A Glimpse of Lindeberg Replacement Method

Leda Wang

Department of Mathematics, USTC.

Abstract

The central limit theorem is a fundamental result in probability theory that describes the behavior of a collection of independent and identically distributed random variables. Specifically, it states that the sum (or mean) of these variables tends towards a normal distribution, even if the individual variables themselves do not follow a normal distribution. While the central limit theorem has been known for over a century, the methods used to prove it have evolved over time. One such method is known as the Lindeberg Method, which was first introduced by the Swedish mathematician Harald Cramér in the 1920s. Cramér's student, H. C. Lindeberg, further developed the method in a series of three papers published between 1920 and 1922. Lindeberg's approach is notable for its simplicity and elegance, making it accessible and easy to understand even for those new to the field. In recent years, there has been a resurgence of interest in the Lindeberg Method, driven in part by its broad applicability to a range of fields including random matrices, high-dimensional Gaussian approximations, and more. In this paper, we provide a comprehensive overview of the Lindeberg Method, including its history, key concepts, and applications. Additionally, we examine recent advancements that have expanded the scope of the Lindeberg Method and opened up exciting new research directions. We hope that this paper will serve as a valuable resource for researchers and students interested in probability theory, statistics, and related fields.

Keywords: Lindeberg Methods, Random Matrix Theory, Gaussian Approximation, Central Limit Theorem, High-Dimensional Probability, Berry-Esséen Bound, Limit Theorem

Contents

1 Introduction

The term, central limit theorem, has been consistently used since the work [19] by Georg Pólya in 1920's entitled *On the central limit theorem of probability calculation and the moment problem whenever Gaussian density occurs as a limit distribution in a stochastic model.* Pólya's phrasing highlights the fundamental meaning of this group of limit theorems.

In probability theory, the term "central" refers to the behavior of the distribution center in a random model. This concept is especially significant when considering the central limit theorem, which pertains to the properties of partial sums of independent random variables. After removing outliers, these sums exhibit a Gaussian limit distribution, where the behavior of the distribution center plays a crucial role.

The central limit theorem is of fundamental importance not just in probability theory, but in many fields that rely on statistical methods. Its significance lies in its ability to describe the behavior of averages, or more generally, linear combinations of independent random variables. A crucial aspect of the central limit theorem is the notion of the distribution center, also known as a central tendency. This refers to the symmetric point around which the distribution is centered and represents the most probable outcome. Conversely, the behavior of the distribution's tails is relatively less consequential since extreme values have a minimal impact on the overall distribution. In contemporary probability theory, the focus primarily revolves around comprehending the behavior of the distribution center concerning the central limit theorem. This enables us to gain a better understanding of the distribution's properties as a whole, leading to more effective analysis and modeling of complex systems.

The central limit theorem has a fascinating historical background. The initial version of this theorem was proposed by the French mathematician Abraham de Moivre, who published a remarkable article in 1733. In his work, he utilized the normal distribution to approximate the distribution of the number of heads obtained from multiple coin tosses. Despite its groundbreaking nature, De Moivre's discovery was nearly forgotten until it was revived by the renowned French mathematician Pierre-Simon Laplace. Laplace further developed De Moivre's work by approximating the binomial distribution with the normal distribution. However, it was not until the late nineteenth century that the true significance of the central limit theorem became apparent. In 1901, the Russian mathematician Aleksandr Lyapunov provided a general definition of the theorem and offered precise mathematical proof of its workings. Since then, the central limit theorem has been recognized as a cornerstone of probability theory, holding immense importance in the field.

The Lyapunov condition also satisfies the Lindeberg condition and, consequently, the central limit theorem's validity follows. Lyapunov used the characteristic function as a methodology to prove probability theory, but his work remained unknown outside of Russia for a prolonged period. Subsequently, the historical narrative transitions to Lindeberg's contribution, whose method of proof was impressively straightforward.

The Finnish mathematician, Earl Waldemar Lindeberg (1876-1932), was born and raised in Helsinki. He was educated at the University of Helsinki and Paris and specialized in studying partial differential equations, ultimately earning his doctorate in 1902. When composing his first work [14] on the central limit theorem 1920, he did not know the results of Lyapunov but did know the weaker results of von Mises. In 1922 Lindeberg wrote the works [15, 16] in which his method and the condition named after him were fully developed.

Paul Lévy wrote 1925 his famous book *Calcul des probabiliés*, in which he presented a particular form of Lindeberg's proof. However, Lévy decisively relied on characteristic functions, similar to Lyapunov's approach. This could be the reason why the Lindeberg method found limited representation in some traditional textbooks. An extension of Lindeberg's Central Limit Theorem, known as the Feller condition, was later proven by Feller himself. Feller hypothesized that Lévy's proof substituted the Lindeberg method with the utilization of Fourier theory. Le Cam's appreciation of the Lindeberg Method in [13] and Pollard's comment on it at the end of the article [13] are impressive. Pollard concludes his comment that Lindeberg's argument still has something to offer.

From Laplace to the mid-20th century, the Central Limit Theorem (CLT) has played a crucial role in bridging various cultural and intellectual aspects of probability theory. In classical probability theory, the CLT served as a "natural law," offering insights into the order inherent in the normal distribution amidst the complex interplay of individual random variables. During the mid-19th century, the CLT gained increasing mathematical significance, initially serving as an illustration of specific analytical theories and techniques. Subsequently, with Lyapunov's contributions, the CLT underwent a transformation into an independent mathematical concept studied for its intrinsic value, with implications for other branches of mathematics.

While preserving the fundamental classical structures such as the independence of added random variables, the CLT was occasionally generalized to encompass non-normal limit laws and weakened forms of independence. Consequently, it became a subject of enduring interest in modern mathematics. It is therefore intriguing to note that, after nearly a century, the Lindeberg Method, which existed in the past, is experiencing a renaissance. In the following chapters, we delve into this exciting revival and explore its implications.

In Chapter 2, we present the original Lindeberg method for calculating partial sums of independent random variables. Despite its simplicity, this method is often overlooked in many textbooks. We strongly encourage readers to explore the Lindeberg proof method, not only for its elegance but also for the potential modifications it offers, which can provide deeper insights beyond the original approach.

Chapter 3 provides examples of random partial sums and martingales, illustrating the evolution and extension of the Lindeberg method in various application fields.

Moving forward, Chapter 4 focuses on the remarkable work of Chatterjee, while Chapter 5 is dedicated to the groundbreaking findings of Tao and Vu in the field of random matrix theory. These chapters highlight significant advancements that have built upon Lindeberg's method, showcasing its enduring intellectual influence.

In Chapter 6, we delve into the major breakthroughs in multivariate Berry-Esséen bounds, starting with Bentkus' contributions. Chernozhukov and others have made a remarkable discovery by demonstrating that a specific class of hyperrectangles can significantly limit the corresponding distance at a logarithmic rate, which depends on p. These works are deeply rooted in Lindeberg's method and further underscore his enduring intellectual legacy.

2 Primitive Approach in Lindeberg's works

2.1 Proof of the classical Central Limit Theorem

We will now introduce Lindeberg's elegant proof method in the simplest scenario of probability theory, which involves real-valued independent random variables $X_1, X_2, \cdots$. To simplify our analysis, we assume $\mathbb{E}(X_i) = 0$ without loss of generality. Also, denote $\Phi(x)$ is the cumulative distribution function of a normal random variable, and $\varphi(x)$ is the probability density function of it, i.e. $\Phi(x) = \int_{-\infty}^{x} \varphi(y)\mathrm{d}y$ and $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$.

We also denote the variance of the random variable X_i with $\sigma_i^2 = \mathbb{E}[(X_i - \mathbb{E}X_i)^2]$. Then, due to the assumed independence of the random variable, the variance of $\sum_{i=1}^{n} X_i$ is $B_n^2 = \sum_{i=1}^{n} \sigma_i^2$. Also denote $W_n = \frac{1}{B_n} \sum_{i=1}^{n} X_i$.

The question of the validity of a central limit theorem in this situation is the question of the conditions attached to the random variable X_i, s.t.

$$\frac{1}{B_n} \sum_{i=1}^{n} X_i \xrightarrow{D} Z \tag{1}$$

where $Z \sim \mathscr{N}(0,1)$.

We note that $\mathbb{E}(W_n) = 0$ since every X_i is a centered random variable, and $\mathbb{E}(W_n^2) = 1$ since the scaling factor, so the random variable W_n for each $n \geq 1$ in the first two moments $\mathbb{E}(W_n)$ and $\mathbb{E}(W_n^2)$ correspond to the moments of the canonical Gaussian distribution. We'll come back to this moment matching later.

Specially, we first consider the case where the random variables X_i have the same distribution and assume variance $\mathbb{E}(X_i^2) = \mathbb{E}(X_1^2) = 1$. Then we have the well-known fact that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{D} Z. \tag{2}$$

Also, suppose $Z_1, Z_2, \cdots, Z_n$ are i.i.d. with the same distribution of $\mathscr{N}(0,1)$ and they are all independent of $X_1, X_2, \cdots, X_n$. Then we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \overset{D}{=} Z, \tag{3}$$

and since we only consider the information of distribution, we can assume that equality holds actually more than an equality in law. This is important that this observation is a characteristic property of the normal distribution.

Lindeberg's method, as proposed, entails a progressive substitution of the summands X_i in the sum W_n with normally distributed variables Z_i. This substitution facilitates the approximation of the sum by Z and leads to a concise proof of the central limit theorem. What's more, Lindeberg's approach enables the extension of the central limit theorem to non-identically distributed random variables, provided that they adhere to the condition named after Lindeberg himself.

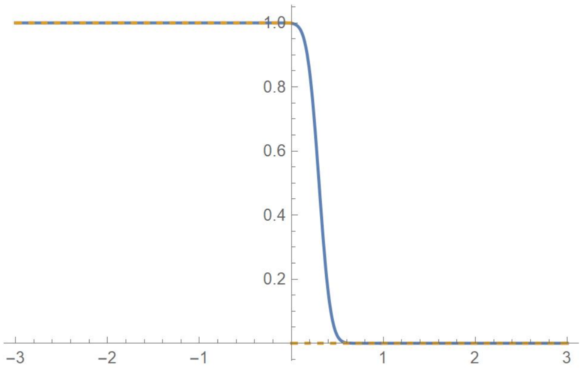

Fig. 1 Smooth approximation to an indicator function

A small technical preparation is the following: we write $\mathbb{P}(W_n \leq x) = \mathbb{E}(1\{W_n \leq x\})$ with the notation of the indicator function $1\{W_n \leq x\}$, which assumes the value 1 on the event $W_n \leq x$ and otherwise the value 0. We want to show that

$$\lim_{n\to\infty} \mathbb{E}\left(1\{W_n \leq x\}\right) = \mathbb{E}\left(1\{Z \leq x\}\right). \tag{4}$$

For this purpose, we approximate the indicator function $1(-\infty, x]$ by a sufficiently smooth function f. To achieve our goals, it is sufficient to choose an f which is three times differentiable and whose derivatives are constant and limited. These test functions always exist by a mollifier convolution, and we can select one such function, as illustrated in the image below 1. Also, by the Portmanteau Theorem, for such a class consisting of all f, the weak convergence is equivalent to the convergence of expectation over this class, i.e., it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}f(W_n) = \mathbb{E}f(Z) \tag{5}$$

Define a function as a type of Taylor remaining terms

$$g(t) = \sup_{x\in\mathbb{R}} \left\{f(x+t) - f(x) - f'(x)t - 1/2f''(x)t^2\right\}, \tag{6}$$

we know it can be easily bounded by Taylor expansion. As a result, there exists some constant $K > 0$ depending on the global supreme of some derivatives of f such that

$$g(t) \leq Kt^2 \wedge |t|^3, \tag{7}$$

thus

$$|f(x+t_1) - f(x+t_2) - f'(x)(t_1 - t_2) - 1/2f''(t_1^2 - t_2^2)| \leq g(t_1) + g(t_2). \tag{8}$$

Now, turns back to our original topic. We define

$$Y_k = \sum_{i=1}^{k-1} X_i + \sum_{j=k+1}^{n} Z_j, \tag{9}$$

then $\frac{1}{\sqrt{n}}(Y_n + X_n) = W_n$ and $\frac{1}{\sqrt{n}}(Y_1 + Z_1) = Z$. As a result, we can approach it step-by-step:

$$\mathbb{E}\left(f(W_n) - f(Z)\right) = \mathbb{E}\left(f(\frac{1}{\sqrt{n}}(Y_n + X_n)) - f(\frac{1}{\sqrt{n}}(Y_1 + Z_1))\right) \tag{10}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left(f(\frac{1}{\sqrt{n}}(Y_i + X_i)) - f(\frac{1}{\sqrt{n}}(Y_i + Z_i))\right) \tag{11}$$

Now we note that for all i, Y_i is independent of X_i and Z_i, so

$$\mathbb{E}\left(f'(\frac{1}{\sqrt{n}}Y_i)(\frac{1}{\sqrt{n}}(X_i - Z_i))\right) = \mathbb{E}\left(f'(\frac{1}{\sqrt{n}}Y_i)\right)\mathbb{E}\left(\frac{1}{\sqrt{n}}(X_i - Z_i))\right). \tag{12}$$

The above term is zero because $\mathbb{E}(X_i) = \mathbb{E}(Z_i) = 0$. For a similar reason,

$$\mathbb{E}\left(f''(\frac{1}{\sqrt{n}}Y_i)(\frac{1}{\sqrt{n}}(X_i^2 - Z_i^2))\right) = 0. \tag{13}$$

Now we can bring in the control over the remaining terms:

$$|\mathbb{E}(f(W_n) - f(Z))| \leq \sum_{i=1}^{n} \Bigg|\mathbb{E}\Bigg(f(\frac{1}{\sqrt{n}}(Y_i + X_i)) - f(\frac{1}{\sqrt{n}}(Y_i + Z_i)) \tag{14}$$

$$- f'(\frac{1}{\sqrt{n}}Y_i)\frac{1}{\sqrt{n}}(X_i - Z_i) - \frac{1}{2}f''(\frac{1}{\sqrt{n}}Y_i)\frac{1}{n}(X_i^2 - Z_i^2)\Bigg)\Bigg| \tag{15}$$

$$\leq \sum_{i=1}^{n} \mathbb{E}\left(g(\frac{X_i}{\sqrt{n}}) + g(\frac{Z_i}{\sqrt{n}})\right) \tag{16}$$

$$= n\mathbb{E}\left(g(\frac{X_1}{\sqrt{n}})\right) + n\mathbb{E}\left(g(\frac{Z_1}{\sqrt{n}})\right) \tag{17}$$

We apply the estimate for g and get by decomposition of the integral

$$\mathbb{E}\left(g(\frac{X_1}{\sqrt{n}})\right) \leq K\left(\int_{|X_1|\leq\epsilon\sqrt{n}} |\frac{X_1}{\sqrt{n}}|^3 \mathrm{dP} + \int_{|X_1|>\epsilon\sqrt{n}} |\frac{X_1}{\sqrt{n}}|^2 \mathrm{dP}\right) \tag{18}$$

and notice that

$$\int_{|X_1|\leq\epsilon\sqrt{n}} |\frac{X_1}{\sqrt{n}}|^3 \mathrm{dP} \leq \frac{\epsilon}{n}\int_{|X_1|\leq\epsilon\sqrt{n}} X_1^2 \mathrm{dP} \leq \frac{\epsilon}{n} \tag{19}$$

So

$$n\mathbb{E}(g(\frac{X_1}{\sqrt{n}})) \leq K\left(\epsilon + \int_{|X_1|>\epsilon\sqrt{n}} X_1^2 \mathrm{dP}\right) \tag{20}$$

And $\{|X_1| > \epsilon\sqrt{n}\}$ decreases monotonically to an empty set. By the monotone convergence theorem, and by the arbitrary of $\epsilon > 0$, we can conclude $n\mathbb{E}(g(\frac{X_1}{\sqrt{n}}))$ tends to 0. For the case of Y, it is almost the same, and we just need to replace all X with Y. That finishes our proof.

2.2 Some Advantages of Lindeberg's Approach

Upon closer examination of the aforementioned proof, it becomes evident that the assumption of identically distributed random variables is not essential. Through the presented arguments, we unveil the second crucial observation made by Lindeberg, known as the *Lindeberg condition.* If the random variables are not distributed identically, we still have

$$|\mathbb{E}(f(W_n) - f(Z))| \leq \sum_{i=1}^{n} \mathbb{E}\left(g(\frac{X_i}{\sqrt{B_n}}) + g(\frac{Z_i}{\sqrt{B_n}})\right) \tag{21}$$

Here the Z_i are independent, normally distributed random variables with expectation value 0 and variance σ_i^2. Using the identical integral decomposition allows for the estimation of the following result:

$$\sum_{i=1}^{n} \mathbb{E}\left(g(\frac{X_i}{\sqrt{B_n}})\right) \leq K\sum_{i=1}^{n}\int_{|X_i|\leq\epsilon\sqrt{B_n}} |\frac{X_i}{\sqrt{B_n}}|^3 \mathrm{dP} + K\sum_{i=1}^{n}\int_{|X_i|>\epsilon\sqrt{B_n}} |\frac{X_i}{\sqrt{B_n}}|^2 \mathrm{dP}$$

The first term can now be estimated analogously by $K\epsilon$. For the second term, we have to demand that this expression converges towards zero for $n \to \infty$, and this is exactly the Lindeberg condition. If we then show that the Gaussian variables Z_i fulfill this condition, which we do not explain here because it is easy to compute and do some estimation and Gaussian random variables' tail density decays very fast, we have proven the Lindeberg Central Limit Theorem:

Theorem 1 (Lindeberg Central Limit Theorem, 1922)**.** *There are independent, real-value random variables* $X_1, X_2, \cdots$ *with* $\mathbb{E}(X_i) = 0$ *and* $\sigma_i^2 = \mathbb{E}(X_i^2) > 0$ *for each* i*. Assume* $Z \sim \mathscr{N}(0,1)$*. If*

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{i=1}^{n} \int_{|X_i|>\epsilon\sqrt{B_n}} X_i^2 \mathrm{dP} = 0 \tag{22}$$

applies for all $\epsilon > 0$*, then we have*

$$\frac{1}{B_n}\sum_{i=1}^{n} X_i \xrightarrow{D} Z. \tag{23}$$

It's worth mentioning that William Feller proved in [11] that a kind of inversion of Lindeberg's theorem applies. Assuming that

$$\max_{k\leq n} \sigma_k/B_n \to 0, \tag{24}$$

then Lindeberg Condition can be both sufficient and necessary.

An additional analysis of the preceding proof reveals that the Lindeberg method provides information regarding the convergence rate in the central limit theorem. In this regard, we will focus solely on the specific scenario of identically distributed random variables X_i. In fact, in his first work in 1920's [14], Lindeberg placed a stronger condition on the random variable X_i. He demanded the third absolute moment $\mathbb{E}|X_i|^3$ is finite for each X_i. Only in the works from 1922 [16] did he weaken this condition to the above condition.

To be more concise, if we look at the special case of identically distributed random variables and assume $\mathbb{E}|X_i|^3 \leq \infty$, then we can estimate the error term by

$$\mathbb{E}g(\frac{X_1}{\sqrt{n}}) \leq \frac{K}{\sqrt{n^3}}\mathbb{E}|X_1|^3 \tag{25}$$

and

$$|\mathbb{E}(f(W_n) - f(Z))| = K\mathcal{O}(\frac{1}{\sqrt{n}}\mathbb{E}|X_1|^3) \tag{26}$$

where K is larger than $\|f'''\|$ by our construction of g.

We notice that (26) is a weak version of a convergence rate in the sense of the theory of Berry and Esséen. Weak means that this result for the described test function class does not imply the supreme of difference between two distribution functions is also in the order of $n^{-1/2}$. However, it is known that the optimal convergence rate for the Kolmogorov distance in our situation is $\mathcal{O}(n^{-1/2})$. This is called the Berry Esséen rate. We will return to that in Chapter 3.

Furthermore, it is clear that by considering higher-order terms in the Taylor expansion of f and ensuring the finiteness and alignment of the higher moments of X_i with those of Z_i, we can improve the convergence rate. This aspect will play a crucial role in our exploration of the theory of random matrices and high-dimensional assumptions in subsequent chapters.

Ultimately, one may ponder why the central limit theorem exclusively yields the normal distribution as the limit distribution. It is apparent that when independent random variables themselves follow a normal distribution, their sum also follows a normal distribution. This property, referred to as the unlimited divisibility of the normal distribution in literature, is well-known. However, it is worth noting that there exists a class of distributions that share this property, including the normal distribution. So, what sets the normal distribution apart and renders it dominant in this context? It is important that with the above Z_i the random variable $\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$ is redistributed $\mathscr{N}(0,1)$, i.e. on the distribution level and the selected scale $\frac{1}{\sqrt{n}}$ the $\mathscr{N}(0,1)$ is a fixed point of the map $(X_1, \cdots, X_n) \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$. Its outstanding role results in a certain sense from the fact that it is the only such fixed point.

3 Martingale Central Limit Theorem

In modern probability theory, the term martingale has become quite essential. Initially, the term was used only to formalize the idea of a fair game. Many sequences of random variables are a martingale, such as the one-dimensional symmetrical error on Z, partial sums of independent random variables as considered in Chapter 2, products of independent random variables when the factors have expected value 1, sizes of populations in various branching models, etc. The core of the definition of a martingale is the concept of conditional expectation, we refer to a standard textbook Durrett's Chap. 4[10] for some background information.

Generally, $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space, and $\mathscr{G}$ is a sub-σ algebra of $\mathscr{F}$. Then $\mathbb{E}(X|\mathscr{G})$ to an integrable X denotes the random variable that is $\mathscr{F}$-measurable and

$$\int_A \mathbb{E}(X|\mathscr{G})\mathrm{dP} = \int_A X\mathrm{dP} \tag{27}$$

applies to all $A \in \mathscr{G}$. This random variable exists and is $\mathbb{P}$-almost certainly uniquely determined. Furthermore, for a sequence of random variables X_i, $\mathscr{F}_i = \sigma(X_1, \cdots, X_i)$ is the σ-algebra generated by the first i random variables. Then (X_i) is a martingale with respect to $\mathscr{F}_i$ if any X_i is integrable and $\mathbb{E}(X_{i+1}|\mathscr{F}_i) = X_i$ almost surely holds.

It follows immediately that $\mathbb{E}(X_{i+j}|\mathscr{F}_i) = X_i$ also applies to each $j \in \mathbb{N}$. The definition of martingale can be equivalent to the definition of martingale differences $\Delta_i = X_i - X_{i-1}$ that satisfies $\mathbb{E}(\Delta_i|\mathscr{F}_i) = 0$ $\mathbb{P}$-a.s.. So instead, we can construct a martingale by using a martingale difference sequence, i.e. if X_i is a martingale difference sequence with respect to $\mathscr{F}_i$ where $\mathscr{F}_i = \sigma(X_1, \cdots, X_i)$, then $S_n = \sum_{i=1}^n X_i$ is a martingale with respect to $\mathscr{F}_n$. We want to investigate the boundary behavior of the martingale B_n. There are several results in Peter Hall's book [12], for example, when some types of Lindeberg condition are satisfied and the martingale difference sequence has a consistent variance, then we have the asymptotic normality. We focus our interest on the Berry-Esséen type bound of Martingale by using Lindeberg replacement strategy, which was given by Bolthausen in [3]:

Theorem 2 (Bolthausen, 1982[3])**.** *For $0 < \alpha \leq \beta < \infty$, $0 < \gamma < \infty$, and X_i is a martingale difference sequence. Denote $\sigma_i^2 = \mathbb{E}(X_i^2|\mathscr{F}_{i-1})$ and $\bar{\sigma}_i^2 = \mathbb{E}(X_i^2)$, we assume $\sigma_i^2 = \bar{\sigma}_i^2$ a.s. hold for every i. Also, $B_n^2 := \sum_{i=1}^n \bar{\sigma}_i^2$. Assume $\alpha \leq \bar{\sigma}_i^2 \leq \beta$ and $\|X_i\|_3 \leq \gamma$ for all $1 \leq i \leq n$, then we have*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(\frac{S_n}{B_n} \leq t) - \Phi(t)| \leq Ln^{-1/4} \tag{28}$$

The outline of the presented proof suggests that it cannot achieve a convergence rate better than $n^{-1/4}$. However, compared to the rate for a partial sum of independent and identically distributed random variables, as discussed in Chapter 2, this rate is relatively weak. It is worth noting that Bolthausen has made an intriguing discovery, demonstrating that this rate is indeed optimal in the case of a martingale. Although we will not delve into the details here, as it diverges from our primary focus, this finding adds an interesting perspective to our main topics.

Proof. In addition to $X = (X_1, \cdots, X_n)$ given as in the sentence, we consider independent normally distributed centered random variables $Z_1, \cdots, Z_n, \xi$ with variances $\mathbb{E}(Z_j^2) = \sigma_j^2, \mathbb{E}(\xi^2) = \sqrt{n}$.

First of all, we will replace S_n/B_n with $S_n/B_n + \xi/B_n$, i.e. we will add a Gaussian variable ξ/B_n with small variance; according to the premise of the theorem, $\frac{1}{\beta\sqrt{n}} \leq \mathbb{E}(\xi/B_n)^2 \leq \frac{1}{\alpha\sqrt{n}}$, and intuitively, this small perturbation should not have a huge impact. Actually, there is a lemma from [3] that shows

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(\frac{S_n}{B_n} \leq t) - \Phi(t)| \leq 2 \sup_{t\in\mathbb{R}} |\mathbb{P}(\frac{S_n}{B_n} + \frac{\xi}{B_n} \leq t) - \Phi(t)| + cn^{-1/4}\alpha^{-1/2}. \tag{29}$$

The lemma states that the error can be bounded by $\|\mathbb{E}(\frac{\xi}{B_n})^2\|_\infty^{1/2}$. Now we push $\mathbb{P}(\frac{\sum_{i=1}^n Z_i}{B_n} + \frac{\xi}{B_n} \leq t)$ into it, and by the same lemma,

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(\frac{S_n}{B_n} \leq t) - \Phi(t)| \leq 2 \sup_{t\in\mathbb{R}} |\mathbb{P}(\frac{S_n}{B_n} + \frac{\xi}{B_n} \leq t) - \mathbb{P}(\frac{\sum_{i=1}^n Z_i}{B_n} + \frac{\xi}{B_n} \leq t)| + cn^{-1/4}\alpha^{-1/2}. \tag{30}$$

with some absolute constant c. Now it can be done by the famous trick from Lindeberg. For each $1 \leq k \leq n$, denote

$$U_k := \frac{\sum_{i=1}^{k-1} X_i}{B_n}, W_k := \frac{\sum_{i=k+1}^n Z_i}{B_n} + \frac{\xi}{B_n} \tag{31}$$

and split the difference into

$$\mathbb{P}(\frac{S_n}{B_n} + \frac{\xi}{B_n} \leq t) - \mathbb{P}(\frac{\sum_{i=1}^n Z_i}{B_n} + \frac{\xi}{B_n} \leq t) \tag{32}$$

$$= \sum_{k=1}^n (\mathbb{P}(U_k + W_k + X_k/B_n \leq t) - \mathbb{P}(U_k + W_k + Z_k/B_n \leq t)). \tag{33}$$

A brilliant step follows: Bolthausen takes advantage of the fact that W_k is normally distributed with expectation value 0 and variance $\lambda_k^2 = (\sum_{i=k+1}^n \bar{\sigma}_j^2)/B_n^2$, so W_k/λ_k is normally distributed. Since W_k is independent of U_k, X_k, and Z_k, the above sum can be written as

$$\sum_{k=1}^n (\mathbb{E}\Phi(\frac{t - U_k}{\lambda_k} - \frac{X_k}{\lambda_k B_n}) - \mathbb{E}\Phi(\frac{t - U_k}{\lambda_k} - \frac{Z_k}{\lambda_k B_n})) \tag{34}$$

Notice that Φ is basically a test function and we look at the Taylor expansion of the last sum as usual with some $0 \leq \theta_k, \theta_k' \leq 1$:

$$\sum_{k=1}^n \mathbb{E}\left(\left(-\frac{X_k}{\lambda_k B_n} + \frac{Z_k}{\lambda_k B_n}\right)\varphi\left(\frac{t - U_k}{\lambda_k}\right) + \left(\frac{X_k^2}{2\lambda_k^2 B_n^2} - \frac{Z_k^2}{2\lambda_k^2 B_n^2}\right)\varphi'\left(\frac{t - U_k}{\lambda_k}\right)\right. \tag{35}$$

$$\left. - \frac{X_k^3}{6\lambda_k^3 B_n^3}\varphi''\left(\frac{t - U_k}{\lambda_k} - \theta_k\frac{X_k}{\lambda_k B_n}\right) + \frac{Z_k^3}{6\lambda_k^3 B_n^3}\varphi''\left(\frac{t - U_k}{\lambda_k} - \theta_k'\frac{Z_k}{\lambda_k B_n}\right)\right) \tag{36}$$

In the case of independent random variables, the first two summands disappeared because of independence and our assumptions. Here, we can argue with conditional expectation values by regarding the first term as

$$\mathbb{E}\left(\mathbb{E}\left(\left(-\frac{X_k}{\lambda_k B_n}+\frac{Z_k}{\lambda_k B_n}\right)\varphi\left(\frac{t-U_k}{\lambda_k}\right)\mid \mathscr{F}_{k-1}\right)\right) \tag{37}$$

Now U_k is measurable with respect to $\mathscr{F}_{k-1}$, so it can be factored from the conditional expectation. The first summand disappears because $\mathbb{E}(X_k|\mathscr{F}_{k-1})=0$, and the second is obvious because of the independence. The same factorization argument can be used and the second summand, where $\mathbb{E}(X_k^2|\mathscr{F}_{k-1})=\sigma_k^2$, and it is almost surely equal to $\mathbb{E}Z_k^2=\bar{\sigma}_k^2$. Since φ and its derivatives are limited and $\max_{1\leq j\leq n}\|X_j\|_3\leq\gamma$ according to the premise, it follows overall

$$\sup_{t\in\mathbb{R}}\left|P\left(S_n/B_n\leq t\right)-\Phi(t)\right|\leq c\sum_{k=1}^{n}\lambda_k^{-3}B_n^{-3}+c'n^{-1/4} \tag{38}$$

for constants c, c', which depend only on $\alpha, \beta,$ and γ. Now we can see the choice of variance $\mathbb{E}\xi^2=\sqrt{n}$, because it guarantees the sum can be accurately controlled by $n^{-1/4}$.

□

Once again, a straightforward yet elegant proof of the central limit theorem is presented, specifically in the context of a martingale setting. Remarkably, this proof also yields an optimal convergence rate without requiring further justification. Few other proof methods exhibit such elegance, making it highly recommended as an extension of the Lindeberg Method. In fact, this can be seen as the foundation of the so-called 'implicit smoothing' method, which has been employed in numerous papers addressing high-dimensional Gaussian approximation problems.

4 Lindeberg Method in Random Matrix Theory

4.1 Chatterjee's Invariance Principle

The Lindeberg method's robustness in the face of changes in the limit distribution, as observed in Chapter 3, serves as the starting point for its far-reaching development in the last fifteen years, which has rightly been described as its renaissance. This resurgence can be traced back to the works of Chatterjee [4, 5]. Chatterjee's work makes a significant innovation by considering the boundary distribution of more general functions f of a random vector $X=(X_1,\cdots,X_n)$. This function can take the classical form $f(X)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i$, but it is not limited to this shape. Furthermore, even though the limit vector is Gaussian in our most important example, now there is no requirement for it to have a Gaussian structure. Chatterjee formulates the following statement.

Theorem 3 (Chatterjee, 2005[4]). *Denote* $\mathbf{X}=(X_1,\ldots,X_n)$ *and* $\mathbf{Y}=(Y_1,\ldots,Y_n)$ *are two independent vectors of independent random variables and satisfying that for* *each i, $\mathbb{E}X_i = \mathbb{E}Y_i$ and $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2 \leq \infty$. Let $\gamma = \max\{\mathbb{E}|X_i|^3, \mathbb{E}|Y_i|^3, 1 \leq i \leq n\} \leq \infty$.*

Let $f : \mathbb{R}^n \to \mathbb{R}$ be thrice differentiable in each argument. If we set $U = f(\mathbf{X})$ and $V = f(\mathbf{Y})$, then for any thrice differentiable $g : \mathbb{R} \to \mathbb{R}$ and any $K > 0$,

$$|\mathbb{E}g(U) - \mathbb{E}g(V)| \leq C_1(g)\lambda_2(f)\sum_{i=1}^n[\mathbb{E}(X_i^2; |X_i| > K) + \mathbb{E}(Y_i^2; |Y_i| > K)] \tag{39}$$

$$+ C_2(g)\lambda_3(f)\sum_{i=1}^n[\mathbb{E}(|X_i|^3; |X_i| \leq K) + \mathbb{E}(|Y_i|^3; |Y_i| \leq K)] \tag{40}$$

where $C_1(g) = \|g'\|_\infty + \|g''\|_\infty$ and $C_2(g) = \frac{1}{6}\|g'\|_\infty + \frac{1}{2}\|g''\|_\infty + \frac{1}{6}\|g'''\|_\infty$, and

$$\lambda_r(f) := \sup\{|\partial_i^p f(\mathbf{x})|^{r/p} : 1 \leq i \leq n, 1 \leq p \leq r, \mathbf{x} \in I^n\}. \tag{41}$$

As mentioned in Chapter 2, if we can choose a suitable K, we use $\mathbb{E}(|X_i|^3; |X_i| \leq K) \leq K\mathbb{E}(X_i^2)$ to control the error.

The function f in this sentence plays the same role as in Chapter 2. If you choose $f(X) = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$, then it is easy to compute $\lambda_2(f) = n^{-1}$ and $\lambda_3(f) = n^{-3/2}$.If X_i's and Y_i's are i.i.d., and $\mathbb{E}X_i = \mathbb{E}Y_i = 0$ and $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2 = 1$ for all i, then we choosing $K = \epsilon\sqrt{n}$ and by Theorem 3:

$$|\mathbb{E}g(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i) - \mathbb{E}g(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i)| \leq C_1(g)[\mathbb{E}(X_1^2; |X_1| > \epsilon\sqrt{n}) \tag{42}$$

$$+ \mathbb{E}(Y_1^2; |Y_1| > \epsilon\sqrt{n})] + 2C_2(g)\epsilon \tag{43}$$

As the number of terms, denoted by n, tends to infinity, the aforementioned result establishes the classical Central Limit Theorem (CLT) since it holds true for all values of ϵ greater than zero. In addition to this fundamental result, if we have the additional conditions that $\mathbb{E}|X_1|^3 \vee \mathbb{E}|Y_1|^3 < \infty$, we can derive an explicit error bound as well:

$$|\mathbb{E}g(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i) - \mathbb{E}g(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i)| \leq \frac{C_2(g)[\mathbb{E}|X_1|^3 + \mathbb{E}|Y_1|^3]}{\sqrt{n}} \tag{44}$$

Now we turn to the proof of this theorem.

Proof. We choose f and g to be fixed, let $h = g \circ f$. By the chain rule,

$$\partial_i^2 h(\mathbf{x}) = g'(f(\mathbf{x}))\partial_i^2 f(\mathbf{x}) + g''(f(\mathbf{x}))(\partial_i f(\mathbf{x}))^2, \tag{45}$$

$$\partial_i^3 h(\mathbf{x}) = g'(f(\mathbf{x}))\partial_i^3 f(\mathbf{x}) + 3g''(f(\mathbf{x}))\partial_i f(\mathbf{x})\partial_i^2 f(\mathbf{x}) + g'''(f(\mathbf{x}))(\partial_i f(\mathbf{x}))^3. \tag{46}$$

So for every i and $\mathbf{x}$, $|\partial_i^2 h(\mathbf{x})| \leq C_1\lambda_2(f)$ and $|\partial_i^3 h(\mathbf{x})| \leq 6C_2\lambda_3(f)$, where $C_1 = \|g'\|_\infty + \|g''\|_\infty$ and $C_2 = \frac{1}{6}\|g'\|_\infty + \frac{1}{2}\|g''\|_\infty + \frac{1}{6}\|g'''\|_\infty$.

Second, for $0 \le i \le n$, we define $\mathbf{Z}_i := (X_1, \ldots, X_{i-1}, X_i, Y_{i+1}, \ldots, Y_n)$ and $\mathbf{W}_i := (X_1, \ldots, X_{i-1}, 0, Y_{i+1}, \ldots, Y_n)$. For $1 \le i \le n$, define

$$R_i := h(\mathbf{Z}_i) - X_i \partial_i h(\mathbf{W}_i) - \frac{1}{2} X_i^2 \partial_i^2 h(\mathbf{W}_i), \tag{47}$$

$$T_i := h(\mathbf{Z}_{i-1}) - Y_i \partial_i h(\mathbf{W}_i) - \frac{1}{2} Y_i^2 \partial_i^2 h(\mathbf{W}_i). \tag{48}$$

By the bounds on the third partials of h, and using Taylor Expansion to the third order, it is obvious that $|R_i| \le C_2 \lambda_3(f) |X_i|^3$ and $|T_i| \le C_2 \lambda_3(f) |Y_i|^3$. Also, we can use the second order term to bound them by $|R_i| \le C_1 \lambda_2(f) |X_i|^2$ and $|T_i| \le C_1 \lambda_2(f) |Y_i|^2$. For every i, X_i, Y_i and $\mathbf{W}_i$ are independent. So

$$\mathbb{E}(X_i \partial_i f(\mathbf{W}_i)) - \mathbb{E}(Y_i \partial_i f(\mathbf{W}_i)) = \mathbb{E}(X_i - Y_i) \mathbb{E}(\partial_i f(\mathbf{W}_i)) = 0. \tag{49}$$

Similarly, $\mathbb{E}(X_i^2 \partial_i^2 f(\mathbf{W}_i)) - \mathbb{E}(Y_i^2 \partial_i^2 f(\mathbf{W}_i)) = 0$. Combining all these results, for every $K > 0$,

$$\begin{aligned}
|\mathbb{E}g(U) - \mathbb{E}g(V)| &= \left| \sum_{i=1}^n \mathbb{E}(h(\mathbf{Z}_i) - h(\mathbf{Z}_{i-1})) \right| \\
&= \Bigg| \sum_{i=1}^n \mathbb{E}(X_i \partial_i h(\mathbf{W}_i) + \frac{1}{2} X_i^2 \partial_i^2 h(\mathbf{W}_i) + R_i) \\
&\qquad - \sum_{i=1}^n \mathbb{E}(Y_i \partial_i h(\mathbf{W}_i) + \frac{1}{2} Y_i^2 \partial_i^2 h(\mathbf{W}_i) + T_i) \Bigg| \\
&\le C_1 \lambda_2(f) \sum_{i=1}^n \left[\mathbb{E}(X_i^2; |X_i| > K) + \mathbb{E}(Y_i^2; |Y_i| > K) \right] \\
&\quad + C_2 \lambda_3(f) \sum_{i=1}^n \left[\mathbb{E}(|X_i|^3; |X_i| \le K) + \mathbb{E}(|Y_i|^3; |Y_i| \le K) \right].
\end{aligned}$$

□

It is important to consider the benefits of such a generalization and the insights it can provide. One particular application of interest lies in exploring Theorem 3, which intersects with the field of random matrix theory, a prominent area within modern probability theory. By delving into this application, we aim to deepen our understanding and uncover the implications within this specialized domain.

4.2 Pastur's Condition for Wigner's semicircle law

A random matrix is a matrix that entries are real or complex random variables. We focus on symmetric $N \times N$ matrix M_N. In this setting, we have a symmetric $M_N = (X_{i,j})_{1 \le i,j \le N}$, where $X_{i,j}, i \le j$ are i.i.d. random variables and $X_{j,i} = X_{i,j}, i \le j$. Obviously, their eigenvalues are all real and we can ask about their distribution. For the same reason as in the classical central limit theorem, we usually consider the scaled matrix $\frac{1}{\sqrt{N}}M_N$.

We define the empirical spectral distribution (ESD) of the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ of $\frac{1}{\sqrt{N}}M_N$ as

$$F_N(t) = \frac{1}{N}\#\{i : \lambda_i \leq t\}$$

In his work [21], Wigner examined Bernoulli distributed variables $X_{i,j}$ and demonstrated that F_N converges in probability to the renowned *semicircle law* with a density function of $\frac{1}{2\pi}\sqrt{4-x^2}$ on the interval $[-2,2]$. He further recognized in [22] that this result holds not only for normally distributed $X_{i,j}$ but also for other distributions. The validity of the semicircle law for independent and identically distributed $X_{i,j}$ (under appropriate moment conditions) was established by Arnold [1]. In the language of physicists, the universality of the semicircle law was demonstrated by first identifying laws applicable to specific distributions of $X_{i,j}$, typically assumed to be normally distributed, and then investigating whether these laws depend on the specific distribution. Several studies have explored scenarios involving dependent $X_{i,j}$. Our objective is to investigate the conditions under which the Wigner semicircle law can be derived as the limit distribution of F_N from the $X_{i,j}$. Through the application of Theorem 3, we aim to uncover a suitable Lindeberg condition in this context.

A standard tool for identifying the limiting spectral distribution, i.e. LSD, of a sequence of random matrices is the Stieltjes transform. To be short, we can say that the ESDs of a sequence $\{A_N\}_{N=1}^{\infty}$ of random real symmetric matrices converge in probability to a probability distribution G if and only if

$$\forall z \in \mathbb{C}\backslash\mathbb{R}, \ \frac{1}{N}\operatorname{Tr}((A_N - zI_N)^{-1}) \xrightarrow{P} \int_{-\infty}^{\infty} \frac{1}{x-z} dG(x)$$

where I_N is the identity matrix of order N. The expression on the right corresponds to the Stieltjes transform of G at the point z, while the expression on the left corresponds to the Stieltjes transform of the empirical spectral distribution (ESD) of A_N at z. The Stieltjes transform plays a crucial role in our technique due to its remarkable smoothness as a function of the matrix entries. This smoothness property makes it highly advantageous for our analysis and application.

By employing the aforementioned techniques, we will showcase the effectiveness of Pastur's condition, which is recognized as the least stringent criterion known for establishing convergence to the semicircle law. Through our analysis, we will demonstrate the sufficiency of this condition and its significance in characterizing convergence behavior.

To begin, let $z = u + iv \in \mathbb{C}$, where $v \neq 0$. Denote $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\mathbf{x}) = \frac{1}{N}\operatorname{Tr}((A(\mathbf{x}) - zI)^{-1}).$$

And $G : \mathbb{R}^n \to \mathbb{C}^{N\times N}$, $G(\mathbf{x}) := (A(\mathbf{x}) - zI)^{-1}$. Note that all eigenvalues of $A(\mathbf{x}) \in \mathbb{R}$, so $\det(A(\mathbf{x}) - zI) \neq 0$. Furthermore, it is worth noting that in matrix theory, when we perform matrix inversion, we compute the classical adjoint and divide it by the determinant. This process entails expressing the elements of the inverse as rational functions of the elements of the original matrix. Consequently, the Stieltjes transform G exhibits smoothness along each coordinate, owing to the rational nature of its elements. This smoothness property is instrumental in our analysis and contributes to the effectiveness of our approach.

Because $(A(\mathbf{x}) - zI)G(\mathbf{x}) = I$ for every $\mathbf{x}$, so $1 \leq i \leq j \leq N$, $\frac{\partial}{\partial x_{i,j}}[(A - zI)G] \equiv 0$, then

$$\frac{\partial G}{\partial x_{i,j}} = -G\frac{\partial A}{\partial x_{i,j}}G. \tag{50}$$

Similarly, higher order derivatives have the same property and take trace we have

$$\begin{aligned}
\frac{\partial f}{\partial x_{i,j}} &= -\frac{1}{N}\operatorname{Tr}(\frac{\partial A}{\partial x_{i,j}}G^2), &(51)\\
\frac{\partial^2 f}{\partial x_{i,j}^2} &= \frac{2}{N}\operatorname{Tr}(\frac{\partial A}{\partial x_{i,j}}G\frac{\partial A}{\partial x_{i,j}}G^2), &(52)\\
\frac{\partial^3 f}{\partial x_{i,j}^3} &= -\frac{6}{N}\operatorname{Tr}(\frac{\partial A}{\partial x_{i,j}}G\frac{\partial A}{\partial x_{i,j}}G\frac{\partial A}{\partial x_{i,j}}G^2). &(53)
\end{aligned}$$

Now we need to bound these terms. For an $N \times N$ complex matrix $B = ((b_{i,j}))$, the Frobenius norm of B is $\|B\| := (\sum_{i,j} |b_{i,j}|^2)^{1/2}$. It's easy to see that this norm satisfies these canonical properties:

1. $|\operatorname{Tr}(BC)| \leq \|B\|\|C\|$.
2. $\|\cdot\|$ has unitary invariance.
3. For a normal matrix B and its eigenvalues $\lambda_1, \ldots \lambda_N$, $\max\{\|BC\|, \|CB\|\} \leq \max_{1\leq i\leq N} |\lambda_i| \cdot \|C\|$ holds for any C.

Note that G as well as the derivatives of A are normal matrices, and the eigenvalues of G are bounded by $|v|^{-1}$ where $v = \operatorname{Im} z$. Also, the eigenvalues of $\partial A/\partial x_{i,j}$ are bounded by $N^{-1/2}$, because $\partial A/\partial x_{i,j}$ is the matrix which only has non-zero term $N^{-1/2}$ at the (i,j) and (j,i) positions.

So the elements of G^2 are bounded by $|v|^{-2}$. As a result,

$$\left\|\frac{\partial f}{\partial x_{i,j}}\right\|_\infty \leq 2|v|^{-2}N^{-3/2}. \tag{54}$$

Similarly,

$$\left\|\frac{\partial^2 f}{\partial x_{i,j}^2}\right\|_\infty \leq \frac{2}{N}\left\|\frac{\partial A}{\partial x_{i,j}}\right\|\left\|G\frac{\partial A}{\partial x_{i,j}}G^2\right\| \leq 4|v|^{-3}N^{-2}. \tag{55}$$

And,

$$\left\|\frac{\partial^3 f}{\partial x_{i,j}^3}\right\|_\infty \leq 12|v|^{-4}N^{-5/2}. \tag{56}$$

In summary,

$$\lambda_2(f) \leq 4\max\{|v|^{-4}, |v|^{-3}\}N^{-2}, \tag{57}$$

$$\lambda_3(f) \leq 12\max\{|v|^{-6}, |v|^{-4}\}N^{-5/2}. \tag{58}$$

Let $\mathbf{X} = (X_{i,j})_{1\leq i\leq j\leq N}$ and $\mathbf{Y} = (Y_{i,j})_{1\leq i\leq j\leq N}$ which entries are independent centered random variables with variance 1, we call this as normalization. Denote $U = \operatorname{Re} f(\mathbf{X})$, $V = \operatorname{Re} f(\mathbf{Y})$, and $g : \mathbb{R} \to \mathbb{R}$ be a thrice differentiable function. Now $\operatorname{Re} f$ is a smooth function and $\lambda_r(\operatorname{Re} f) \leq \lambda_r(f)$ for each r. Take $K = \epsilon\sqrt{N}$, Theorem 3 can tell us that $|\mathbb{E}g(U) - \mathbb{E}g(V)|$ can be bounded by a constant rate depending only on g and v of

$$N^{-2} \sum_{1\leq i\leq j\leq N} [\mathbb{E}(X_{i,j}^2; |X_{i,j}| > \epsilon\sqrt{N}) + \mathbb{E}(Y_{i,j}^2; |Y_{i,j}| > \epsilon\sqrt{N})] + \epsilon. \tag{59}$$

The imaginary part is similar.

According to Wigner's Theorem for Gaussian scenarios, we observe that convergence to the semicircle law can occur when the random variables $X_{i,j}$ are independent, centered, normalized, and satisfy the condition

$$\forall\epsilon > 0, \lim_{N\to\infty} N^{-2} \sum_{1\leq i\leq j\leq N} \mathbb{E}(X_{i,j}^2; |X_{i,j}| > \epsilon\sqrt{N}) = 0. \tag{60}$$

This condition precisely corresponds to Pastur's condition. It is satisfied, for instance, when the random variables $X_{i,j}$ are independent and identically distributed (i.i.d.), centered, and normalized. It is important to note that although it may resemble Lindeberg's condition for the central limit theorem, it is not identical.

It is important to highlight that Theorem 3 has other intriguing applications, such as in the theory of spin glasses and the study of the Sherrington-Kirkpatrick model. These models hold great significance in modern probability theory. However, due to the length constraints of this discussion, we will not delve into these applications here.

5 Moment Conditions in Random Matrix Theory

We recall the situation in Chapter 4. We have a symmetric matrix $M_N := (X_{i,j}/\sqrt{N})_{1\leq i,j\leq N}$ where $X_{i,j}$ for $i \leq j$ are independent random variables. The semicircle law provides a universal limit distribution for the global statistics $F_N(t) := \frac{1}{N}\#\{i : \lambda_i \leq t\}$ under moment conditions.

Since the inception of random matrix analysis, numerous other statistics derived from the random eigenvalues λ_i have been explored. One example is the distribution of gaps between consecutive eigenvalues, which examines the frequency of occurrences where $\lambda_{i+1} - \lambda_i \leq s$ for $1 \leq i \leq N$. Another area of interest is the correlation among k eigenvalues in the limit, known as the k-point correlation function. Questions regarding the distribution of individual eigenvalues and the limit distribution of the joint distribution of k eigenvalues $(\lambda_{i_1}, \cdots, \lambda_{i_k})$ also arise. These statistics fall under the category of local eigenvalue statistics, and their investigation tends to be more complex than that of global statistics. Global statistics encompass quantities such as the determinant of the matrix M_N or the count of eigenvalues λ_i that fall within a given interval.

We now consider a $N \times N$ Wigner Hermitian matrix $W_N = (X_{i,j})_{1\leq i,j\leq N}$, which is a Hermitian matrix with independent entries $X_{i,j}$ and $X_{j,i} = \bar{X}_{i,j}$. For $i < j$, the entries $X_{i,j}$ are identically distributed and follow a centered and normalized complex value distribution. When $i = j$, the entries $X_{i,i}$ are equally distributed with an expectation value of 0 and a variance of σ^2. The real and imaginary parts of the random variables are independent. Additionally, there exists a constant C independent of i, j, and N such that $\mathbb{E}|X_{i,j}|^C \leq D$ for every i and j, where D is another constant independent of i, j, and N. To ensure that the eigenvalues of the matrix are within the interval $[-2, 2]$, we scale the matrix by $M_N = \frac{1}{\sqrt{N}} W_N$. This scaling preserves the constrained interval for the eigenvalues. Alternatively, if we scale the matrix by $A_N = \sqrt{N} W_N$, the spacing between two eigenvalues remains roughly constant. It is important to note that a special case of the Wigner Hermitian matrix arises when the entries are normally distributed. In this case, the diagonal entries are real random variables following $\mathscr{N}(0, 1)$ distribution, while the off-diagonal entries are complex random variables following $\mathscr{N}(0, 1/2) + i\mathscr{N}(0, 1/2)$ distribution. This ensemble is known as the Gaussian Unitary Ensemble (GUE) because the distribution of W_N remains invariant under unitary matrix conjugation. The main advantage of the GUE is that the common distribution of unordered eigenvalues can be expressed by Ginibre's formula

$$\varrho(\lambda_1, \cdots, \lambda_N) = Z_N^{-1} \prod_{1\leq i<j\leq N} |\lambda_i - \lambda_j|^2 \exp(-\frac{1}{2}\sum_{i=1}^{N} \lambda_i^2) \tag{61}$$

where Z_N is a suitable standardization so-called partition function. In recent years, several studies have utilized Ginibre's formula to explore the limiting distribution of different local statistics associated with the eigenvalues of the Gaussian Unitary Ensemble (GUE). These investigations have employed various techniques, such as determinant point processes and orthogonal polynomial methods, to analyze the distribution. Although the details of these techniques are beyond the scope of this discussion, they have played a significant role in advancing our understanding of the GUE's eigenvalue statistics.

Here are some notable results for GUE matrices. The smallest eigenvalue of a GUE matrix, denoted by $\lambda_1(M_N)$, converges to the Tracy-Widom distribution. More precisely, $(\lambda_1(M_N) + 2)N^{2/3}$ converges to this distribution, which is a well-known probability distribution. Let $N_I(M_N)$ be the number of eigenvalues of the GUE matrix M_N that lie in the interval I. If $\mathrm{Var}(N_I(M_N))$ tends to infinity as $N \to \infty$, then the standardized random variable $(N_I(M_N) - \mathbb{E}(N_I(M_N)))/\sqrt{\mathrm{Var}(N_I(M_N))}$ approaches a standard normally distributed random variable. Locally, for a sequence $k(N)$ such that $k(N)/N$ converges to a constant c in the range $(0, 1)$ as $N \to \infty$, the random variable $(\lambda_k(N) - \alpha(k(N)))/\beta(k(N))$ converges to a standard normally distributed random variable. Here, $\alpha(k(N))$ and $\beta(k(N))$ are suitable scales that indicate the expected location and standard deviation of $\lambda_k(N)$, respectively. These results provide valuable insights into the behavior of eigenvalues in GUE matrices and their convergence to specific probability distributions.

For a considerable time, there has been a belief that the aforementioned distribution laws, along with many others, hold universally not only for Wigner Hermitian matrices but also for broader classes of matrices. Proving this universality has been a challenging task, leading to numerous mathematically intricate works. However, in this context, we should highlight the significant contribution made by Tao and Vu in their groundbreaking work [20], where they were able to address certain questions regarding universality. Remarkably, their work builds upon the foundations of the Lindeberg Method, showcasing its enduring relevance and applicability in advancing our understanding of these complex problems.

Consider two independent Wigner Hermitian matrices M_N and M'_N. For various statistics F, $\mathbb{E}(F(M_N)) - \mathbb{E}(F(M'_N))$ should be relatively small. Consider the matrix $\tilde{M}_N$ formed from M_N by replacing either one of the diagonal entries $X_{i,i}$ of M_N by the corresponding entry $X'_{i,i}$ of M'_N, or one of the non-diagonal entries $X_{i,j}$ of M_N is replaced by the corresponding entry $X'_{i,j}$ of M'_N (and thus $X_{j,i}$ by $X'_{j,i}$). If it can now be shown that $\mathbb{E}(F(M_N)) - \mathbb{E}(F(\tilde{M}_N)) = \mathrm{o}(1/n)$ uniformly when replacing a diagonal element and $\mathbb{E}(F(M_N)) - \mathbb{E}(F(\tilde{M}_N)) = \mathrm{o}(1/n^2)$ uniformly when replacing a non-diagonal element, the Lindeberg replacement approach would show that $\mathbb{E}(F(M_N)) - \mathbb{E}(F(M'_N)) = \mathrm{o}(1)$.

In Chapter 2, we introduced a GUE matrix M'_N to gradually replace the elements of the original matrix with Gaussian-distributed elements. Building on that, Chapter 4 suggests that the same result holds for a general matrix M'_N belonging to the same matrix class as M_N. The key insight provided by Tao and Vu's fourth moments theorem is that by matching the first four moments of the matrix entries, we can achieve the desired convergence. But why specifically four moments?

To understand this heuristic choice, let's refer back to Chapter 2.1. There, we assumed that the first two moments of the sum X_i with a $\mathscr{N}(0,1)$ distribution were matched, and we observed an error of size $\mathcal{O}(1/N^{3/2})$ when exchanging a single element. Consequently, performing this exchange n times led to an error rate of $\mathcal{O}(1/N^{1/2})$. For each subsequent moment of agreement with $\mathscr{N}(0,1)$, we obtained an improvement in the error rate by $\mathcal{O}(1/N^{1/2})$. Therefore, if we were to achieve agreement up to the fourth moment, replacing a term would introduce an error of size $\mathcal{O}(1/N^{5/2})$. Since we are replacing approximately N^2 terms with matrices, this level of agreement would be sufficient to obtain an overall bound of $\mathrm{o}(1)$. Hence, it is not surprising that the entries of M_N and M'_N coincide in moments up to the second order on the diagonal and up to the fourth order off the diagonal.

We first give a precise definition of the correspondence of moments:

Definition 1 (moment matching)**.** *Two complex random variables ξ and ξ' match to order k if for every $m, l \geq 0$ s.t. $m + l \leq k$,*

$$\mathbb{E}\mathrm{Re}(\xi)^m\mathrm{Im}(\xi)^l = \mathbb{E}\mathrm{Re}(\xi')^m\mathrm{Im}(\xi')^l. \tag{62}$$

Definition 2 (Condition **C0**)**.** *A random Hermitian matrix $M_N = (X_{i,j})_{1\leq i,j\leq N}$ is said to obey* condition **C0** *if*

- *The $X_{i,j}$ are independent (but not necessarily identically distributed) for $1 \leq i \leq j \leq N$, and have mean zero and variance* 1.
- *(Uniformly sub-exponential) There exist constants $C, C' > 0$ s.t.*

$$\mathbb{P}(|X_{i,j}| \geq t^C) \leq \exp(-t) \tag{63}$$

for all sufficiently large $t \geq C'$ and $1 \leq i, j \leq N$.

Then our main theorem comes.

Theorem 4 (Four Moment Theorem, Tao and Vu [20])**.** *Let $M_N = (X_{i,j})_{1 \leq i,j \leq N}$ and $M'_N = (X'_{i,j})_{1 \leq i,j \leq N}$ be two random matrices satisfying* **C0**. *We assume that for any $1 \leq i < j \leq N$, the entries $X_{i,j}$ and $X'i, j$ match up to the fourth order, and for any $1 \leq i \leq N$, the entries $X_{i,i}$ and $X'_{i,i}$ match up to the second order. Let $A_N = \sqrt{N} M_N$ and $A'_N = \sqrt{N} M'_N$. We introduce a small positive constant c_0, and for any $0 < \epsilon < 1$ and $k \geq 1$, the following condition holds: If $G : \mathbb{R}^k \to \mathbb{R}$ is a smooth function that satisfies the derivative bounds*

$$|\nabla^j G(x)| \leq N^{c_0} \tag{64}$$

for all $0 \leq j \leq 5$ and $x \in \mathbb{R}^k$. Then for any $\epsilon N \leq i_1 < i_2 \cdots < i_k \leq (1-\epsilon)N$, and for N sufficiently large depending on ϵ, k, we have

$$|\mathbb{E}(G(\lambda_{i_1}(A_N), \dots, \lambda_{i_k}(A_N))) - \mathbb{E}(G(\lambda_{i_1}(A'_N), \dots, \lambda_{i_k}(A'_N)))| \leq N^{-c_0}. \tag{65}$$

An overview of a whole class of various fourth moment theorems can be found in [20]. We outline the major approach. We form the matrix $\tilde{M}_N$ of M_N by taking a single entry $X_{i,j}, i < j$ of M_N replaced by the entry $X'_{i,j}$ of M'_N, and the corresponding location $X_{j,i}$, to keep $\tilde{M}_N$ hermitian. One important technical observation is that $\tilde{M}_N$ is no longer a Wigner matrix because the entries are not necessarily identically distributed. We look at $\tilde{A}_N = \sqrt{N}\tilde{M}_N$ and want to proof

$$\mathbb{E}(G(\lambda_{i_1}(A_N), \dots, \lambda_{i_k}(A_N))) - \mathbb{E}(G(\lambda_{i_1}(\tilde{A}_N), \dots, \lambda_{i_k}(\tilde{A}_N))) = \mathcal{O}(N^{-5/2+\mathcal{O}(c_0)})$$

We now think $A_N = A(X_{i,j})$ and $\tilde{A}_N = \tilde{A}(X'_{i,j})$ as functions of $X_{i,j}$ and $X'_{i,j}$. We have $A(t) := A(0) + t\mathbf{e}_i\mathbf{e}_j^* + \bar{t}\mathbf{e}_j\mathbf{e}_i^*$, where $A(0)$ is a Wigner matrix where one entry and its adjoined entry is zero, $\mathbf{e}_i$ is the canonical orthonormal basis of $\mathbb{C}^N$. Note that A_N is a matrix amplified with a scale number $\sqrt{N}$, so we concern the case $t = \mathcal{O}(N^{1/2+\mathrm{o}(1)})$. We claim

$$F(t) = \mathbb{E}(G(\lambda_{i_1}(A(t)), \dots, \lambda_{i_k}(A(t))))$$

And as a result, we want to have

$$\mathbb{E}F(X_{i,j}) - \mathbb{E}F(X'_{i,j}) = \mathcal{O}(N^{-5/2+\mathcal{O}(c_0)})$$

So the problem turns to that how does the exchange of a matrix element change its eigenvalue. Suppose that we have Taylor expansions of the form

$$\lambda_{i_l}(A(t)) = \lambda_{i_l}(A(0)) + \sum_{j=1}^{4} c_{l,j} t^j + \mathcal{O}(n^{-5/2+\mathcal{O}(c_0)}) \tag{66}$$

for all $t = \mathcal{O}(n^{\mathcal{O}(c_0)})$ and $l = 1, \dots, k$, where the Taylor coefficients $c_{l,j}$ have size $c_{l,j} = \mathcal{O}(n^{-j/2+\mathcal{O}(c_0)}$. Then by Taylor expansion and the gradient control, we have

$$F(t) = F(0) + \sum_{j=1}^{4} f_j t^j + O(n^{-5/2+O(c_0)})$$

for the function $F(t)$, where the coefficients f_j have size $f_j = \mathcal{O}(n^{-j/2+\mathcal{O}(c_0)})$. Setting t equal to $X_{i,j}$ and taking expectations, and noting that the Taylor coefficients f_j depend only on F and $A(0)$ and is thus independent of $X_{i,j}$, we conclude that

$$\mathbb{E}F(X_{i,j}) = \mathbb{E}F(0) + \sum_{k=1}^{4} (\mathbb{E}f_j)(\mathbb{E}X_{i,j}^k) + \mathcal{O}(n^{-5/2+\mathcal{O}(c_0)})$$

and similarly for $\mathbb{E}F(X'_{i,j})$. If $X_{i,j}$ and $X'_{i,j}$ have matching moments to fourth order, this gives the result we want.

It remains to establish (66). We abbreviate i_l simply as i. By Taylor's theorem, it would suffice to show that

$$\frac{d^j}{dt^j} \lambda_i(A(t)) = \mathcal{O}(n^{-j/2+\mathcal{O}(c_0)}) \tag{67}$$

for $j = 1, \dots, 5$. There remains some technically great details. This is where the real difficulty lies and is overcome with the help of Hadamard's variation formulas. If $\lambda_i(A(t))$ denotes the i-th vector of an orthogonal base of eigenvectors of $A(t)$, the following applies:

$$\frac{d}{dt} \lambda_i(A(t)) = u_i(A(t))^* A'(0) u_i(A(t)).$$

$$\frac{d^2}{dt^2} \lambda_i(A(t)) = -2 \sum_{j \neq i} \frac{|u_i(A(t))^* A'(0) u_j(A(t))|^2}{\lambda_j(A(t)) - \lambda_i(A(t))}$$

The results on delocalization indicate that in the last expression, the numerator is at least of the order $\mathcal{O}(n^{-1+o(1)})$ with a very high probability. However, to evaluate the denominator, we need to establish that the eigenvalues of M_N are almost surely distinct. This relies on the presence of a gap property among the eigenvalues and a localized version of the semicircle law. It's important to note that the formulas for the higher derivatives of $\lambda_i(A(t))$ become significantly more complex in this context. Details on this have been elaborated in [20].

Using appropriate choices of test functions, Tao and Vu successfully established that the limit distributions, such as the Tracy-Widom distribution for the smallest eigenvalue, observed in GUE matrices can be extended to all Wigner-Hermite matrices. Furthermore, they have made significant progress in deriving other universal fluctuations, including moderate deviations principles for individual eigenvalues, the relative number of eigenvalues within an interval, and the determinant of Wigner-Hermite matrices. While we won't delve into the specifics here, these advancements have expanded our understanding of random matrices and their statistical properties.

6 A New Approach in High Dimensional CLT

6.1 Breakthrough in High Dimensional settings

We now consider the case that p is much much larger than n. $X_1, \dots, X_n$ are independent random vectors in $\mathbb{R}^p$ where $p \geq 3$. Assume $X_i = (X_{i1}, \dots, X_{ip})'$, every X_i is centered and $\mathbb{E}[X_{ij}^2] < \infty$ for all i,j. Consider the normalized sum

$$S_n^X := (S_{n1}^X, \dots, S_{np}^X)' := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i. \tag{68}$$

Let $Y_1, \dots, Y_n$ be independent centered Gaussian r.v.s in $\mathbb{R}^p$ s.t. every Y_i has the same covariance matrix as X_i. Similarly,

$$S_n^Y := (S_{n1}^Y, \dots, S_{np}^Y)' := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i. \tag{69}$$

We consider bounding the below quantity as a distance of distribution

$$\rho_n(\mathcal{A}) := \sup_{A \in \mathcal{A}} |\mathbb{P}(S_n^X \in A) - \mathbb{P}(S_n^Y \in A)|, \tag{70}$$

where $\mathcal{A}$ is a class of Borel sets in $\mathbb{R}^p$. There has been extensive research on bounding $\rho_n(\mathcal{A})$, with a particular focus on explicitly capturing the dependence on p in the bounds. The main interest lies in understanding the rate at which $p = p_n \to \infty$ can grow while ensuring that $\rho_n(\mathcal{A})$ tends to zero. In particular, Bentkus [2] established one of the most well-known results which says that if $X_1, \dots, X_n$ are i.i.d. with $\mathbb{E}[X_i X_i'] = I$,

$$\rho_n(\mathcal{A}) \leq C_p(\mathcal{A}) \frac{\mathbb{E}[\|X_1\|^3]}{\sqrt{n}}, \tag{71}$$

where $C_p(\mathcal{A})$ is a constant depends only on p and $\mathcal{A}$. In a sense, this can be seen as a high-dimensional counterpart of the Berry-Esseen Theorem. However, what often frustrates and confuses us is that the best classical results only hold when p grows at most polynomially with respect to n. Breaking through this limitation, Chernozhukov et al. in [7] made significant progress in overcoming the so-called "curse of large dimension." They demonstrated in their paper that Gaussian approximation is achievable for normalized maximum statistics, even when $\log(p)$ grows at a polynomial rate with respect to n.

For the sake of brevity, we will focus mainly on the core conclusions, leaving most of the remaining theorems unmentioned. We think the key lemma of [7] is of vital importance.

Define

$$\varrho_n := \sup_{y\in\mathbb{R}^p, v\in[0,1]} |\mathbb{P}\left(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y \leq y\right) - \mathbb{P}(S_n^Y \leq y)|, \tag{72}$$

where the r.v.s $Y_1,\ldots,Y_n$ are independent of the random vectors $X_1,\ldots,X_n$, and recall that $M_n(\phi) := M_{n,X}(\phi) + M_{n,Y}(\phi)$ for $\phi \geq 1$, where

$$M_{n,X}(\phi) := n^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\max_{1\leq j\leq p}|X_{ij}|^3 1\left\{\max_{1\leq j\leq p}|X_{ij}| > \sqrt{n}/(4\phi\log p)\right\}\right]. \tag{73}$$

Obviously ρ_n is a specified version of ϱ_n and our lemma derives a tight bound of ρ_n.

Theorem 5 (Key Lemma)**.** *Suppose that there exists some constant $b > 0$ such that $n^{-1}\sum_{i=1}^{n}\mathbb{E}[X_{ij}^2] \geq b$ for all $j = 1,\ldots,p$. Then ϱ_n satisfies the following inequality for all $\phi \geq 1$:*

$$\varrho_n \lesssim \frac{\phi^2\log^2 p}{n^{1/2}}\left\{\phi L_n\varrho_n + L_n\log^{1/2}p + \phi M_n(\phi)\right\} + \frac{\log^{1/2}p}{\phi}$$

up to a constant K that depends only on b.

The proof is based on the Slepian-Stein method proposed in [6], which improves upon Lindeberg's method. This improved method is commonly applied to high-dimensional Gaussian approximation problems. However, the original idea, swapping one variable for another in high-dimensional settings does not lead to a precise bound, because intuitively this may spend a lot of steps. Therefore, we introduce parameters to interpolate between the variables and analyze their behavior as the parameters vary between 0 and 1. The Taylor expansion is a natural way to bound the difference, in which we only keep one term, leveraging the independence property and considering the local behavior of the random variables. We believe that this idea is related to the classical Lindeberg method, as the leave-one-out method is originally derived from Lindeberg's work.

Let $W_1,\ldots,W_n$ be a copy of $Y_1,\ldots,Y_n$. We may assume that $X_1,\ldots,X_n$, $Y_1,\ldots,Y_n$, and $W_1,\ldots,W_n$ are independent. Consider $S_n^W := n^{-1/2}\sum_{i=1}^{n}W_i$. Then $\mathbb{P}(S_n^Y \leq y) = \mathbb{P}(S_n^W \leq y)$, so that

$$\varrho_n = \sup_{y\in\mathbb{R}^p, v\in[0,1]} |\mathbb{P}\left(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y \leq y\right) - \mathbb{P}(S_n^W \leq y)|. \tag{74}$$

For any $y \in \mathbb{R}^p$ and $v \in [0,1]$, we choose $\beta := \phi\log p$, and denote

$$\varrho_n = \sup_{y\in\mathbb{R}^p, v\in[0,1]} |\mathbb{P}\left(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y \leq y\right) - \mathbb{P}(S_n^W \leq y)|. \tag{75}$$

The function $F_\beta(w)$ satisfies the property below, which means that F_β is a good approximation of maximum:

$$0 \leq F_\beta(w) - \max_{1\leq j\leq p}(w_j - y_j) \leq \beta^{-1}\log p = \phi^{-1}, \text{ for all } w \in \mathbb{R}^p. \tag{76}$$

Actually, the construction of F_β comes from statistical physics where it comes from the free energy of a spin-glass system.

Next, we select a smooth approximation function $g_0 : \mathbb{R} \to [0,1]$ and define it such that $g_0(t) = 1$ for $t \leq 0$ and $g_0(t) = 0$ for $t \geq 1$. We further define $g(t)$ as $g(t) := g_0(\phi t)$, and

$$m(w) := g(F_\beta(w)),\ w \in \mathbb{R}^p. \tag{77}$$

We notice that m is exactly a version of the smooth approximation of the indicator of maximum. And the l_1 norm of thrice derivative of m can be bounded well:

$$\textstyle\sum_{j,k,l=1}^p |m_{jkl}(w)| \lesssim (\phi^3 + \phi\beta + \phi\beta^2) \lesssim \phi\beta^2, \tag{78}$$

$$|m_{jkl}(w)| \lesssim |m_{jkl}(w+\tilde{w})| \lesssim |m_{jkl}(w)|, \tag{79}$$

where the inequality (79) holds for all $w, \tilde{w} \in \mathbb{R}^p$ with $\max_{1\leq j\leq p}|\tilde{w}_j|\beta \leq 1$. Define the functions

$$h(w,t) := 1\left\{-\phi^{-1} - t/\beta < \max_{1\leq j\leq p}(w_j - y_j) \leq \phi^{-1} + t/\beta\right\},\ w \in \mathbb{R}^p, t > 0, \tag{80}$$

$$\omega(t) := \frac{1}{\sqrt{t} \wedge \sqrt{1-t}},\ t \in (0,1).$$

The proof consists of two steps. In the first step, we show that

$$|\mathbb{E}[\mathcal{I}_n]| \lesssim \frac{\phi^2 \log^2 p}{n^{1/2}}\left(\phi L_n \varrho_n + L_n \log^{1/2} p + \phi M_n(\phi)\right) \tag{81}$$

where $\mathcal{I}_n := m(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y) - m(S_n^W)$. In the second step, we use anti-concentration inequality together with the above bound to complete the proof.

Step 1. Consider the Slepian's interpolant $Z(t) := \sum_{i=1}^n Z_i(t),\ t \in [0,1]$, where

$$Z_i(t) := \frac{1}{\sqrt{n}}\left\{\sqrt{t}(\sqrt{v}X_i + \sqrt{1-v}Y_i) + \sqrt{1-t}W_i\right\}. \tag{82}$$

Obviously,

$$\mathcal{I}_n = m(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y) - m(S_n^W) = \int_0^1 \frac{dm(Z(t))}{dt}dt. \tag{83}$$

We denote $Z^{(i)}(t)$ as the remaining term for $Z(t)$, i.e. $Z^{(i)}(t) := Z(t) - Z_i(t)$. Finally, define

$$Z_i(t) := \frac{1}{\sqrt{n}}\left\{\sqrt{t}(\sqrt{v}X_i + \sqrt{1-v}Y_i) + \sqrt{1-t}W_i\right\}. \tag{84}$$

as a type of its derivative.

Now, by chain rule and Taylor Expansion, we have

$$\mathbb{E}[\mathcal{I}_n] = \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathbb{E}[m_j(Z)\dot{Z}_{ij}]dt = \frac{1}{2}(I + II + III), \tag{85}$$

where

$$I := \sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathbb{E}[m_j(Z^{(i)})\dot{Z}_{ij}]dt, \tag{86}$$

$$II := \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathbb{E}[m_{jk}(Z^{(i)})\dot{Z}_{ij}Z_{ik}]dt, \tag{87}$$

$$III := \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 (1-\tau)\mathbb{E}[m_{jkl}(Z^{(i)}+\tau Z_i)\dot{Z}_{ij}Z_{ik}Z_{il}]d\tau dt. \tag{88}$$

Due to the independence of $Z^{(i)}$ from the centered $\dot{Z}_{ij}$ variables, we have $I = 0$. Similarly, using the independence and the assumption of the same second moment, we find that $II = 0$. Thus, it is enough to bound III. To achieve this, we employ a truncation technique, which allows us to complete the proof. Denote

$$\chi_i := 1\left\{\max_{1\le j\le p}|X_{ij}|\vee|Y_{ij}|\vee|W_{ij}| \le \sqrt{n}/(4\beta)\right\},\ i = 1,\dots,n \tag{89}$$

and $III = III_1 + III_2$, where

$$III_1 := \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 (1-\tau)\mathbb{E}[\chi_i m_{jkl}(Z^{(i)}+\tau Z_i)\dot{Z}_{ij}Z_{ik}Z_{il}]d\tau dt, \tag{90}$$

$$III_2 := \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 (1-\tau)\mathbb{E}[(1-\chi_i) m_{jkl}(Z^{(i)}+\tau Z_i)\dot{Z}_{ij}Z_{ik}Z_{il}]d\tau dt. \tag{91}$$

For III_2, we have some insights that it should be bound by some tail expectations, then

$$|III_2| \lesssim (M_{n,X}(\phi) + M_{n,Y}(\phi))\phi\beta^2/n^{1/2} = M_n(\phi)\phi\beta^2/n^{1/2}. \tag{92}$$

To bound III_1, we use another truncation to control the term of τZ_i in m_{jkl}, which is rather tricky and we suggest the reader see [7] as a reference; we split the integral again and conclude that

$$|III_1| \lesssim \frac{\phi\beta^2 L_n}{n^{1/2}}(\varrho_n + \phi^{-1}\log^{1/2}p) \lesssim \frac{\phi^2\log^2 p}{n^{1/2}}(\phi L_n\varrho_n + L_n\log^{1/2}p), \tag{93}$$

where $\beta = \phi\log p$.

Step 2. This is a much shorter and easier step. For $V_n := \sqrt{v}S_n^X + \sqrt{1-v}S_n^Y$,

$$\begin{aligned}\mathbb{P}(V_n \le y - \phi^{-1}) &\le \mathbb{P}(F_\beta(V_n) \le 0) \le \mathbb{E}[m(V_n)] \\ &\le \mathbb{P}(F_\beta(S_n^W) \le \phi^{-1}) + (\mathbb{E}[m(V_n)] - \mathbb{E}[m(S_n^W)]) \\ &\le \mathbb{P}(S_n^W \le y + \phi^{-1}) + |\mathbb{E}[\mathcal{I}_n]| \\ &\le \mathbb{P}(S_n^W \le y - \phi^{-1}) + C\phi^{-1}\log^{1/2} p + |\mathbb{E}[\mathcal{I}_n]|,\end{aligned}$$

the last inequality is also because of the anti-concentration result. It is similar to the other direction. After collecting all these inequalities, we derive this result. □

Now we can derive some application results by using this lemma. By specifying a particular property of X_{ij}, such as being uniformly bounded in a sub-Gaussian sense, we can obtain more specific convergence rates. Meanwhile, this result can be extended from the class of hyperrectangles to other classes. We consider classes of simple convex sets and derive bounds that are similar under certain conditions. Although it is not a real challenge to extend the results to simple convex sets, the size of this class in high-dimensional spaces is significant. By considering sparsely convex sets as well, we can obtain similar bounds. These classes can be beneficial in statistics, as sparse models and techniques have played a crucial role in recent years. Also, due to the convenience of calculation, this result can be similarly extended to Bootstrap cases, which needs more techniques but leads to some reliability of data-dependent methods.

6.2 Randomized version of Lindeberg Replacement

Chernozhukov, etc. did not stop there but continued to make further contributions to strengthen the rate of this estimation. In this section, we will discuss this iterative randomized Lindeberg replacement method developed in [8], which to some extent, improves the existing estimation rate and also extends to more classes of bootstrap methods.

We first introduce what is the meaning of iterative randomized Lindeberg method. Define

$$\varrho_\epsilon = \sup_{y\in\mathbb{R}^p} \left|\mathbb{P}\left(S_{n,\epsilon}^V \le y\right) - \mathbb{P}\left(S_n^Z \le y\right)\right|, \tag{94}$$

where

$$S_{n,\epsilon}^V = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\epsilon_i V_i + (1-\epsilon_i)Z_i) \quad \text{and} \quad S_n^Z = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i.$$

To analyze the problem, we employ a random process in $0, 1^n$, which is composed of $\epsilon^0, \dots, \epsilon^D \in 0, 1^n$. These vectors are independent of Z and V and we seek to establish recursive bounds for ρ_{ϵ^d} for $d = 0, \dots, D$. In order to construct such a sequence of random vectors, we follow these steps:

- We first choose $D = [4\log n] + 1$, which determines the number of steps to use. We initialize $\epsilon^0 = (1, \dots, 1)$.
- We generate $U_1, \dots, U_D$ as independent and identically distributed uniform random variables on the interval $[0, 1]$ and independent of Z and V's.

- Given ϵ^{d-1} and $U_1, \dots, U_D$, we set $\epsilon_i^d = 0$ if $\epsilon_i^{d-1} = 0$, and generate $\epsilon_i^d i \in I_{d-1}$ using i.i.d. Bernoulli(U_d) random variables where $I_{d-1} = \{i = 1, \dots, n : \epsilon_i^{d-1} = 1\}$. This process is regarded as one step of replacement.

Obviously, ϵ^d has these properties:

(i) for every $i = 1, \dots, n$, $\epsilon_i^d = 0$ if $\epsilon_i^{d-1} = 0$;

(ii) for $I_{d-1} = \{i = 1, \dots, n \colon \epsilon_i^{d-1} = 1\}$, $\{\epsilon_i^d\}_{i \in I_{d-1}}$ are exchangeable conditional on ϵ^{d-1} and satisfy

$$\mathbb{P}\left(\sum_{i \in I_{d-1}} \epsilon_i^d = s \mid \epsilon^{d-1}\right) = \frac{1}{|I_{d-1}| + 1}, \quad \text{for all } s = 0, \dots, |I_{d-1}| \tag{95}$$

which is a classical result from some Bayesian viewpoints.

The recursive inequality is established by linking S_{n,ϵ^d}^V with S_n^Z using the randomized Lindeberg method, first introduced by [9]. In contrast to our prior work in [7], where we employed the Slepian-Stein method to connect S_{n,ϵ^0}^V with G, the randomized Lindeberg method enables matching of moments up to the third order for both S_{n,ϵ^d}^V and S_n^Z. This results in an improved power of $\log(pn)$, leading to a more substantial increase in the sample size n.

We always assume V and Z to be independent centered r.v.s in $\mathbb{R}^p$. Also, some technical conditions are stated. For example, for every component, the mean of the fourth moments of V and Z should be uniformly bounded. Also, they should be uniformly sub-Gaussian as well as Z have some properties like anti-concentration results.

Theorem 6. *Suppose that some mild conditions are satisfied. Then for every $d < D$ and $\phi > 0$ s.t.*

$$C_p B_n \phi \log^2(pn) \leq \sqrt{n}, \tag{96}$$

restricted on $\mathcal{A}_d$,

$$\begin{aligned} \varrho_{\epsilon^d} \lesssim{}& \frac{\sqrt{\log p}}{\phi} + \delta + \frac{B_n^2 \phi^4 \log^5(pn)}{n^2} + \left(\mathbb{E}[\varrho_{\epsilon^{d+1}} \mid \epsilon^d] + \frac{\sqrt{\log p}}{\phi} + \delta\right) \\ &\times \left(\frac{\mathcal{B}_{n,1,d} \phi^2 \log p}{\sqrt{n}} + \frac{\mathcal{B}_{n,2,d} \phi^3 \log^2 p}{n} + \frac{B_n^2 \phi^4 \log^3(pn)}{n}\right), \end{aligned}$$

Where

$$\begin{aligned} \mathcal{A}_d = {}& \left\{\max_{1 \leq j,k \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i^d (\mathcal{E}_{i,jk}^V - \mathcal{E}_{i,jk}^Z) \right| \leq \mathcal{B}_{n,1,d}\right\} \\ &\bigcap \left\{\max_{1 \leq j,k,l \leq p} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i^d (\mathcal{E}_{i,jkl}^V - \mathcal{E}_{i,jkl}^Z) \right| \leq \mathcal{B}_{n,2,d}\right\}. \end{aligned}$$

The intuition behind the iterative randomized Lindeberg method can be explained as follows. When approximating $\mathbb{E}[g(X_1 + \cdots + X_n)]$ using $\mathbb{E}[g(Y_1 + \cdots + Y_n)]$, the traditional Lindeberg method constructs an interpolation path where X_i's are replaced with Y_i's one-by-one in a given order. The method then utilizes Taylor's expansion to demonstrate the small change in expectation due to each replacement. On the other hand, the randomized Lindeberg method, introduced in [9], replaces X_i's with Y_i's in a randomly selected order, leading to significant benefits in the final bound. To enhance this approach, we examine the coefficients of Taylor's expansion and apply the randomized Lindeberg method again for further approximations. This iterative method introduces new coefficients, and the process is repeated until the approximation error reaches an acceptable level. Our research demonstrates that the iterative randomized Lindeberg method significantly improves the final bound. We provide a comparison of this method with the randomized Lindeberg method presented in [9] and the Slepian-Stein method used in [6, 7] in advance of Lemma 6.

And after this, we have our main result which shows some improvement of the convergence rate:

Theorem 7 (Distributional Approximation via Iterative Randomized Lindeberg Method, [8])**.** *Under mild assumptions, if*

$$\max_{1\le j,k\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbb{E}[V_{ij}V_{ik}]-\mathbb{E}[Z_{ij}Z_{ik}])\right|\le C_m B_n\sqrt{\log(pn)} \tag{97}$$

and

$$\max_{1\le j,k,l\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbb{E}[V_{ij}V_{ik}V_{il}]-\mathbb{E}[Z_{ij}Z_{ik}Z_{il}])\right|\le C_m B_n^2\sqrt{\log^3(pn)} \tag{98}$$

for some constant C_m. Then

$$\sup_{y\in\mathbb{R}^p}\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_i\le y\right)-\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\le y\right)\right|\le C\left(\left(\frac{B_n^2\log^5(pn)}{n}\right)^{1/4}+\delta\right),$$

where C is a constant depending only on our assumptions and C_m.

The proof of this theorem is rather tricky. We use some kind of induction and the reader can see [8] for more detailed proof.

6.3 Optimal Rate for High-Dimensional Settings

We have reached some near $n^{-1/4}$ rates. For some possible developments, we know that generally, the $n^{-1/2}$ rate in the classical Berry-Esseen theorem is optimal. And Lopes in [17] modified the result to get this optimal rate in Gaussian approximations.

The utilization of smoothing techniques plays a pivotal role in the proofs. These techniques involve the use of a smooth function $\psi:\mathbb{R}^p\to\mathbb{R}$ that is dependent on a set $A\subset\mathbb{R}^p$, such that $\mathbb{E}[\psi(B_n)]\approx\mathbb{P}(B_n\in A)$. While these techniques are crucial, they come with the disadvantage of introducing an additional smoothing error $|\mathbb{P}(B_n\in A)-\mathbb{E}[\psi(B_n)]|$, which needs to be balanced with errors arising from other approximations. Often, this balancing process becomes a bottleneck for the overall rate of Gaussian approximation, and finding appropriate trade-offs can be challenging.

To tackle this challenge, we utilize a smoothing function introduced in the Lindeberg interpolation framework, which avoids introducing any additional smoothing error. This approach assumes that $X_1, \ldots, X_n$ are non-Gaussian, while $Y_1, \ldots, Y_n$ are Gaussian. It takes into account the probability $\mathbb{P}(\sum_{i=1}^{k} X_i + \sum_{j=k+1}^{n} Y_j \in A)$, which can be expressed as $\mathbb{E}[\tilde{\psi}(\sum_{i=1}^{k} X_i)]$, where $\tilde{\psi}$ is a specific smooth random function defined with respect to $Y_{k+1}, \ldots, Y_n$. This smoothing technique shares some similarities with the method proposed by [3], where the cumulative distribution function of the normal distribution was employed to create an exact and unbiased approximation of the difference. The resulting estimation requires that each component of X is uniformly sub-Exponential or sub-Gaussian. For brevity, we will focus on the case where they are uniformly sub-Exponential.

We introduce this main theorem:

Theorem 8 (Optimal Gaussian approximation, Lopes [17]). *There is an absolute constant $C > 0$, s.t. the following holds for all n, p: Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be centered i.i.d. r.v.s, and $\nu_q = \max_{1 \le j \le p} \|X_{1j}/\sqrt{\mathrm{var}(X_{1j})}\|_{\psi_1}$ is finite. Let ρ be the smallest eigenvalue of the correlation matrix of X_1, and $\rho > 0$. $Y \in \mathbb{R}^p$ be the Gaussian version with $\mathbb{E}[YY^\top] = \mathbb{E}[X_1 X_1^\top]$. Then,*

$$\sup_{A \in \mathscr{R}} \left| \mathbb{P}\big(\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^{n} X_i \in A\big) - \mathbb{P}(Y \in A) \right| \;\le\; \frac{C \nu_q^4 \log^6(pn) \log(n)}{\rho^{3/2}\, n^{1/2}}, \tag{99}$$

Sketch of Proof. Denote $S_{k:k'}(X) = n^{-1/2}(X_k + \cdots + X_{k'})$ and similar for Y, where they are i.i.d. copies of X, and

$$\mathsf{D}_k = \sup_{A \in \mathscr{R}} \Big| \mathbb{P}(S_{1:k}(X) \in A) - \mathbb{P}(S_{1:k}(Y) \in A) \Big| \tag{100}$$

as well as

$$\delta_k^X(A) = \mathbb{P}\Big(S_{1:k-1}(X) + \tfrac{1}{\sqrt{n}} X_k + S_{k+1:n}(Y) \in A\Big) - \mathbb{P}\Big(S_{1:k-1}(X) + S_{k+1:n}(Y) \in A\Big) \tag{101}$$

as well as for Y, So

$$\mathbb{P}(S_{1:n}(X) \in A) - \mathbb{P}(S_{1:n}(Y) \in A) \;=\; \sum_{k=1}^{n} \{\delta_k^X(A) - \delta_k^Y(A)\}. \tag{102}$$

Then define

$$\delta_k \;=\; \sup_{A \in \mathscr{R}} |\delta_k^X(A) - \delta_k^Y(A)|, \tag{103}$$

The proof of the major theorem is quite complicated, and we suggest the reader to see [17] for further reading.

The key is to bound δ_k, and thus, to bound their sum; in fact, the major lemma says for every $n \geq 3$, $p \geq 1$, and $k \in \{2, \dots, n-1\}$,

$$\delta_k \ \leq \ \frac{c\nu_q^3 (\log(p))^{\frac{9}{2}}}{\epsilon_k^3 \, n^{3/2}} \Big(\epsilon_k \log(pn) \sqrt{\tfrac{n}{k-1}} \ + \ \mathsf{D}_{k-1} \ + \ \tfrac{1}{pn} \Big). \tag{104}$$

Here we estimate δ_k by implicit smoothing. To know what exactly this term means, denote $\zeta \sim N(0, I_p)$ and for any fixed $s \in \mathbb{R}^p$, $A \in \mathscr{R}$, and $\epsilon > 0$, define

$$\varphi_\epsilon(s, A) = \mathbb{P}(s + \epsilon\zeta \in A). \tag{105}$$

Actually, $\varphi_\epsilon(\cdot, A)$ is a smoothed version of the indicator $s \mapsto 1\{s \in A\}$, when ϵ tends to 0. Next, for each $k = 1, \dots, n-1$, define

$$\epsilon_k = \sqrt{\tfrac{n-k}{n}} \sqrt{\rho}. \tag{106}$$

We use ϵ_k to decompose the Gaussian r.v. $S_{k+1:n}(Y)$ that

$$S_{k+1:n}(Y) \ \overset{\mathcal{L}}{=} \ \epsilon_k V_{k+1} \ + \ \sqrt{\tfrac{n-k}{n}} W_{k+1}, \tag{107}$$

where $V_{k+1} \sim N(0, I_p)$ and $W_{k+1} \sim N(0, R - \rho I_p)$ are independent with X and Gaussian r.v.s, and we know all eigenvalues of R is larger than ρ. Now defined

$$\hat{A}_{k+1} = \Big\{ x - \sqrt{\tfrac{n-k}{n}} W_{k+1} \Big| x \in A \Big\}, \tag{108}$$

then it's easy to see that

$$\mathbb{P}\Big(S_{1:k}(X) + S_{k+1:n}(Y) \in A\Big) \ = \ \mathbb{E}\Big[\varphi_{\epsilon_k}(S_{1:k}(X) \, , \, \hat{A}_{k+1})\Big]. \tag{109}$$

which means φ_{ϵ_k} is an unbiased version of smooth approximation on the randomly shifted set. As a result,

$$\delta_k^X(A) = \mathbb{E}\Big[\varphi_{\epsilon_k}\Big(S_{1:k-1}(X) + \tfrac{1}{\sqrt{n}} X_k, \hat{A}_{k+1}\Big) - \varphi_{\epsilon_k}\Big(S_{1:k-1}(X), \hat{A}_{k+1}\Big)\Big]. \tag{110}$$

By Taylor's Theorem, we can split the above into three parts, and the first and second moments agreed, by independence, we can subtract Y's version from it and get $\delta_k^X(A) - \delta_k^Y(A) \ = \ \mathbb{E}[R_k^X(A)] - \mathbb{E}[R_k^Y(A)]$. where

$$R_k^X(A) = \tfrac{(1-\tau)^2}{2} \Big\langle \nabla^3 \varphi_{\epsilon_k} \Big(S_{1:k-1}(X) + \tfrac{\tau}{\sqrt{n}} X_k, \hat{A}_{k+1}\Big), \ n^{-3/2} X_k^{\otimes 3} \Big\rangle \tag{111}$$

with τ Uniform on [0,1] and independent of all others. Therefore, $\delta_k \leq \sup_{A \in \mathscr{R}} \mathbb{E}[|R_k^X(A)|] + \sup_{A \in \mathscr{R}} \mathbb{E}[|R_k^Y(A)|]$, and we use Holder's inequality to bound $\sup_{A \in \mathscr{R}} \mathbb{E}[|R_k^X(A)|]$. Actually, this turns to consider the event $E_k(\epsilon_k) \ = \ \Big\{ S_{1:k-1}(X) \in$

$\partial\tilde{A}_{k+1}(\epsilon_k)\Big\}$. We know $\sup_{(s,A)\in\mathbb{R}^p\times\mathscr{R}} \|\nabla^3\varphi_{\epsilon_k}(s,A)\|_1 \le \frac{c\log^{3/2}(p)}{\epsilon_k^3}$ is a global bound, so on the event $E_k(\epsilon_k)$, it turns to bound some tail expectation of X. This can turn to anti-concentration and sub-exponential results. Besides, outside the event $E_k(\epsilon_k)$, the gradient can be relatively small:

$$\sup\left\{\|\nabla^3\varphi_{\epsilon_k}(s,A)\|_1 \,\middle|\, A\in\mathscr{R} \text{ and } s\in(\mathbb{R}^p\setminus\partial A(\epsilon_k))\right\} \le \tfrac{c}{\epsilon_k^3 pn}.$$

Combining the two cases, we get the bound for δ_k, and we can see that it can be dominated by some term related to D_{k-1}, so the induction hypothesis can be used; it turns to control some sum, and we can choose a proper m to get an optimal bound for D_n. Detailed proof can be found in [17].

□

Our exemplary bound provides significant implications, such as utilizing it to perform inference on mean vectors or refining the precision bounds of numerous bootstrap methods. This is of paramount importance, as the accuracy of such procedures is crucial in ensuring the validity and reliability of statistical analyses. By leveraging our efficient bound, researchers can have greater confidence in their results and draw more accurate conclusions from their data.

7 Other Interesting Results: CLT of Random Sum

In this chapter, we describe further results derived from the Lindeberg Replacement Method. We look at the situation from Chapter 1 again and look at partial sums of independent and not identically distributed random variables, where now the number of summation terms is random. The fact that the limit distribution is not restricted solely to the Gaussian distribution piques our interest. These findings can potentially be applied to real-life problems, such as finance, economics, and risk management, where the underlying assumptions are not strictly Gaussian and require a more comprehensive understanding of the limit distribution.

Assume there are independent centered random variables X_i with $\sigma_i^2 = \mathbb{E}(X_i^2) < \infty$. Furthermore, N is a random variable with values in $\mathbb{N}$ with $\mathrm{Var}(N) < \infty$, and it is chosen independently of the sigma field $\sigma(X_i : i\in\mathbb{N})$. We are interested in the randomized partial sum

$$W_N := \frac{1}{\sqrt{\mathbb{E}B_n^2}}\sum_{i=1}^{N} X_i \tag{112}$$

where $B_n^2 := \sum_{i=1}^N \sigma_i^2$.

Random partial sum is a much-studied object in probability theory. They occur in the theory of branching processes, models of mathematical biology, as well as in economics and risk theory.

For example, If N_n is about the number of male offspring in the n-th generation, and if the j-th of these offspring has $X_j^{(n+1)}$ sons, then $N_{n+1} = \sum_{i=1}^{N_n} X_i^{(n+1)}$. Then it is called a Galton-Watson trial. Under what conditions at the moments of X_i and N can convergence in distribution be demonstrated and will a central limit theorem apply? In order to somewhat simplify the arguments, assume $\sigma_i^2 = 1$ for all i, then $\mathbb{E}B_n^2 = \mathbb{E}N$.

We then examine the limit distribution of $W_N = \frac{1}{\sqrt{\mathbb{E}N}} \sum_{i=1}^N X_i$. If the Z_i are again independent $\mathscr{N}(0,1)$ distributed random variables, independent of N, we investigate $\mathbb{E}(f(W_N) - f(Z^{(N)}))$ using Lindeberg's sum decomposition, where $Z^{(N)} := \frac{1}{\sqrt{\mathbb{E}N}} \sum_{i=1}^N Z_i)$. But now we do not know the distribution of $Z^{(N)}$. For what N, the limit of $Z^{(N)}$ is normally distributed? For each test function f chosen as in Chapter 2, below is followed by conditional expectations

$$\mathbb{E}(f(W_N) - f(Z^{(N)})) = \sum_{n \in \mathbb{N}} \mathbb{P}(N = n)\mathbb{E}(f(W_n) - f(Z^{(n)})) \tag{113}$$

Therefore, $|\mathbb{E}(f(W_N) - f(Z^{(N)}))| \leq T_1 + T_2$, where

$$T_1 \leq 2K\epsilon \sum_{n \geq 1} \mathbb{P}(N = n)\frac{n}{\mu} = 2K\epsilon \tag{114}$$

and

$$T_2 \leq \frac{K}{\mu} \sum_{n \geq 1} \mathbb{P}(N = n) \sum_{i=1}^{n} (\int_{|X_i| > \epsilon\sqrt{\mu}} X_i^2 \mathrm{dP} + \int_{|Z_i| > \epsilon\sqrt{\mu}} Z_i^2 \mathrm{dP}) \tag{115}$$

As for the i.i.d. case, we can especially derive

$$T_2 \leq K(\int_{|X_1| > \epsilon\sqrt{\mu}} X_1^2 \mathrm{dP} + \int_{|Z_1| > \epsilon\sqrt{\mu}} Z_1^2 \mathrm{dP}) \tag{116}$$

To control T_2 on an arbitrarily small scale, we need to let the event $\{|X_1| > \epsilon\sqrt{\mu}\}$ tend to a null event monotonically. In other words, when $\mu = \mathbb{E}N \to \infty$, we can have a similar limit result for random partial sums. It is not surprising that we examine a random sum of N summands since the expected number of summands would increase as in our classical limit analysis. In more general cases, by Fubini's Theorem, we have

$$T_2 \leq \frac{K}{\mu} \sum_{i \geq 1} \mathbb{P}(N \geq i)(\int_{|X_i| > \epsilon\sqrt{\mu}} X_i^2 \mathrm{dP} + \int_{|Z_i| > \epsilon\sqrt{\mu}} Z_i^2 \mathrm{dP}) \tag{117}$$

So it needs to become small while $\mathbb{E}N \to \infty$. However, for non-identically distributed cases, we note the scalar term of W_N and $Z^{(N)}$ should be changed into $\mathbb{E}(B_n^2)$. In order to find some consistency with the i.i.d. case, we assume $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \sigma_i^2 = \sigma^2$ with some $\sigma > 0$. In this case, the increasing rate of $\mathbb{E}(B_n^2)$ to $\mathbb{E}N$ is asymptotically fixed. As well as for T_1, we need to mention that

$$T_1 \leq 2K\epsilon \sum_{n \geq 1} \mathbb{P}(N = n)\frac{\sum_{i=1}^n \sigma_i^2}{\mu} \tag{118}$$

So there must exist some $M > 0$ s.t. $T_1 \leq 2K\epsilon M$. To sum up, we get the following theorem, which should be thought as a simple conditional version of the primitive Lindeberg's result:

Theorem 9. *Suppose there are independent centered random variables $X_1, \cdots$ with $\sigma_i^2 := \mathbb{E}X_i^2 > 0$for each i, and $Z_1, \cdots$ be some independently Gaussian copies with $Z_i \sim \mathscr{N}(0, \sigma_i^2)$. Furthermore, N is a random variable with values in $\mathbb{N}$, independent of all other elements with $\mathbb{E}N \to \infty$. Suppose $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \sigma_i^2 = \sigma^2$ and Lindeberg condition holds for every $\epsilon > 0$:*

$$\frac{1}{\mathbb{E}N}\sum_{i\geq 1}\mathbb{P}(N \geq i)(\int_{|X_i|>\epsilon\sqrt{\mathbb{E}N}} X_i^2 \mathrm{dP} + \int_{|Z_i|>\epsilon\sqrt{\mathbb{E}N}} Z_i^2 \mathrm{dP}) \longrightarrow 0 \tag{119}$$

Then with our above definition of W_N and $Z^{(N)}$, we have

$$|\mathbb{E}(f(W_N) - f(Z^{(N)}))| \to 0 \tag{120}$$

Due to the ubiquitous of its application, the case of a geometrically distributed random variable N was always considered. Geometric distribution means if $P(N = n) = (1-p)^{n-1}p$ for some $0 < p < 1$ and $n \in \mathbb{N}$. In the model of the independent coin toss, this is the probability that success occurred for the first time at time n. It applies $\mathbb{E}(N) = 1/p$ and $\mathrm{Var}(N) = (1-p)/p^2$. So we look at the case $p \to 0$ to find a limit for the distribution of W_N. Since $P(N \geq i) = (1-p)^{i-1}$, the Lindeberg condition 119 here is:

$$\lim_{p\to 0}\sum_{i\geq 1}(1-p)^{i-1}p\int_{|X_i|>\epsilon p^{-1/2}} X_i^2 \mathrm{dP} = 0 \tag{121}$$

In [18] it was further shown that 121 applies to the normally distributed Z_i if one additionally demands that $\lim_{n\to\infty} n^{-\gamma}\sigma_n^2 = 0$ applies to a $0 < \gamma < 1$. Especially, if we consider the case of identically distributed random quantities X_i with $\sigma_i^2 = 1$, then it follows immediately that $|\mathbb{E}(f(W_N) - f(Z^{(N)}))| \to 0$.

In [18], it was shown that the limit of the characteristic function $\mathbb{E}(\exp(itZ(N)))$ when $p \to 0$ is $\frac{1}{1+t^2/2}$. The characteristic function of a distribution clearly determines itself, and we can see the limit distribution with this specified characteristic function has the density

$$f(x) = \frac{1}{\sqrt{2}}\exp(-\sqrt{2}|x|) \tag{122}$$

We call this a Laplace distribution.

The possible class of limit distributions is described by means of the characteristic function of $Z^{(N)}$ as follows. If $Z_i \sim \mathscr{N}(0, \sigma_i^2)$ are independent, $\mathbb{E}(\exp(itZ(N))) = \sum_{i\geq 1}\mathbb{P}(N = n)\exp(-\frac{t^2}{2\mathbb{E}N}B_n^2)$, and $B_n^2/n \to \sigma^2$, it's easy to see the characteristic function converges to $\frac{1}{1+\sigma^2t^2/2}$, which represents a scaled Laplace distribution.

It's quite an amazing result, for if N is deterministic, i.e. there exists some n_0 s.t. $P(N = n_0) = 1$, then $Z^{(N)}$ is normally distributed. For any random variable N with values in $\mathbb{N}$, we do not have this information, and other limit distributions may occur, as we have shown above. But we can still take conditional expectation on N and use some similar approach like Lindeberg's. That's the charm of this method!

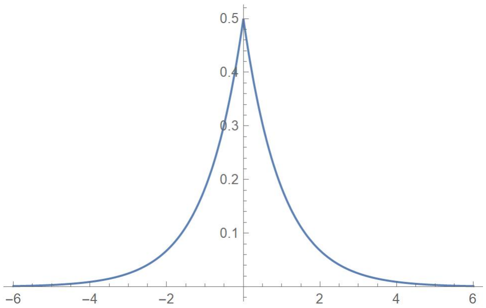

Fig. 2 PDF of a Laplace Distribution

The Lindeberg method was thus successfully presented for a non-central limit theorem. This remains an indication that why a geometrically distributed number of summands does not have a central limit set. The variance of the number of summands is $(1-p)/p^2$ and thus grows for $p \to 0$ faster than the expected value. The variation of the number of summands is therefore too large to allow a central behavior in the sense of a central limit theorem.

Finally, in the case of random partial sums W_N, we can also derive convergence rates from the Lindeberg proof. For the situation of independent and identically distributed X_i, we denote $\alpha = \mathbb{E}|X_1|^3 \leq \infty$ and again consider only the case $\sigma_i^2 = 1$. Using this to estimate g and similarly, we have

$$|\mathbb{E}(f(W_N) - f(Z^{(N)}))| = K\mathcal{O}(\frac{\alpha}{\sqrt{\mathbb{E}N}}) \tag{123}$$

In the case of geometric sums, the convergence rate is $p^{1/2}$ therefore.

8 Conclusion

This paper provides a comprehensive overview of the Lindeberg Method, a technique used to prove the central limit theorem. The method was developed by H. C. Lindeberg in a series of papers in the 1920s, and has been widely used due to its simplicity and versatility. The paper first provides an introduction to the central limit theorem, which describes the behavior of collections of independent and identically distributed random variables. It then goes on to describe the history and development of the Lindeberg Method, which has been applied in various fields including random matrices and high-dimensional Gaussian approximations.

In addition to introducing key concepts and applications of the Lindeberg Method, the paper also explores recent advancements in the field. These developments have broadened the scope of the method and opened up new research directions. We especially consider Chernozhokov's breakthrough of Gaussian Approximation in High-Dimensional settings. Overall, the paper is a valuable resource for researchers and students in probability theory, statistics, and related fields who seek to develop a deeper understanding of the central limit theorem and the Lindeberg Method.

Acknowledgments. The author acknowledges Professor Dang-Zheng Liu, USTC for selfless guidance and dedication.

References

[1] L. Arnold. On the asymptotic distribution of the eigenvalues of random matrices. Technical report, WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CENTER, 1967.

[2] V. Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

[3] E. Bolthausen. Exact convergence rates in some martingale central limit theorems. *The Annals of Probability*, pages 672–688, 1982.

[4] S. Chatterjee. A simple invariance theorem. *arXiv preprint math/0508213*, 2005.

[5] S. Chatterjee. A generalization of the lindeberg principle. 2006.

[6] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. 2013.

[7] V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. 2017.

[8] V. Chernozhuokov, D. Chetverikov, K. Kato, and Y. Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586, 2022.

[9] H. Deng and C.-H. Zhang. Beyond gaussian approximation: Bootstrap for maxima of sums of independent random vectors. 2020.

[10] R. Durrett. Probability: theory and examples. 49, 2019.

[11] W. Feller. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung. pages 167–205, 2015.

[12] P. Hall and C. C. Heyde. Martingale limit theory and its application. 2014.

[13] L. Le Cam. The central limit theorem around 1935. *Statistical science*, pages 78–91, 1986.

[14] J. W. Lindeberg. Über das exponentialgesetz in der wahrscheinlichkeitsrechnung. 1920.

[15] J. W. Lindeberg. äber das gauss' sche fehlergesetz. *Scandinavian Actuarial Journal*, 1922(1):217–234, 1922.

[16] J. W. Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922.

[17] M. E. Lopes. Central limit theorem and bootstrap approximation in high dimensions: Near 1/n rates via implicit smoothing. *The Annals of Statistics*, 50(5):2492–2513, 2022.

[18] J. Pike and H. Ren. Stein's method and the laplace distribution. *arXiv preprint arXiv:1210.5775*, 2012.

[19] G. Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8(3-4):171–181, 1920.

[20] T. Tao and V. Vu. Random matrices: the universality phenomenon for wigner ensembles. *Modern aspects of random matrix theory*, 72:121–172, 2014.

[21] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, pages 548–564, 1955.

[22] E. P. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, pages 325–327, 1958.