A product-CLT and its application in invariance principle of random projection

Juntao Duan[*], Ionel Popescu[†], Fan Zhou [‡]

July 25, 2022

Abstract

Johnson-Lindenstrauss lemma states random projections can be used as a topology preserving embedding technique for fixed vectors. In this paper, we try to understand how random projections affect probabilistic properties of random vectors. In particular we prove the distribution of inner product of two independent random vectors $X, Z \in \mathbb{R}^n$ with i.i.d. entries is preserved by random projection $S : \mathbb{R}^n \to \mathbb{R}^m$. More precisely,

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

This is achieved by proving a general central limit theorem (product-CLT) for $\sum_{k=1}^n X_k Y_k$, where $\{X_k\}$ is a martingale difference sequence, and $\{Y_k\}$ has dependency within the sequence. We also obtain the rate of convergence in the spirit of Berry-Esseen theorem.

Keywords: Johnson-Lindenstrauss lemma; random projection; Central limit theorem; dependent; invariance; inner product, Berry-Esseen; rate of convergence

1 Introduction

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 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 and high noise to

[*]Corresponding author. Georgia Institute of Technology, School of Mathematics, juntaoduan@gmail.com

[†]University of Bucharest, Mathematics and Computer Science, Institute of Mathematics of the Romanian Academy, ioionel@gmail.com

[‡]Baidu Research, fanzhou@baidu.com

signal ratio. It is therefore necessary to reduce the dimensionality of data while preserving relevant structures.

The celebrated Johnson-Lindenstrauss lemma [14] 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. Since then random projections has been found very useful in many applications such as signal processing and machine learning. For example fast Johnson-Lindenstrauss random projections is used to approximate K-nearest neighbors to speed up computation [13, 1]. Random sketching uses random projection to reduce sample sizes in regression model and low rank matrix approximation [17]. Random projected features can be used to create low dimensional base classifiers which are combined as robust ensemble model [7]. Practitioners found applications of random projection in privacy and security [15]. Before we begin to state our problem, let us state the Johnson Lindenstrauss lemma [6].

Lemma 1 (Johnson and Lindenstrauss)**.** *Given a set of vectors* $\{u_1, \cdots, u_k\}$ *in* $\mathbb{R}^n$*, for any* $m \geq 8\varepsilon^{-2} \log k$*, there exists a linear map* $A : \mathbb{R}^n \to \mathbb{R}^m$ *such that*

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

Given two fixed vectors $X, Z \in \mathbb{R}^n$, by Johnson-Lindenstrauss lemma, we can find a random projections $A : \mathbb{R}^n \to \mathbb{R}^m$ such that the projected distance $\|AX - AZ\|$ has only a small distortion of the original distance $\|X - Z\|$. More precisely,

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

Equivalently, this property can be reformulated as random projections preserves the inner product of two vectors (Equivalence can be obtained by elementary computation and polarization identity). 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 $A : \mathbb{R}^n \to \mathbb{R}^m$ such that

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

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

$$|\langle AX, AZ\rangle - \langle X, Z\rangle| \leq O(\frac{1}{\sqrt{m}})\|X\|\|Z\| \leq O(\frac{1}{\sqrt{m}})(\|X\|^2 + \|Z\|^2)$$

The natural extension is to consider random vectors X, Z. Then we may ask what random projections do to random vectors? Is there an invariance phenomenon in the distribution sense? Closeness in distribution usually boils down to the difference of the cumulative distribution function. If we look at inner product, then we will be interested in controlling

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

In this work we try to address some of these question. In particular, how distributions of randomly projected random vectors changes. We obtain an invariance principle for independent random vectors very similar to the inner product form of Johnson-Lindenstrauss lemma but extended to the distribution sense. Our contributions in this paper includes:

1. We proved random projections preserves distribution of inner product of independent random vectors with i.i.d. entries. Roughly speaking, two orthogonal random vectors in high dimension remains orthogonal in the randomly projected lower dimensional space.

2. We also quantitatively characterize the distortion of distribution introduced by random projection. 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. A central limit theorem is established for random variables with 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 dependence structure.

The rest of the paper is structured as follows. We first state the main theorems in section 2. Then we prove product-CLT in section 3 and obtain the rate of convergence in section 4. Along the way, we will discuss some equivalent conditions for product-CLT theorems. In section 5, we prove the theorems concern invariance principle of random projections.

1.1 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 [12] 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 [5]. 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: $\frac{1}{n}\sum_k \mathbb{E}x_k^2 I(|x_k| > \varepsilon\sqrt{n}) \to 0$ *for all* $\varepsilon > 0$.

Then

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

The exact rate of convergence is obtained by [3]: 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 [11] and [16]. There is another line of research considering mixing weak dependence which is extensively discussed in [4] and [8]. 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 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*

$$\mathbb{E}[|X_k|^3] < A < \infty, \quad \mathbb{E}[|Y_k|^3] < A < \infty, \quad \forall k \tag{2.1}$$

Further assume

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

where $\mathcal{F}_k$ is the filtration generated by the (martingale difference) sequence $\{X_k\}$.
Assume $\{Y_k\}$ satisfies

$$\frac{1}{n}\sum_{k=1}^{n} Y_k^2 \xrightarrow{p} 1 \tag{2.3}$$

Then we have the following CLT

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_k Y_k \to G$$

where G is the standard Gaussian random variable.

Remarks. *The product-CLT can be viewed as an extension of Martingale-CLT theorem 1. If X is a 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 dependence but satisfies a law of large number condition, then the dot product X^TY has a CLT. This extension is useful because no other CLT can deal with a sequence $\{X_iY_i\}$ with unknown dependence. As we will see in the proof of invariance principle of random projections, there is no way to apply martingale-CLT directly. 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.*

In principle, one can replace the third order moment condition eq. (2.1) by the Lindeberg condition. But we prefer it to keep the argument compact. After proving the theorem, we will also give a few conditions that guarantees eq. (2.3)

Moreover, we are interested in the rate of convergence which will need control of higher order moments. Indeed, 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, except LLN of Y_i^2, in theorem 2 holds. Further assume if rate of convergence for LLN of Y_k^2 is controlled by the following condition*

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

where ε_n converges to zero. Then we have

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

where G is the standard normal random variable.

We then use the product-CLT theorem to obtain invariance of the distribution of inner product of randomly projected or embedded random vectors.

Theorem 4 (Random matrix 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* $S : \mathbb{R}^n \to \mathbb{R}^m$ *with i.i.d. 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*

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

Theorem 5 (Random matrix invariance principle)**.** *Given the same moment assumptions as in theorem 4, the following bounds hold,*

$$\sup_t \left| \mathbb{P}\left(\frac{1}{\sqrt{m^2 n + mn^2}} X^T S^T S Z < t\right) - \mathbb{P}(G < t) \right| \leq O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \tag{2.5}$$

$$\sup_t \left| \mathbb{P}\left(\frac{1}{\sqrt{m^2 n + mn^2}} X^T S^T S Z < t\right) - \mathbb{P}\left(\frac{1}{\sqrt{n}} X^T Z < t\right) \right| \leq O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \tag{2.6}$$

where G *is a standard normal random variable.*

3 Proof of 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

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

Our strategy is to split the difference into two parts

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

then show both are small.

First step, let us try to show

$$\mathbb{E}[f(\frac{1}{\sqrt{n}} S_n)] - \mathbb{E}[f(\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i Y_i)] \to 0$$

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 $\mathcal{F}_{k-1}$ the first order terms match.

$$\mathbb{E}[f'(U_k)\frac{1}{\sqrt{n}} X_k Y_k] = \mathbb{E}[f'(U_k)\frac{1}{\sqrt{n}} Y_k \mathbb{E}[X_k|\mathcal{F}_{k-1}]] = 0$$

$$\mathbb{E}[f'(U_k)\frac{1}{\sqrt{n}} Z_k Y_k] = \mathbb{E}[f'(U_k)\frac{1}{\sqrt{n}} Y_k \mathbb{E}[Z_k]] = 0$$

Similar argument shows second 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 will only introduce the difference of the order $n^{-1/2}$

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

Now it suffices to show

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i Y_i \to \mathcal{N}(0,1)$$

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

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

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

□

3.2 Alternative assumptions

Here we discuss a variation of the condition eq. (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 approximated with empirical data.

Proposition 2. *In theorem 2, if Y_k satisfied,*

$$\mathbb{E}Y_k^2 \to 1, \quad \mathbb{E}[|Y_k|^4] < C < \infty, \qquad \forall k \in \mathbb{N}$$

Further assume the mixed second moments satisfy

$$\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[Y_i^2 Y_j^2] \xrightarrow{n\to\infty} 1 \tag{3.1}$$

Then the following LLN holds

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

Proof: By Chebyshev's inequality,

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

Notice

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

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

$$\mathbb{P}(\frac{1}{n}|\sum_i (Y_i^2-1)| > \varepsilon) \to 0$$

This implies there is a weak law of large number for the sequence Y_n^2, namely

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

□

In practice one will only need to verify that the average $\frac{1}{n^2}\sum_{i,j} Y_i^2Y_j^2$ is close to 1. It turns out that the mixed second moments condition is equivalent to a fourth moment convergence condition.

Proposition 3. *In proposition 2, the condition eq.* (3.1) *is equivalent to*

$$\mathbb{E}\left[P_n^4\right] \to 3, \qquad \textit{where } P_n = \frac{1}{\sqrt{n}}\sum_i X_iY_i \tag{3.2}$$

Proof:

$$\mathbb{E}[P_n^4] = n^{-2} \sum_{1\le i_1,\dots i_4\le n} \mathbb{E}X_{i_1}Y_{i_1}\dots X_{i_4}Y_{i_4}$$

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 vanish. Therefore the only surviving terms must be either all indices the same or indices appear as pairs. Namely

$$\mathbb{E}[P_n^4] = n^{-2}\left[\sum_{1\le i\le n} \mathbb{E}X_i^4Y_i^4 + 3\sum_{1\le i\ne j\le n} \mathbb{E}X_i^2Y_i^2X_j^2Y_j^2\right]$$

where the factor 3 is because the pairs have three cases $\{(i_1 = i_2, i_3 = i_4), (i_1 = i_3, i_2 = i_4), (i_1 = i_4, i_2 = i_4)\}$.

Then notice $\mathbb{E}(X_i^2X_j^2) = \mathbb{E}[\mathbb{E}[X_i^2X_j^2|\mathcal{F}_{min(i,j)}]] = 1$, we find

$$\mathbb{E}X_i^2Y_i^2X_j^2Y_j^2 = \mathbb{E}(X_i^2X_j^2)\mathbb{E}Y_i^2Y_j^2 = \mathbb{E}Y_i^2Y_j^2$$

Combining with the assumption that fourth moment is bounded we see

$$\mathbb{E}[P_n^4] \to 3 \Longleftrightarrow \frac{1}{n^2}\sum_{i\ne j}\mathbb{E}Y_i^2Y_j^2 \to 1$$

□

4 Proof of theorem 3

The proof will be several steps. First we record a variation formula of Gaussian density in lemma 4. Then we use the variation formula to rewrite the error term by introducing a standard normal variable in lemma 5. Then we use Lindeberg type argument to reduce the control of error to control of two terms. One term is a telescopic sum which we will control in lemma 6 with the moments information. The other term is the difference of two cumulative distribution functions (cdfs) that are close to normal cdfs which we will control in lemma 7 with the LLN property of Y_i^2, namely condition eq. (2.4).

Lemma 4. *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*

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

Then

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

Proof: See for example [3] □

Lemma 5. *Denote*

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

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

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

Proof:
By lemma 4 we have

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

Again by lemma 4, we see

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

□

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 5

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

where

$$\begin{aligned}\Delta_t &= \mathbb{P}\left(\frac{\xi + \sum Z_i Y_i}{\sqrt{n}} \le t\right) - \mathbb{P}\left(\frac{\xi}{\sqrt{n}} + G \le t\right) \\ T_m &= t\sqrt{n} - \sum_{i=1}^{m} X_i Y_i - \sum_{i=m+1}^{n} Z_i Y_i\end{aligned}$$

Therefore with lemma 6 controlling the part of telescopic sum and lemma 7 controlling $\sup_t |\Delta_t|$ (which we will prove later in section 4.1), we see,

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

Then by lemma 5, we conclude the desired result

$$\sup_t \left|\mathbb{P}\left(\frac{\sum X_i Y_i}{\sqrt{n}} \le t\right) - \Phi(t)\right| \le O(\varepsilon_n \vee \frac{1}{\sqrt{n}})$$

□

Remarks. *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 [3]. Martingale CLT is derived under a slightly weaker condition on variance, which only requires $\frac{1}{n}\sum_k \mathbb{E}[X_k^2|\mathcal{F}_{k-1}] \to 1$ instead of our condition that $\mathbb{E}[X_k^2|\mathcal{F}_{k-1}]$ to be constant 1 for all k.*

4.1 Proof of lemma 6 and lemma 7

Lemma 6. *If $\mathbb{E}X_k^3 < A < \infty, \mathbb{E}Y_k^3 < A < \infty, \forall k$ then there is a constant c*

$$\sup_t \left| \mathbb{E}\left[\sum_{m=1}^{n} \Phi\left(T_m\right) - \Phi\left(T_{m-1}\right) \right] \right| \leq \frac{c}{\sqrt{n}} \tag{4.1}$$

Proof: Let $U_k = T_k - X_kY_k = T_{k-1} - Z_kY_k$, then

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

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

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

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

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

□

Lemma 7. *If condition eq.* (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 argument in lemma 5, we can removing the same variation term, normal random variable $\frac{\xi}{\sqrt{n}}$ in Δ_t. So for some constant c_0,

$$\sup_t |\Delta_t| \leq 2 \sup_t \left| \mathbb{P}\left(\frac{\sum Z_iY_i}{\sqrt{n}} \leq t \right) - \mathbb{P}\left(G \leq t\right) \right| + \frac{c_0}{\sqrt{n}}$$

$$\sup_t \left| \mathbb{P}\left(\frac{\sum Z_iY_i}{\sqrt{n}} \leq t \right) - \mathbb{P}(G \leq t) \right| = \sup_t \mathbb{E}\left| \mathbb{P}(G \leq t\sqrt{\frac{n}{\sum Y_i^2}}) - \mathbb{P}(G \leq t) \right|$$

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

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

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

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

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

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 eq. (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

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

□

4.2 Discussion on the assumptions

A natural question is whether the condition eq. (2.4) is necessary for theorem 3. We will first show the condition eq. (2.4) for lemma 7 is sharp by obtaining a lower bound for $\sup_t \mathbb{E}h(t)$. This implies the technique we used in proving theorem 3 is delicate enough to squeeze out any unnecessary relaxation.

Proposition 8.

$$\sup_t \left| \mathbb{P}\left(\frac{\sum Z_i Y_i}{\sqrt{n}} \le t \right) - \mathbb{P}(G \le t) \right| := \sup_t \mathbb{E}h(t) \ge O(\mathbb{E}\left[1 \wedge \left| \sqrt{\frac{\sum Y_k^2}{n}} - 1 \right| \right]) \tag{4.2}$$

Proof: Take $t = 1$ we find

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

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

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

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

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

$$|\frac{1}{S_n} - 1| \ge \frac{1}{2}[1 \wedge |S_n - 1|]$$

Combining all together,

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

□

Next we will use i.i.d. X_i, Y_i as an example to show the rate of convergence obtained from theorem 3 is the same as Berry-Esseen in classical CLT. This implies the condition eq. (2.4) is sharp for this specific example. Note this does not imply condition eq. (2.4) is sharp in general. However, any nontrivial improvement would require more restrictive assumptions.

Proposition 9. *In theorem 3, condition eq.* (2.4) *is sharp if X_i, Y_i are i.i.d. sequences with mean zero and variance one.*

Proof: 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

$$\mathbb{P}(\frac{1}{\sqrt{n}}\sum_{k=1}^{n} X_k Y_k < t) = \mathbb{P}(G < t) + O(\frac{1}{\sqrt{n}}) \quad \forall t \in \mathbb{R}$$

Let us derive the same result from theorem 3. For i.i.d. Y_i mean zero and variance one, since $\sqrt{n}(\sum Y_i^2/n - 1) \to \mathcal{N}(0,1)$ and $(\sqrt{\sum Y_i^2/n} + 1) \to 2$ in probability and we can apply Slutsky's theorem,

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

Therefore condition eq. (2.4) is satisfied with

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

Then theorem 3 gives the same conclusion as Berry-Esseen. □

A more intuitive control of the LLN of Y_k^2 would be controlling the tail probability directly, which will not be sharp.

Proposition 10. *In theorem 3, condition eq.* (2.4) *can be replaced by*

$$\mathbb{P}\left(\left|\sqrt{\frac{\sum Y_k^2}{n}} - 1\right| > O(\varepsilon_n)\right) \le O(\varepsilon_n) \tag{4.3}$$

where $\varepsilon_n \to 0$. This condition is stronger than eq. (2.4)*. In other words, it is sufficient for theorem 3 but not necessary.*

Proof: Let $S_n = \sqrt{\frac{\sum Y_k^2}{n}}$. Assume eq. (4.3) holds.

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

To show condition eq. (4.3) is stronger than condition eq. (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,

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

Since $\sqrt{n}(\sum Y_i^2/n - 1) \to \mathcal{N}(0,1)$ and $(\sqrt{\sum Y_i^2/n} + 1) \to 2$ in probability and we can apply Slutsky's theorem. Therefore condition eq. (2.4) is satisfied

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

However, condition eq. (4.3) is not satisfied since

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

□

5 Random projection

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 all X_i, Z_i independent with each other. Let $\mathbb{E}X_i = \mathbb{E}Z_i = 0, \quad \mathbb{E}X_i^2 = \mathbb{E}Z_i^2 = 1, \quad \mathbb{E}|X_i|^3 \vee \mathbb{E}|Z_i|^3 < A < \infty$. 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 [2, 10, 9] tells us

$$\sup_t \left|\mathbb{P}\left(\frac{1}{\sqrt{n}} X^T Z < t\right) - \mathbb{P}\left(\mathcal{N}(0,1) < t\right)\right| \leq O\left(\frac{1}{\sqrt{n}}\right) \tag{5.1}$$

Consider a random matrix $S : \mathbb{R}^n \to \mathbb{R}^m$ whose entries has 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 depend on both m and n. Moreover, we will need to derive the rate of convergence in the spirit of Berry-Esseen theorem, namely find

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

If we try to use existing CLT that deals 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^TSZ 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

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} := \frac{1}{a_{n,m}} S^T S Z$$

is also satisfied. Denote i-th column of S as S_i, then $Y_i = \frac{1}{a_{n,m}} S_i^T SZ$. 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 the following lemma 11. In the proof we will follow proposition 2 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 eq. (2.4).

5.1 Random matrix preserves inner product

Lemma 11. *Given $m, n \to \infty$, $a_{n,m} = \sqrt{m^2 + mn}$, 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*

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

then we have

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

and

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

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

For the second moment,

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

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

$$\mathbb{E}[y_1^2] = \frac{1}{a_{n,m}^2}\left[\sum_{1\le i\le m, 2\le p\le n} \mathbb{E}S_{i,1}^2 S_{i,p}^2 \mathbb{E}z_p^2 + \sum_{1\le i,j\le m} \mathbb{E}S_{i,1}^2 S_{j,1}^2 \mathbb{E}z_1^2\right]$$

$$= \frac{1}{(m^2+mn)}[m(n-1) + (m\mathbb{E}S_{1,1}^4 + m^2 - m)]$$
$$= 1 + \frac{\mathbb{E}S_{1,1}^4 - 2}{m+n}$$
$$= 1 + O\left(\frac{1}{m+n}\right) \to 1 \tag{5.2}$$

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

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

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 $\mathbb{E}S^2_{i_1,1}S^2_{j_1,1}S_{i_2,2}S_{i_2,q}S_{j_2,2}S_{j_2,q}z_1^2z_q^2$. 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), $\mathbb{E}S^2_{i_1,1}S_{j_1,1}S_{j_1,q}S_{i_2,2}S_{i_2,1}S_{j_2,2}S_{j_2,q}z_1^2z_q^2$. 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 $\mathbb{E}S_{i_1,1}S_{i_1,q}S_{j_1,1}S_{j_1,q}S^2_{i_2,2}S^2_{j_2,2}z_q^2z_2^2$. 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.
$\mathbb{E}S_{i_1,1}S_{i_1,2}S_{j_1,1}S_{j_1,2}\ S_{i_2,2}S_{i_2,q}S_{j_2,2}S_{j_2,q}z_q^2z_2^2$. 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 $\mathbb{E}S^2_{i,1}S^2_{i,2}S_{i,p_1}S_{i,p_2}S_{i,q_1}S_{i,q_2}$. 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 $m(n-2)\mathbb{E}S^4_{1,1}\mathbb{E}z_1^4$ 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

$$(m^2 - m)[(n-2)^2 - (n-2)] + (m^2 - m)(n-2)\mathbb{E}z_1^4$$

$$=m^2n^2 - mn^2 + (\mathbb{E}z_1^4 - 5)m^2n + O(mn + m^2)$$

Adding all the cases together we obtain

$$\mathbb{E}[y_1^2 y_2^2] = \frac{1}{(m^2+mn)^2}[m^4 + 2m^3n + m^2n^2 + O(m^3 + m^2n + mn^2)] \tag{5.3}$$

$$=1 + \frac{O(m^3 + m^2n + mn^2)}{(m^2+mn)^2} \tag{5.4}$$

$$=1 + O\left(\frac{1}{m} + \frac{1}{m+n}\right) \to 1 \tag{5.5}$$

Therefore,

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

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

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

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.

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

□

Proof: **(theorem 4 Random matrix inner product CLT)**

Combing lemma 11 with the product-CLT theorem 2, we conclude random projection preserves product-independence, namely given conditions in **theorem 4** we have

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

□

Now we shall discuss the rate of convergence. Obtaining the exact rate is usually very hard since one has to compute the exact rate of convergence for the law of large number statement on Y_k^2 (namely condition eq. (2.4)), which is not practically computable if no further information given. However it is possible to carry out an argument (for example using relaxations or proposition 10) to obtain an upper bound.

Proof: **(theorem 5 Random matrix invariance principle)**
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,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\sum Y_k^2}{n} - 1\right)^2\right] &= \frac{1}{n^2}\left[\sum_{k=1}^{n} \mathbb{E}(Y_k^2 - 1)^2 + \sum_{i \neq j}^{n} \mathbb{E}(Y_i^2 - 1)(Y_j^2 - 1)\right] \\ &= \frac{1}{n}\mathbb{E}(Y_1^2 - 1)^2 + \frac{n^2 - n}{n^2}\mathbb{E}(Y_1^2 - 1)(Y_2^2 - 1)\end{aligned}$$

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 11.

$$\begin{aligned}\mathbb{E}[Y_2^2] = \mathbb{E}[Y_1^2] &= 1 + O\left(\frac{1}{m+n}\right) \\ \mathbb{E}[Y_1^4] &\leq 4 + O\left(\frac{1}{m} + \frac{1}{m+n}\right) \leq 4 + O\left(\frac{1}{m}\right) \\ \mathbb{E}Y_1^2 Y_2^2 &= 1 + O\left(\frac{1}{m} + \frac{1}{m+n}\right) = 1 + O\left(\frac{1}{m}\right)\end{aligned}$$

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 eq. (2.5). Then combining Berry-Esseen inequality eq. (5.1) and triangle inequality, we obtain eq. (2.6). □

5.2 Simulation

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

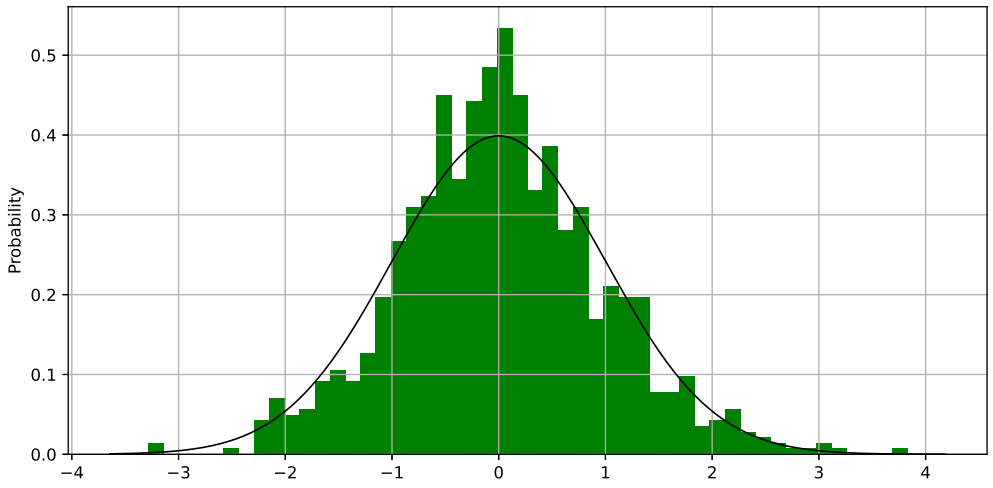

Figure 1: Random projected inner product (m=10, n=100)

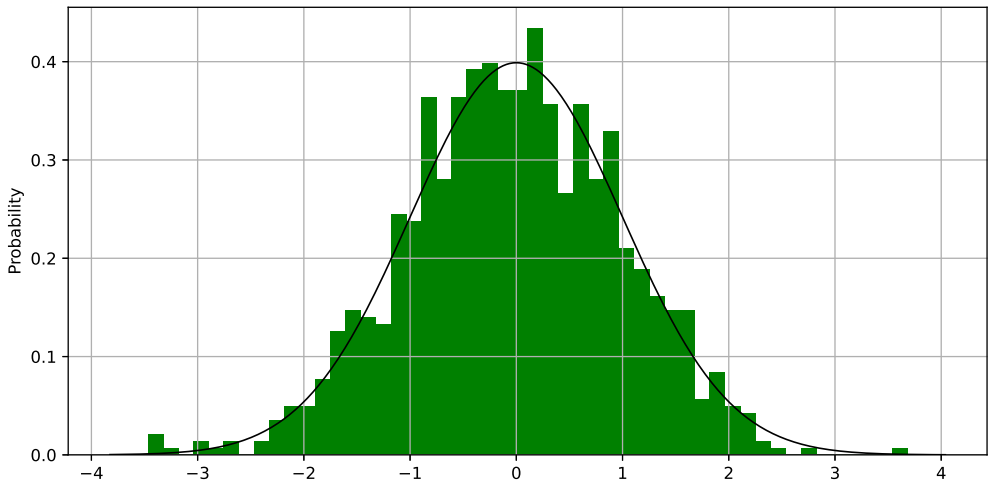

Figure 2: Random projected inner product (m=500, n=5000)

We have fig. 1 and fig. 2 plotted histograms of 1000 samples of the projected inner product $\frac{1}{\sqrt{m^2n+mn^2}}X^TS^TSZ$ 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 $m > n$.

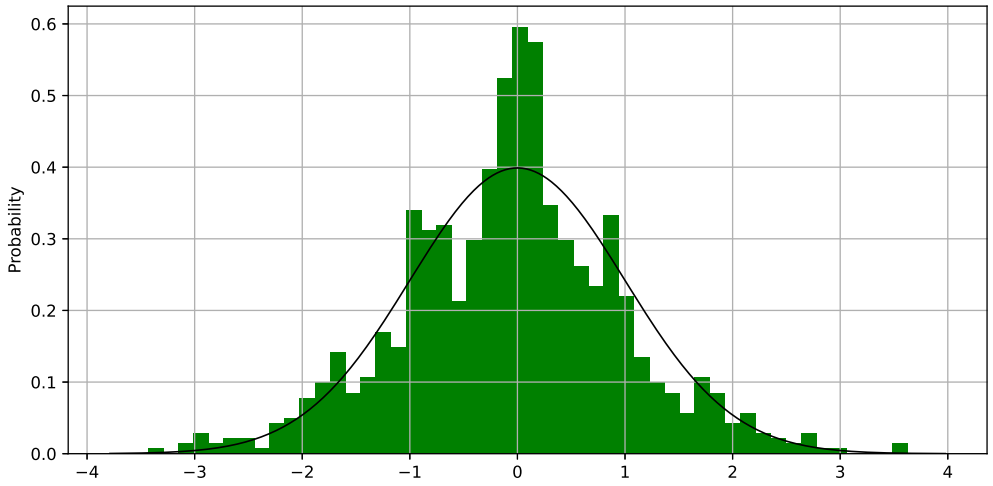

Figure 3: Random embedded inner product (m=500, n=50)

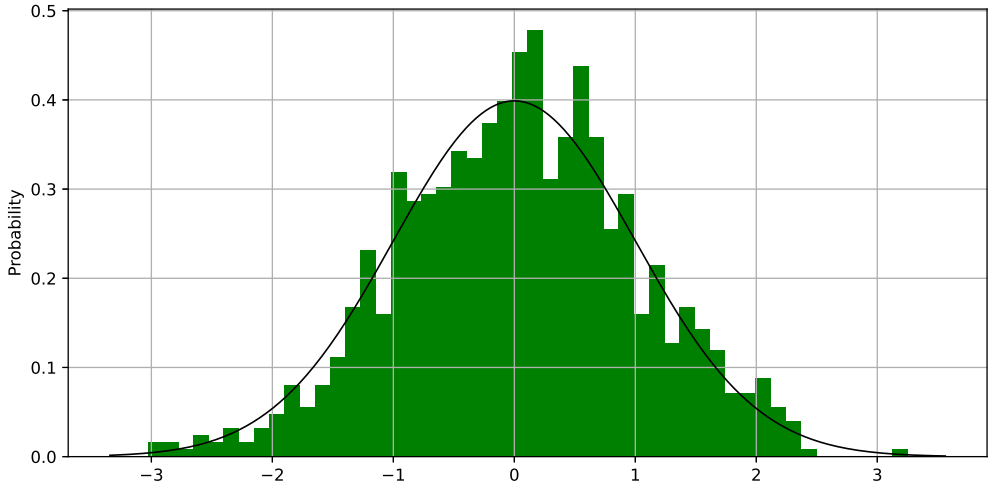

Figure 4: Random embedded inner product (m=5000, n=500)

Again fig. 3 and fig. 4 plotted histograms of 1000 samples of the embedded inner product $\frac{1}{\sqrt{m^2n+mn^2}}X^TS^TSZ$. 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.

5.3 Discussion and open questions

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 $m = 1, n \to \infty$. Then let $X \in \mathbb{R}^n$ be Gaussian vector, $Z \in \mathbb{R}^n$ be Rademacher vector and let $S : \mathbb{R}^n \to \mathbb{R}$ 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

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

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 $\mathcal{N}(0,1)$. $\frac{1}{\sqrt{n+1}} SZ$ converges to another $\mathcal{N}'(0,1)$ with error $O(\frac{1}{\sqrt{n}})$. 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 $\frac{1}{\sqrt{1^2 n + 1 n^2}} X^T S^T S Z$ differ by $O(1) = O(\frac{1}{\sqrt{m}})$.

The bound eq. (2.5) in general can not be improved if there is no additional assumption. $O(\frac{1}{\sqrt{n}})$ is necessary as it is in Berry-Esseen. $O(\frac{1}{\sqrt{m}})$ 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 $S : \mathbb{R}^n \to \mathbb{R}^m$ 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 7 and proposition 8, we have the lower bound.

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

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 $O(\frac{1}{\sqrt{m}})$ is obtained.

On the other hand, it is not clear whether eq. (2.6) can be improved. In some cases, $O(\frac{1}{\sqrt{n}})$ is not necessary. For example, if we let $m \to \infty, n = 1$, then

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

In the original Johnson-Lindenstrauss lemma, the number of vectors p can be arbitrary ($p \ge 2$) and the error has a factor $\log p$. So far, we only discussed the case $p = 2$. 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 another future work.

References

[1] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

[2] Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

[3] E. Bolthausen. Exact convergence rates in some martingale central limit theorems. *Ann. Probab.*, 10(3):672–688, 08 1982.

[4] Richard C. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2(none):107 – 144, 2005.

[5] B. M. Brown. Martingale central limit theorems. *Ann. Math. Statist.*, 42(1):59–66, 02 1971.

[6] Michael Burr, Shuhong Gao, and Fiona Knoll. Optimal bounds for johnson-lindenstrauss transformations. *The Journal of Machine Learning Research*, 19(1):2920–2941, 2018.

[7] Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

[8] Jrme Dedecker, Paul Doukhan, Lang Gabriel, Jos Len, Sana Louhichi, and Clmentine Prieur. *Weak Dependence: With Examples and Applications*, volume 190. Springer, 08 2007.

[9] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[10] Carl-Gustav Esseen. On the remainder term in the central limit theorem. *Arkiv för Matematik*, 8(1):7–15, 1969.

[11] Erich Haeusler. On the rate of convergence in the central limit theorem for martingales with discrete and continuous time. *Ann. Probab.*, 16(1):275–299, 01 1988.

[12] Wassily Hoeffding and Herbert Robbins. The central limit theorem for dependent random variables. *Duke Math. J.*, 15(3):773–780, 09 1948.

[13] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[14] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26, 1984.

[15] Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1):92–106, 2005.

[16] Jean-Christophe Mourrat. On the rate of convergence in the martingale central limit theorem. *Bernoulli*, 19(2):633–645, 05 2013.

[17] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.