

$$\mathbb{E}[\mathcal{N}(0, 1)]^{2k+2} = (2k + 1) \mathbb{E}[\mathcal{N}(0, 1)]^{2k}$$

■

4.2.2 Simulation

We first give some simulations to show the random projected norm converges to normal distribution.

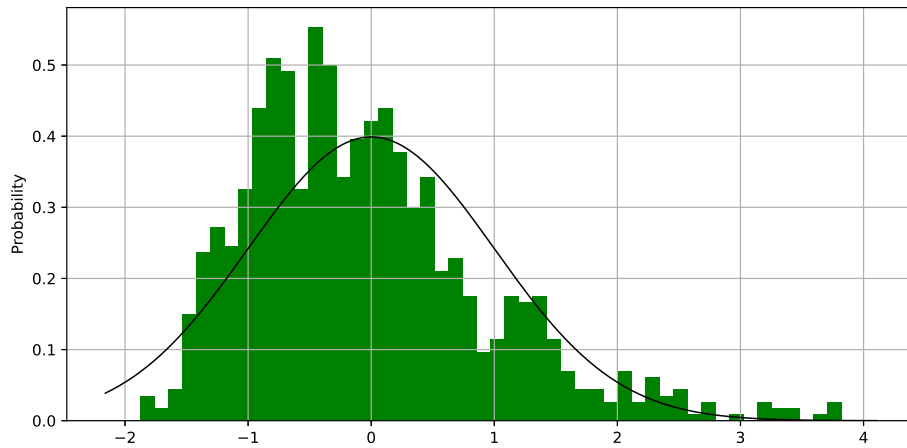


Figure 4.1: Random projected norm ($m=10, n=100$)

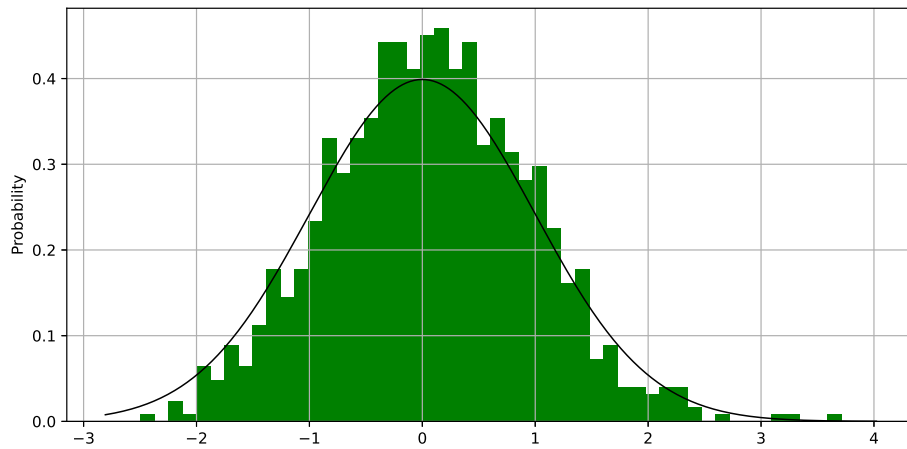


Figure 4.2: Random projected norm ($m=500, n=5000$)

Figure 4.1 and Figure 4.2 plotted histograms of 1000 samples of the projected norm

$\frac{1}{\sqrt{\sigma^2 m^2 n + 2mn^2 + \xi mn}} X^T S^T S Z$ with different dimension settings. The random variables we used for X, S, Z are standard normal random variables. As dimension m, n increases, the convergence improves.

Next we give simulations for random embedded norms where $m > n$. Even though we do not have a CLT in this setting but the Bernstein type (mixed sub-Gaussian and sub-exponential) concentration behavior we proved in section 4.1 is still relevant.

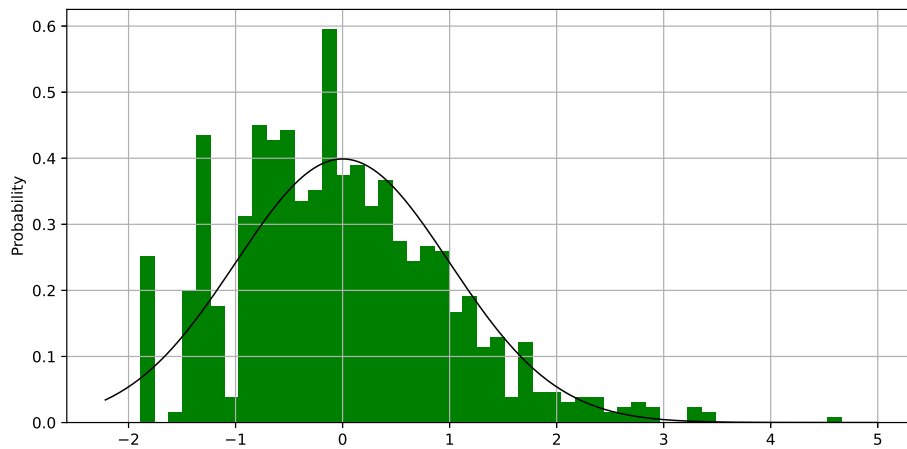


Figure 4.3: Random embedded norm ($m=200, n=20$)

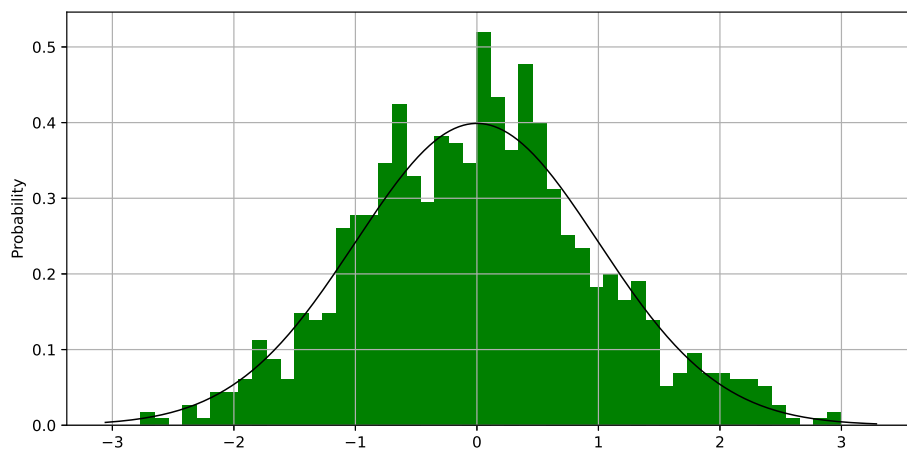


Figure 4.4: Random embedded norm ($m=2000, n=200$)

Figure 4.3 and Figure 4.4 plotted histograms of 1000 samples of the random embedded norm. The random variables we used for X, S, Z take discrete values $\{-2.5, 0, 2.5\}$ with probability $\{0.08, 0.84, 0.08\}$. The kurtosis of such random variable is 6.25 which is larger than standard normal random variable. Again as dimension m, n increases, the histogram converges to a standard normal shape.

4.2.3 Possibility of extending CLT to random embedding $m \geq O(n)$

Figure 4.3 and Figure 4.4 shows it is very likely there is a CLT for random embedded norms where $m \geq O(n)$. Of course the moment computation would fail because of the complicated dependence structure.

If we view the random transformed norm $\|SX\|^2$ as a trace function on the spectral of product of random matrices, there are potential ways from random matrix theory to overcome the difficult of too many correlated random variables when $m \geq O(n)$.

$$X^T S^T S X = \text{tr}(X^T S^T S X) = \text{tr}(S^T S X X^T)$$

It is clear the random matrix $\frac{1}{\sqrt{n}} X X^T$ has one nonzero eigenvalue which is $\lambda_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i^2 = O(\sqrt{n})$.

From random matrix analysis [31], it is known that the empirical spectral distribution of $A_n := \frac{1}{m} S^T S$ converges to the celebrated Marčenko-Pastur law depends on the parameter $\frac{m}{n} \rightarrow c, c \in (0, \infty)$.

$$\frac{1}{n} \sum_i \delta_{\lambda_i(A_n)} \xrightarrow{m \rightarrow \infty} \mu_{M,c}$$

where $\lambda_i(A_n)$ are eigenvalues of A_n and $\delta_{\lambda_i(A_n)}$ is the Dirac delta function, $\mu_{M,c}$ is the Marčenko-Pastur probability measure. Moreover, [32, 33, 34] showed if a non-negative definite $n \times n$ random matrix B_n has a deterministic limiting distribution F^B , then one can characterize the limiting spectral distribution of the product, $A_n B_n$, converges in distribu-

tion to probability distribution F .

$$\frac{1}{m} \sum_i \delta_{\lambda_i(A_n B_n)} \xrightarrow{n \rightarrow \infty} \mu_F$$

One may start thinking if it is possible to apply the result of product of random matrices to our problem. Obviously, one would replace B_n with $\frac{1}{\sqrt{n}} X X^T$. However the spectral distribution of B_n does not converge properly since $\lambda_1 = O(\sqrt{n})$. And our CLT result is actually on another level of details. One has to first center and standardize spectral of B_n , then see how the fluctuation is interacting with the spectral of A_n . In fact $A_n B_n$ has only one nonzero eigenvalue, we are actually looking at distribution of this single eigenvalue, which usually requires very different techniques to compute. The extreme eigenvalues of full rank random matrices usually converges to Tracy-Widom distribution [35, 36]. In our case, we are looking at a version of this type but the random matrix has certain structure of rank one.

4.3 Conjecture on rate of convergence

In this section we discuss the rate of convergence for the projected or embedded norm. Based on some detailed calculation, we believe the following conjecture should be true.

Conjecture 8 (Random projection of norm rate of invariance). *Given a random vector X in \mathbb{R}^n with i.i.d. entries*

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Let $\mathbb{E} x_1 = 0$, $\mathbb{E} x_1^2 = 1$, $\mathbb{E} x_1^4 = 1 + \sigma^2$ ($0 < \sigma < \infty$). Consider a random matrix $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with independent entries and $\mathbb{E} S_{i,j} = 0$ and $\mathbb{E} S_{i,j}^2 = 1$. Further assume S, X are all independent and $\mathbb{E} S_{1,1}^8 \vee \mathbb{E} |x_1|^6 < c < \infty$. Also let G be a standard normal random

variable. Then we have

$$\sup_t \left| \mathbb{P} \left(\frac{(X^T S^T S X - mn)}{\sqrt{\sigma^2 m^2 n + 2mn^2}} < t \right) - \mathbb{P}(G < t) \right| \leq O \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right) \quad (4.16)$$

$$\sup_t \left| \mathbb{P} \left(\frac{(X^T S^T S X - mn)}{\sqrt{\sigma^2 m^2 n + 2mn^2}} < t \right) - \mathbb{P} \left(\frac{(X^T X - n)}{\sigma \sqrt{n}} < t \right) \right| \leq O \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right) \quad (4.17)$$

Remark. If we use Berry-Essen theorem for random sequence $X^2 - 1 = [x_1^2 - 1, \dots, x_n^2 - 1]$ (the assumptions in Berry-Essen are satisfied since $x_i^2 - 1$ are i.i.d., $\mathbb{E} x_1^2 - 1 = 0$, $\mathbb{E}(x_1^2 - 1)^2 = \sigma^2$ and $\mathbb{E} |x_1^2 - 1|^3 < \infty$.) we find

$$\sup_t \left| \mathbb{P}(G < t) - \mathbb{P} \left(\frac{(X^T X - n)}{\sigma \sqrt{n}} < t \right) \right| \leq O \left(\frac{1}{\sqrt{n}} \right)$$

Then it is tempting to use techniques similar with the proof of Theorem 5 to prove Equation 4.16, then by triangle inequality one concludes Equation 4.17. However, the techniques in the proof of Theorem 5 heavily relies on the fact that we can separate the quantity of interests into two independent parts X and $S^T S Z$. For this conjecture on the rate of norm invariance, there is no such luxury property that we can exploit.

There are some examples suggesting this rate is the correct order. We will try to analyze in detail to see how much distortion is introduced in the projected norm with two examples. From the variance calculation Equation 4.7, we know there is an error term at least the order $O(\frac{1}{\sqrt{m+n}})$. Then let us analyze two special cases $m = 1, n \rightarrow \infty$ and $m \rightarrow \infty, n = 1$.

For $m = 1, n \rightarrow \infty$, we find

$$\begin{aligned} \frac{(X^T S^T S X - mn)}{\sqrt{\sigma^2 m^2 n + 2mn^2}} &= \frac{(\sum_{i=1}^n S_{1,i} x_i)^2 - n}{\sqrt{m} \sqrt{\sigma^2 n + 2n^2}} \\ &= \frac{1}{\sqrt{m}} \sqrt{\frac{n^2}{\sigma^2 n + 2n^2}} \left[\left(\frac{\sum_{i=1}^n S_{1,i} x_i}{\sqrt{n}} \right)^2 - 1 \right] \end{aligned}$$

$$\begin{aligned}
&\approx \frac{1}{\sqrt{m}} \sqrt{\frac{n^2}{\sigma^2 n + 2n^2}} [\mathcal{N}(0, 1)^2 - 1] \\
&= \frac{1}{\sqrt{2m}} [\mathcal{N}(0, 1)^2 - 1]
\end{aligned}$$

Since $m = 1$, $\mathcal{N}(0, 1)^2 - 1$ differ from $\mathcal{N}(0, 1)$ by $O(1)$, we see the error term is $O(\frac{1}{\sqrt{m}})$.

For $m \rightarrow \infty, n = 1$, similarly we compute

$$\begin{aligned}
\frac{(X^T S^T S X - mn)}{\sqrt{\sigma^2 m^2 n + 2mn^2}} &= \frac{x_1^2 (\sum_{i=1}^m S_{i,1}^2) - m}{\sqrt{n} \sqrt{\sigma^2 m^2 + 2m}} \\
&= \frac{x_1^2 (\sum_{i=1}^m S_{i,1}^2 - m) + m(x_1^2 - 1)}{\sqrt{n} \sqrt{\sigma^2 m^2 + 2m}} \\
&\approx \frac{x_1^2}{\sqrt{mn}} \mathcal{N}(0, 1) + \frac{x_1^2 - 1}{\sqrt{\sigma^2 n}} \\
&\approx 0 + \frac{x_1^2 - 1}{\sqrt{\sigma^2 n}}
\end{aligned}$$

In this case the error term is on the scale of $O(\frac{1}{\sqrt{n}})$ since $x_1^2 - 1$ differs from $\mathcal{N}(0, 1)$ by $O(1)$. For large m and n , we believe both $O(\frac{1}{\sqrt{m}})$ and $O(\frac{1}{\sqrt{n}})$ are necessary.

CHAPTER 5

CONCLUSION

Motivated by Johnson-Lindenstrauss lemma that random projection preserves the topological structure of a set of p deterministic vectors simultaneously, we in this work try to extend the study to random matrices, including random projection and embedding, acting on random vectors. In a nutshell, we show given p independent random vectors $X_1, \dots, X_p \in \mathbb{R}^n$ with i.i.d. entries and a random matrix $S : \mathbb{R}^n \rightarrow \mathbb{R}^m$, then the distribution of the inner product structure in \mathbb{R}^n ,

$$\langle X_i, X_j \rangle,$$

is preserved in the projected or embedded space \mathbb{R}^m , or in other words is close in distribution to

$$\langle SX_i, SX_j \rangle$$

In particular chapter 3 justifies for $i \neq j$,

$$\mathbb{P}\left(\frac{X_i^T S^T S X_j}{\sqrt{m^2 n + m n^2}} < t\right) \approx \mathbb{P}\left(\frac{1}{\sqrt{n}} X_i^T X_j < t\right)$$

by establishing a CLT type result in Theorem 4, and invariance properties in Theorem 5. The results hold for both random embedding and projection.

In chapter 4, we justify

$$\mathbb{P}\left(\frac{X_i^T S^T S X_i - mn}{\sqrt{\sigma^2 m^2 n + 2mn^2}} < t\right) \approx \mathbb{P}\left(\frac{X_i^T X_i - n}{\sqrt{n(\mathbb{E} x_1^4 - 1)}} < t\right)$$

by establishing CLT type result in Theorem 7 for random projection, and a concentration of measure result in Theorem 6 for both random embedding and projection.

The results in this work can be used as tools to analyze real world applications involving

random projections and embeddings such as, sketching for regression [9], weight analysis in neural networks [37], minimum variance portfolios [5], etc..

REFERENCES

- [1] J. Duan, I. Popescu, and F. Zhou, “An invariance principle of random projection,” *arXiv: 2106.14825*, 2021.
- [2] J. Duan, “Invariance principle of random projection for the norm,” *arXiv: 2112.00300*, 2021.
- [3] D. L. Donoho *et al.*, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS math challenges lecture*, vol. 1, no. 2000, p. 32, 2000.
- [4] V. Koltchinskii and K. Lounici, “Concentration inequalities and moment bounds for sample covariance operators,” *Bernoulli*, vol. 23, no. 1, pp. 110–133, 2017.
- [5] J. Duan and I. Popescu, “Locov: Low dimension covariance voting algorithm for portfolio optimization,” *arXiv: 2204.00204*, 2022.
- [6] W. B. Johnson and J. Lindenstrauss, “Extensions of lipschitz mappings into a hilbert space 26,” *Contemporary mathematics*, vol. 26, 1984.
- [7] P. Indyk and R. Motwani, “Approximate nearest neighbors: Towards removing the curse of dimensionality,” in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.
- [8] N. Ailon and B. Chazelle, “The fast johnson–lindenstrauss transform and approximate nearest neighbors,” *SIAM Journal on computing*, vol. 39, no. 1, pp. 302–322, 2009.
- [9] D. P. Woodruff, “Sketching as a tool for numerical linear algebra,” *arXiv preprint arXiv:1411.4357*, 2014.
- [10] T. I. Cannings and R. J. Samworth, “Random-projection ensemble classification,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 959–1035, 2017.
- [11] K. Liu, H. Kargupta, and J. Ryan, “Random projection-based multiplicative data perturbation for privacy preserving distributed data mining,” *IEEE Transactions on knowledge and Data Engineering*, vol. 18, no. 1, pp. 92–106, 2005.
- [12] M. Burr, S. Gao, and F. Knoll, “Optimal bounds for johnson-lindenstrauss transformations,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2920–2941, 2018.

- [13] W. Hoeffding and H. Robbins, “The central limit theorem for dependent random variables,” *Duke Math. J.*, vol. 15, no. 3, pp. 773–780, Sep. 1948.
- [14] B. M. Brown, “Martingale central limit theorems,” *Ann. Math. Statist.*, vol. 42, no. 1, pp. 59–66, Feb. 1971.
- [15] E. Bolthausen, “Exact convergence rates in some martingale central limit theorems,” *Ann. Probab.*, vol. 10, no. 3, pp. 672–688, Aug. 1982.
- [16] E. Haeusler, “On the rate of convergence in the central limit theorem for martingales with discrete and continuous time,” *Ann. Probab.*, vol. 16, no. 1, pp. 275–299, Jan. 1988.
- [17] J.-C. Mourrat, “On the rate of convergence in the martingale central limit theorem,” *Bernoulli*, vol. 19, no. 2, pp. 633–645, May 2013.
- [18] R. C. Bradley, “Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions,” *Probability Surveys*, vol. 2, no. none, pp. 107–144, 2005.
- [19] J. Dedecker, P. Doukhan, L. Gabriel, J. León, S. Louhichi, and C. Prieur, *Weak Dependence: With Examples and Applications*. Springer, Aug. 2007, vol. 190.
- [20] A. C. Berry, “The accuracy of the gaussian approximation to the sum of independent variates,” *Transactions of the american mathematical society*, vol. 49, no. 1, pp. 122–136, 1941.
- [21] C.-G. Esseen, “On the remainder term in the central limit theorem,” *Arkiv för Matematik*, vol. 8, no. 1, pp. 7–15, 1969.
- [22] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2019, vol. 49.
- [23] M. Ledoux, *The concentration of measure phenomenon*, 89. American Mathematical Soc., 2001.
- [24] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [25] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.
- [26] T. Tao, “Topics in random matrix theory,” *Graduate Studies in Mathematics*, vol. 132, 2011.

- [27] M. Rudelson and R. Vershynin, “Hanson-wright inequality and sub-gaussian concentration,” *Electronic Communications in Probability*, vol. 18, pp. 1–9, 2013.
- [28] J. A. Tropp, “An introduction to matrix concentration inequalities,” *arXiv preprint arXiv:1501.01571*, 2015.
- [29] D. R. Brillinger, “A note on the rate of convergence of a mean,” *Biometrika*, vol. 49, no. 3/4, pp. 574–576, 1962.
- [30] B. Von Bahr, “On the convergence of moments in the central limit theorem,” *The Annals of Mathematical Statistics*, pp. 808–818, 1965.
- [31] V. A. Marčenko and L. A. Pastur, “Distribution of eigenvalues for some sets of random matrices,” *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.
- [32] Y. Q. Yin, “Limiting spectral distribution for a class of random matrices,” *Journal of multivariate analysis*, vol. 20, no. 1, pp. 50–68, 1986.
- [33] J. W. Silverstein, “Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices,” *Journal of Multivariate Analysis*, vol. 55, no. 2, pp. 331–339, 1995.
- [34] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*. Springer, 2010, vol. 20.
- [35] C. A. Tracy and H. Widom, “Level-spacing distributions and the airy kernel,” *Communications in Mathematical Physics*, vol. 159, no. 1, pp. 151–174, 1994.
- [36] I. M. Johnstone, “On the distribution of the largest eigenvalue in principal components analysis,” *Annals of statistics*, pp. 295–327, 2001.
- [37] C. H. Martin and M. W. Mahoney, “Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning,” *Journal of Machine Learning Research*, vol. 22, no. 165, pp. 1–73, 2021.