Lecture notes on Probability Limit Theory

Yukun He*

June 11, 2026

This is the lecture notes I use for the course "Limit Theory of Probability with Applications" at Fudan.

*Proudly hand-crafted: no AI was involved in the making of this lecture notes.

Contents

0 Introduction

社会化就是异化。

I assume that anyone who reads this lecture notes has already finished the course Probability Theory (and most preferably, also Measure Theory). It is clear that “Probability Limit Theory” differs from “Probability Theory” by the additional “limit” in its name. This is what this course is about.

In a more serious manner, we consider random variables indexed by n, our *fundamental large parameter*, and we would like to understand what happens when n gets large. One of the simplest examples is $S_n = X_1 + X_2 + \cdots + X_n$, where $\{X_n\}_{n\geqslant 1}$ are independent, identically distributed random variables. From the central limit theorem (CLT), we know that under the additional assumptions $\mathbb{E}X_1^2 < \infty$ and $\sigma^2 := \mathbb{E}[(X_1 - \mathbb{E}X_1)^2] > 0$, we have

$$\frac{S_n - B_n}{M_n} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty ,$$

where $M_n = \sigma\sqrt{n}$ and $B_n = n\mathbb{E}X_1$.

0.1. Everything but the CLT

The first natural questions is, what happens if $\mathbb{E}X_1^2 = \infty$? This is the motivation of Section 2. We shall see that, without the existence of the second moment, it is not always possible to find M_n and B_n such that $\frac{S_n - B_n}{M_n}$ converges to a non-trivial distribution; if it does converge, the limit might not be Gaussian, but it can only be in a very special class of distributions.

When the second moment is finite, we can without loss of generality, assume that $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = 1$. The CLT says $S_n/\sqrt{n} \xrightarrow{d} \mathcal{N}(0,1)$. This convergence in distribution means

$$\lim_n \mathbb{P}(S_n/\sqrt{n} \leqslant x) = \mathbb{P}(Z \leqslant x) \tag{0.1}$$

for all $x \in \mathbb{R}$. Here Z denotes the standard normal random variable. The relation (0.1) is *qualitative*, which only holds asymptotically as $n \to \infty$. In Section 3, we shall see a *quantitative* version of the CLT, which says under the additional assupmtion $\mathbb{E}|X_1^3| < \infty$, there exists some $C > 0$ independent of n, such that

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(S_n/\sqrt{n} \leqslant x) - \mathbb{P}(Z \leqslant x)\big| \leqslant Cn^{-1/2} \tag{0.2}$$

for all $n \geqslant 1$. Obviously, (0.2) is stronger than (0.1): it gives a rate of convergence.

The CLT estimates the probability that $S_n \sim x\sqrt{n}$ for bounded x, but it does not tell us what happens when $x \to \infty$ as $n \to \infty$. This is another topic that we will be treating in Section 3, called the *large deviation* theory, where for $x \gg 1$, we estimate the probability

$$\mathbb{P}(S_n \geqslant x\sqrt{n}) .$$

The CLT estimates the distribution of $S_n/\sqrt{n}$, but it does not tell us anything about $S_n(\omega)/\sqrt{n}$ for any particular outcome $\omega \in \Omega$. In fact, we can deduce for CLT that

$$\mathbb{P}\Big(\limsup_n \frac{S_n}{\sqrt{n}} = \infty\Big) = 1 , \quad \text{and analogusly} \quad \mathbb{P}\Big(\liminf_n \frac{S_n}{\sqrt{n}} = -\infty\Big) = 1 .$$

In particular, $S_n/\sqrt{n}$ diverges almost surely. In Section 4, we will show that

$$\limsup_n \frac{S_n}{\sqrt{2n\log\log n}} = 1 \quad \text{almost surely}\,,$$

which is know as the *the law of iterated logarithm* (LIL). The LIL tells us about the extreme asymptotic behavior of S_n.

0.2. The Matrix

Almost everything we know about S_n has its analogue in random matrices, i.e. matrices with random coefficients.

Let $H = H^T \in \mathbb{R}^{N\times N}$ be a real symmetric random matrix. We assume its upper triangular entries $(H_{ij})_{i\leqslant j}$ are independent, $\mathbb{E}H_{ij} = 0$, and $\mathbb{E}H_{ij}^2 = N^{-1}$. This is the model of *Wigner matrices*. In the random matrix setting, we use N as our fundamental large parameter, and we denote the (real) eigenvalues of H by $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$. It is easy to see that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\lambda_i^2 = \frac{1}{N}\mathbb{E}\operatorname{Tr} H^2 = \frac{1}{N}\sum_{i,j=1}^{N}\mathbb{E}H_{ij}H_{ji} = 1\,.$$

In other words, the scaling $\mathbb{E}H_{ij}^2 = N^{-1}$ makes the eigenvalues to have typical size 1. We can actually show that the empirical eigenvalue density converges weakly to the semicircle law almost surely, which is equivalent to

$$\frac{1}{N}\sum_{i=1}^{N} f(\lambda_i) \xrightarrow{a.s.} \int_{-2}^{2}\frac{1}{2\pi}\sqrt{4-x^2}f(x)\mathrm{d}x$$

for any smooth test function f. This can be thought as the law of large numbers for the eigenvalues. In addition, we can prove that

$$\sum_{i=1}^{N} f(\lambda_i) - N\int_{-2}^{2}\frac{1}{2\pi}\sqrt{4-x^2}f(x)\mathrm{d}x \xrightarrow{d} \mathcal{N}(0,\sigma_f^2)\,. \tag{0.3}$$

Note that scaling is very different from the classical CLT: it does not have the normalizing factor $N^{-1/2}$. This is because the eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$ are highly correlated. We can furthermore also prove a convergence rate for the CLT (0.3). There are also results on the large deviations and LIL of the eigenvalues, and some of them just came out recently.

We can also talk about the eigenvectors of H. Since H is real symmetric, we have the spectral decomposition

$$H = \sum_{i=1}^{N}\lambda_i \mathbf{u}_i\mathbf{u}_i^T\,,$$

where each $\mathbf{u}_i \in \mathbb{R}^N$ satisfies $\|\mathbf{u}_i\|_2 = 1$. First thing we can prove about $\mathbf{u}$ is *delocalization*, namely, for any $\varepsilon > 0$, we have

$$\max_k |\mathbf{u}_i(k)| \leqslant CN^{-1/2+\varepsilon}$$

with very high probability. In fact, we can further show that

$$\sqrt{N}\mathbf{u}_i(k) \xrightarrow{d} \mathcal{N}(0,1)$$

for all i and k.

Research topics in random matrix theory normally have natural intuitions from classic probability theory, and the questions are hard enough that we are still working on some of them today. In this course, if time permits, I will introduce some of the results to you, and hopefully, also give some ideas of their proofs.

1 Preliminaries

This section recalls some basics in probability theory that we will use later.

1.1. Probability space, random variable, and probability distribution

Recall that the fundamental element in probability theory is the *probability space* $(\Omega, \mathcal{A}, \mathbb{P})$. Here Ω is the *sample space*, which is the set of all possible outcomes of a random process under consideration. The elements of Ω are denoted by ω. $\mathcal{A}$ is a σ-algebra on Ω, which is a collection of subsets of Ω satisfying

(i) $\Omega \in \mathcal{A}$;

(ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$;

(iii) If $A_i \in \mathcal{A}$ for all $i = 1, 2, ...$, then $\cup_i A_i \in \mathcal{A}$.

The elements of $\mathcal{A}$ are called *events*, and $\mathcal{A}$ is often referred as the *event space*. Finally, a function $\mathbb{P} : \mathcal{A} \to \mathbb{R}$ is a *probability measure* on $\mathcal{A}$, which satisfies

(i) $P(A) \geqslant 0$ for all $A \in \mathcal{A}$;

(ii) $\mathbb{P}(\Omega) = 1$;

(iii) For any $A_i \in \mathcal{A}, i = 1, 2, ...$, if $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)\,. \tag{1.1}$$

Remark 1.1. The property (1.1) is called *countable additivity.* It is a sufficient condition of *finite additivity*, which says for any $A_i \in \mathcal{A}, i = 1, 2, ..., n$, if $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathbb{P}(A_i)\,. \tag{1.2}$$

By setting $A_k = \emptyset$ for all $k \geqslant n + 1$, one can obtain (1.2) from (1.1).

Exercise 1.2. We say a function $\mathbb{P} : \mathcal{A} \to \mathbb{R}$ is upper semicontinuous if for any $B_1 \supset \cdots \supset B_n \supset \cdot$ satisfying $\cap_{i\geqslant 1} B_i = \emptyset$, then $\lim_{n\to\infty} \mathbb{P}(B_n) = 0$. Show that $\mathbb{P}$ is countable additive if and only if $\mathbb{P}$ is additive and upper semicontinuous.

Next, X is said to be a *random variable* on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, if it is a function $X : \Omega \to \mathbb{R}$, and it is measurable w.r.t. $\mathcal{A}$. Let $\mathcal{B}$ denotes the Borel set on $\mathbb{R}$, and set $\mathbb{P}_X : \mathcal{B} \to \mathbb{R}$ as $\mathbb{P}_X(B) := \mathbb{P}(\{\omega : X(\omega) \in B\})$. Then we can induce a new probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ from $(\Omega, \mathcal{A}, \mathbb{P})$.

The function $F(x) := \mathbb{P}(X \leqslant x)$ is the *(cumulative) distribution function* of the random variable X. It has the properties

(i) F is non-decreasing, right-continuous;

(ii) $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to\infty} F(x) = 1$.

Conversely, if a function F satisfying the conditions (i) and (ii) above, there exists a random variable X whose distribution function is F.

If two random variables X and Y have the same distribution function, we write $X \overset{d}{=} Y$. But be careful that X and Y do not need to be on the same probability space.

Roughly speaking, there are 3 basic types of random variables. If there is a countable set $B \in \mathbb{R}$, such that $\mathbb{P}(X = x) > 0$ for all $x \in B$, and $\mathbb{P}(X \in B) = 1$, then X is a *discrete* random variable. If $\mathbb{P}(X \in B) = 0$ for any set $B \in \mathbb{R}$ with Lebesgue measure 0, then X is (absolute) *continuous*; if there exists a 0-measure set $B \in \mathbb{R}$ such that $\mathbb{P}(X \in B) = 1$ and $\mathbb{P}(X = x) = 0$ for all x, then X is *singular*.

For an absolute continuous random variable X, its distribution function $F_X(x)$[1] is also absolute continuous. As a result, for all $x \in \mathbb{R}$ we have

$$F_X(x) = \int_{-\infty}^{x} p_X(t)\mathrm{d}t$$

for some nonnegative function p_X. Here p_X is called the *density function* of X.

By Lebesgue Decomposition Theorem, any distribution function F can be written as

$$F(x) = C_1F_1(x) + C_2F_2(X) + C_3F_3(x)\,,$$

where $C_i \geqslant 0$, $\sum_{i=1}^{3} C_i = 1$, and F_1, F_2, F_3 are the distribution functions of a discrete, absolute continuous, singular random variable respectively.

If $X_1, ..., X_n$ are random variables defined on the same probability space, then $\mathbf{X} = (X_1, ..., X_n)$ is a *random vector* with distribution function

$$F(x_1, ..., x_n) := \mathbb{P}(\{X_1 \leqslant x_1, ..., X_n \leqslant x_n\})\,.$$

The *marginal distribution* of X_i is

$$F_i(x_i) := \lim_{x_j\to\infty,\, j\neq i} F(x_1, ..., x_n) = \mathbb{P}(X_i \leqslant x_i)\,.$$

Events $E_1, ..., E_n \in \Omega$ are *independent* if for every $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$, $\mathbb{P}(\bigcap_{j=1}^{k} E_{i_k}) = \prod_{j=1}^{k} \mathbb{P}(E_{i_j})$. If for any Borel sets $B_1, ..., B_n \in \mathbb{R}$, the events $\{X_1 \in B_1\}, ..., \{X_n \in B_n\}$ are independent, then we say that the random variables $X_1, ..., X_n$ are *independent*. Equivalently, random variables $X_1, ..., X_n$ are independent if $F(x_1, ..., x_n) = \prod_{i=1}^{n} F_i(x_i)$.

If $X_1, ..., X_{m+n}$ are independent random variables, and $f : \mathbb{R}^m \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$ are measurable, then $f(X_1, ..., X_m)$ and $g(X_{m+1}, ..., X_{m+n})$ are independent.

Exercise 1.3. If X_1 and X_2 are independent, with distribution functions F_1 and F_2 respectively, prove that

$$F_{X_1+X_2}(x) = \mathbb{P}(X_1 + X_2 \leqslant x) = \int_{\mathbb{R}} F_1(x-u)\mathrm{d}F_2(u)$$

where then RHS of the above is the Riemann-Stieltjes integral. If we further know that X_1 and X_2 have density functions p_1 and p_2 respectively, then

$$p_{X_1+X_2}(x) = \int_{\mathbb{R}} p_1(x-u)p_2(u)\mathrm{d}u =: p_1 * p_2(x)\,.$$

Here the last notation is the *convolution* of p_1 and p_2.

[1] In the squeal, we shall omit this subscript if the dependence is clear.

1.2. Expectation

Suppose X is a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$. If $\int_\Omega |X| \mathrm{d}\mathbb{P} < \infty$, then we say the *expectation* of X exists, and it is defined as

$$\mathbb{E}X := \int_\Omega X \mathrm{d}\mathbb{P} = \int_\Omega X^+ \mathrm{d}\mathbb{P} - \int_\Omega X^- \mathrm{d}\mathbb{P} = \int_\mathbb{R} x \mathrm{d}F_X(x)\,.$$

Recall that $X^+ := \max\{X, 0\}$ and $X^- := \max\{-X, 0\}$. Here $\int \cdot \,\mathrm{d}\mathbb{P}$ are Lebesgue integrals, and $\int \cdot \,\mathrm{d}F(x)$ is the Riemann-Stieltjes integral. As expectation is essentially integration, it is *linear*, in the sense that for any number a, b, we have

$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y\,,$$

provided that $\mathbb{E}X, \mathbb{E}Y$ exist.

More generally, for measurable $g : \mathbb{R} \to \mathbb{R}$, if $\int_\Omega |g(X)| \mathrm{d}\mathbb{P} < \infty$, then

$$\mathbb{E}g(X) = \int_\Omega g(X) \mathrm{d}\mathbb{P} = \int_\mathbb{R} g(x) \mathrm{d}F_X(x)\,.$$

If X has a density function p_X, we can further write $\mathbb{E}g(X) = \int_\mathbb{R} g(x)p_X(x)\mathrm{d}x$. For $k \geqslant 1$, the *k-th moment* of X is denoted as

$$m_k := \mathbb{E}X^k = \int_\mathbb{R} x^k \mathrm{d}F_X(x)\,.$$

The *variance* of X is denoted as

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}X)^2] = m_2 - m_1^2 \geqslant 0\,.$$

The next result is a handy formula for calculating the expectation.

Theorem 1.4. *Suppose the random variable X is nonnegative almost surely (we will abbreviate it as a.s. in the future). Then $\mathbb{E}X$ exists if and only if $\int_0^\infty \mathbb{P}(X \geqslant x)\mathrm{d}x < \infty$, and*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geqslant x)\mathrm{d}x\,.$$

Proof. As X is nonnegative, it is safe to use Fubini's Theorem, which gives

$$\mathbb{E}X = \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \int_0^\infty \mathbf{1}_{x \leqslant X(\omega)} \mathrm{d}x\, \mathrm{d}\mathbb{P}(\omega) = \int_0^\infty \int_\Omega \mathbf{1}_{x \leqslant X(\omega)} \mathrm{d}\mathbb{P}(\omega)\, \mathrm{d}x = \int_0^\infty \mathbb{P}(X \geqslant x)\mathrm{d}x\,.$$

This finishes the proof. □

Exercise 1.5. In the above proof, it is also OK to use the $\mathbf{1}_{x < X(\omega)}$ in the second step, which leads to the result $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x)\mathrm{d}x$. Let us discard the proof, and simply compare the two results. Why does

$$\int_0^\infty \mathbb{P}(X \geqslant x)\mathrm{d}x = \int_0^\infty \mathbb{P}(X > x)\mathrm{d}x\,?$$

Corollary 1.6. *In general, $\mathbb{E}|X| < \infty$ if and only if $\int_0^\infty \mathbb{P}(X \geqslant x)\mathrm{d}x, \int_{-\infty}^0 \mathbb{P}(X \leqslant x)\mathrm{d}x < \infty$, and*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geqslant x)\mathrm{d}x - \int_{-\infty}^0 \mathbb{P}(X \leqslant x)\mathrm{d}x\,.$$

Corollary 1.7. *For $p \in (0, \infty)$, $\mathbb{E}|X|^p < \infty$ if and only if $\sum_{n=1}^{\infty} \mathbb{P}(|X| \geqslant n^{1/p}) < \infty$. It is also equivalent to $\sum_{n=1}^{\infty} n^{p-1}\mathbb{P}(|X| \geqslant n) < \infty$.*

Proof. By Theorem 1.4 and the change of variable $x = y^p$, we have

$$\mathbb{E}|X|^p = \int_0^{\infty} \mathbb{P}(|X|^p \geqslant x)\mathrm{d}x = \int_0^{\infty} py^{p-1}\mathbb{P}(|X| \geqslant y)\mathrm{d}y\,.$$

Observe that

$$\int_0^{\infty} \mathbb{P}(|X| \geqslant x^{1/p})\mathrm{d}x = \sum_{n=0}^{\infty}\int_n^{n+1} \mathbb{P}(|X| \geqslant x^{1/p})\mathrm{d}x \leqslant \sum_{n=0}^{\infty} \mathbb{P}(|X| \geqslant n^{1/p}) = \sum_{n=1}^{\infty} \mathbb{P}(|X| \geqslant n^{1/p}) + 1$$

and

$$\sum_{n=0}^{\infty}\int_n^{n+1} \mathbb{P}(|X| \geqslant x^{1/p})\mathrm{d}x \geqslant \sum_{n=0}^{\infty} \mathbb{P}(|X| \geqslant (n+1)^{1/p}) = \sum_{n=1}^{\infty} \mathbb{P}(|X| \geqslant n^{1/p})\,.$$

Thus $\int_0^{\infty} \mathbb{P}(|X|^p \geqslant x)\mathrm{d}x = \int_0^{\infty} \mathbb{P}(|X| \geqslant x^{1/p})\mathrm{d}x < \infty$ is equivalent to $\sum_{n=1}^{\infty} \mathbb{P}(|X| \geqslant n^{1/p}) < \infty$. In addition,

$$\int_0^{\infty} y^{p-1}\mathbb{P}(|X| \geqslant y)\mathrm{d}y \leqslant \sum_{n=0}^{\infty}(n+1)^{p-1}\mathbb{P}(|X| \geqslant n) \leqslant 2^p \sum_{n=1}^{\infty} n^{p-1}\mathbb{P}(|X| \geqslant n) + 1$$

and

$$\begin{aligned}\int_0^{\infty} y^{p-1}\mathbb{P}(|X| \geqslant y)\mathrm{d}y &\geqslant \sum_{n=0}^{\infty} n^{p-1}\mathbb{P}(|X| \geqslant n+1) = \sum_{n=1}^{\infty}(n-1)^{p-1}\mathbb{P}(|X| \geqslant n) \\ &\geqslant 2^{-p}\sum_{n=1}^{\infty} n^{p-1}\mathbb{P}(|X| \geqslant n) - 1\,.\end{aligned}$$

Hence $\int_0^{\infty} y^{p-1}\mathbb{P}(|X| \geqslant y)\mathrm{d}y < \infty$ is equivalent to $\sum_{n=1}^{\infty} n^{p-1}\mathbb{P}(|X| \geqslant n) < \infty$. ☐

Lemma 1.8. *For any real random variable X, there is al least one $m \in \mathbb{R}$ such that $\mathbb{P}(X \geqslant m), \mathbb{P}(X \leqslant m) \geqslant 1/2$. In this case m is called a* median *of X.*

Proof. Let $x_1 := \inf\{x : \mathbb{P}(X \leqslant x) \geqslant 1/2\}$, then by the definition of x_1, we have

$$\mathbb{P}(X \leqslant x_1) = \lim_{x \downarrow x_1} \mathbb{P}(X \leqslant x) \geqslant 1/2$$

and

$$\mathbb{P}(X < x_1) = \lim_{x \uparrow x_1} \mathbb{P}(X \leqslant x) \leqslant 1/2\,.$$

Thus $\mathbb{P}(X \geqslant x_1) \geqslant 1 - 1/2 = 1/2$. This makes x_1 a median of the random variable X. ☐

Next we introduce the expectation for random vectors $(X_1, ..., X_n)$. Suppose $g : \mathbb{R}^n \to \mathbb{R}$ is Borel measurable, and the random vector has the distribution function $F(x_1, ..., x_n)$, then

$$\mathbb{E}g(X_1, ..., X_n) = \int_{\mathbb{R}^n} g(x_1, ..., x_n)\mathrm{d}F(x_1, ..., x_n)\,.$$

For two random variables X_i, X_j, their *covariance* and *correlation* are defined by

$$\operatorname{Cov}(X_i, X_j) := \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}X_i \mathbb{E}X_j$$

and

$$\operatorname{Corr}(X_i, X_j) := \frac{\operatorname{Cov}(X_i, X_j)}{(\operatorname{Var}(X_i)\operatorname{Var}(X_j))^{1/2}}$$

respectively (If at least one of X_i and X_j is deterministic, then their correlation is set to be 0). It is easy to see that we always have $\operatorname{Corr}(X_i, X_j) \in [-1, 1]$.

If X_i and X_j are independent, then for any measurable function f, we have

$$\mathbb{E}[f(X_i)f(X_j)] = \mathbb{E}f(X_i)\mathbb{E}f(X_j)\,,$$

provided that the expectations exist (Quick question: how to generalize the above relation to $f(X_i)$ and $g(X_j)$?). In particular, $\operatorname{Cov}(X_i, X_j) = 0$ for independent X_i and X_j.

Finally, for a random variable X, recall the *moment generating function*

$$M_X(t) := \mathbb{E}\,\mathrm{e}^{tX}\,.$$

Its name comes from the identity

$$m_k = \mathbb{E}X^k = (\partial_t^k M_X(t))\big|_{t=0}$$

for all $k \geqslant 1$. For independent X_i and X_j, we have $M_{X_i+X_j}(t) = M_{X_i}(t)M_{X_j}(t)$. The moment generating function also has the following nice property.

Theorem 1.9. *For two random variables X and Y, if there exists $c > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in [-c, c]$, then $X \stackrel{d}{=} Y$.*

However, the moment generating function has a huge disadvantage: it does not always exist. Think about a random variable X with the density function $p_X(x) = C\mathrm{e}^{-\sqrt{|x|}}$, where $x \in \mathbb{R}$, and C is a normalization constant. Then all moments of X exist, but $M_X(t) = \infty$ for all $t \neq 0$. In the next session, we recall the mathematical object that is strictly more handy: the Characteristic function.

1.3. Characteristic function

The *characteristic function* of a random variable X is defined by

$$\varphi_X(t) := \mathbb{E}\mathrm{e}^{\mathrm{i}tX} = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}F_X(x)\,.$$

By triangle inequality, $|\varphi_X(t)| = |\mathbb{E}\mathrm{e}^{\mathrm{i}tX}| \leqslant \mathbb{E}|\mathrm{e}^{\mathrm{i}tX}| = 1$. In other words, the characteristic function always exists. We summarize some useful properties of φ in the next result.

Theorem 1.10. *(i) If X and Y are independent, then $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.*

(ii) The characteristic function is uniformly continuous on $\mathbb{R}$.

(iii) For any $k \geqslant 1$, if $\mathbb{E}|X|^k < \infty$, the kth moment can be represented as

$$m_k = \mathbb{E}X^k = (-\mathrm{i})^k \cdot (\partial_t^k \varphi_X(t))\big|_{t=0}\,.$$

(iv) *[6, Theorem 3.3.11] For any random variable X and any $a<b$, we have*

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{it}\varphi_X(t)\mathrm{d}t=\mathbb{P}(X\in(a,b))+\frac{1}{2}(\mathbb{P}(X\in\{a,b\}))\,.$$

In particular, it implies that $\varphi_X=\varphi_Y$ if and only if $X\overset{d}{=}Y$ (why?).

(v) *[6, Theorem 3.3.14] If $\int_{\mathbb{R}}|\varphi_X(t)|\mathrm{d}t<\infty$, then X has bounded continuous density function, which is given by*

$$p_X(x):=\frac{1}{2\pi}\int_{\mathbb{R}}\mathrm{e}^{\mathrm{i}tx}\varphi_X(t)\mathrm{d}t\,.$$

(vi) *The function $\log\varphi_X(t)$ is well-defined when t is near 0. For $\ell\geqslant 1$, if the ℓth moment of X exists, then the first ℓth derivative of $\log\varphi_X$ is finite near 0. In this case, for $k=1,2,...,\ell$, we set*

$$\mathcal{C}_k(X):=(-\mathrm{i})^k\cdot\left(\partial_t^k\log\varphi_X(t)\right)\Big|_{t=0}$$

to be the kth cumulant of X. It is not hard to see that $\mathcal{C}_1(X)=\mathbb{E}X$ and $\mathcal{C}_2(X)=\operatorname{Var}X$.

Remark 1.11. (i) The converse of Theorem 1.10(i) is not true. If $\varphi_{X+Y}=\varphi_X\cdot\varphi_Y$, then we say that X and Y are *subindependent*. For instance, let $X=Y$ with the density function

$$p(x)=\frac{1}{\pi(1+x^2)}\,,$$

which is the Cauchy distribution of scale 1. Then $X+Y=2X$ has density function

$$\widetilde{p}(x)=\frac{2}{\pi(4+x^2)}\,.$$

One can check that $\varphi_{X+Y}(t)=\varphi_X(t)\cdot\varphi_Y(t)=\mathrm{e}^{-2|t|}$.

However, if $\varphi_{aX+bY}=\varphi_{aX}\cdot\varphi_{bY}$ for all $a,b\in\mathbb{R}$, then X and Y are independent. To see this, for a random vector $\mathbf{X}=(X_1,...,X_k)$, we define its characteristic function at $\mathbf{t}=(t_1,...,t_k)$ by

$$\varphi_{\mathbf{X}}(\mathbf{t}):=\mathbb{E}\exp\Big(\mathrm{i}\sum_{i=1}^{k}t_iX_i\Big)\,.$$

As an extension of Theorem 1.10 (iv) to higher dimensions, we can show that

$$\varphi_{\mathbf{X}}=\varphi_{\mathbf{Y}}\quad\text{if and only if}\quad\mathbf{X}\overset{d}{=}\mathbf{Y}.\tag{1.3}$$

Now let $X'\overset{d}{=}X$, $Y'\overset{d}{=}Y$, and X',Y' independent. Then

$$\varphi_{aX+bY}=\varphi_{aX}\cdot\varphi_{bY}=\varphi_{aX'}\cdot\varphi_{bY'}=\varphi_{aX'+bY'}$$

for all $(a,b)\in\mathbb{R}^2$. As a result, $(X,Y)\overset{d}{=}(X',Y')$.

(ii) Although the moment generating function and characteristic function uniquely determines the distribution, the same cannot be said for moments. A classical example is the (perturbed) lognormal distribution X_a, which has the density

$$p_a(x)=\frac{1}{\sqrt{2\pi}x}\exp\Big(-\frac{(\log x)^2}{2\pi}\Big)(1+a\sin(2\pi\log x))\,,x\geqslant 0\,.$$

One can check that, $\mathbb{E}X_a^k=\mathbb{E}X_{a'}^k$ for all $a,a'\in[-1,1]$ and $k\geqslant 1$ (see [6, Section 3.3.5]).

(iii) In part (vi) of Theorem 1.10, it is worth mentioning that it is not possible to define $\log z:\mathbb{C}\backslash\{0\}\to\mathbb{C}$ that is even continuous. However, if $f:\mathbb{R}\to\mathbb{C}$ is differentiable and f never vanishes, $\log\circ f$ is differentiable. In this case, the log function takes value on the Riemann surface.

1.4. Convergence in distribution

Let $F, F_1, F_2, ...,$ be the distribution function of $X, X_1, X_2, ...$ respectively. We say that F_n *weakly converges* to F (write $F_n \xrightarrow{w} F$), if for any continuous points x of F, we have

$$\lim_n F_n(x) = F(x)\,. \tag{1.4}$$

In this case, we say that X_n *converges in distribution* to X, and we write $X_n \xrightarrow{d} X$.

Remark 1.12. It is important to only require the convergence at continuous points of F. Consider random variables X_n uniformly distributed on $[0, 1/n]$. Obviously, $X_n \xrightarrow{d} 0 =: X$. We have $F_X(0) = 1$, while $F_{X_n}(0) = 0$ for all n.

Theorem 1.13. *The following conditions are equivalent.*

(i) $X_n \xrightarrow{d} X$.

(ii) For any bounded continuous function f, we have $\lim_n \mathbb{E}f(X_n) = \mathbb{E}f(X)$. Here the continuous condition can also be replaced with differentiable or smooth.

In reality, the next result is used more often.

Theorem 1.14. *Let $X_1, X_2, ...$ be random variables with distribution function $F_1, F_2, ...$*

(i) If there exits a random variable X with distribution F, such that $F_n \xrightarrow{w} F$, then $\lim_n \varphi_{X_n}(t) = \varphi_X(t)$ for all $t \in \mathbb{R}$.

(ii) If $\lim_n \varphi_{X_n}(t)$ exists for all $t \in \mathbb{R}$ and $\lim_n \varphi_{X_n}$ is continuous at 0, then there exists a random variable X with distribution function F, such that

$$\lim_n \varphi_{X_n}(t) = \varphi_X(t)\,,\ \forall t \in \mathbb{R} \quad \text{and} \quad F_n \xrightarrow{w} F\,.$$

Moreover, (by Arzelà–Ascoli) the convergence of φ_n is uniform on any compact set.

Remark 1.15. (i) Theorem 1.14 (ii) holds in a slightly general form. Let μ_n be a sequence of finite (signed) measures on $\mathbb{R}$ with Fourier–Stieltjes transforms

$$\varphi_n(t) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}\mu_n(x)\,.$$

Suppose φ_n converges to φ, and φ is continuous at 0. Then there exists a finite measure on $\mathbb{R}$ such that $\varphi(t) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}\mu(x)$ and $\mu_n \xrightarrow{w} \mu$. This is known as *continuity theorem for Fourier–Stieltjes transforms.*

(ii) If we go back to the proof of Theorem 1.14 (ii), the condition that $\lim_n \varphi_{X_n}$ is continuous at 0 is essentially due to the *tightness* property. As a counter example, let $X_n := \mathcal{N}(0, n^2)$. Then $\varphi_{X_n}(t) = \exp(-t^2 n^2/2)$. Obviously, (X_n) does not converge, and

$$\lim_{n\to\infty} \varphi_{X_n}(t) = \begin{cases} 1 & \text{if} \quad t = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let us give the proof of Theorem 1.14 (ii) here. We start with the following result.

Theorem 1.16 (Helly)**.** *Let F_n be a sequence of distribution functions on $\mathbb{R}$. Then there exists a subsequence F_{n_k} and a right-continuous, non-decreasing function F on $\mathbb{R}$ such that*

$$\lim_k F_{n_k}(x) = F(x)$$

on all continuous points of F.

Lemma 1.17. *If a sequence of probability measures μ_n on $\mathbb{R}$ is* tight*, i.e.*

$$\forall \varepsilon\,, \quad \exists M > 0\,, \quad \sup_n \mu_n\big([-M, M]^c\big) \leqslant \varepsilon\,,$$

then any subsequential limit of $F_n(x) := \mu_n((-\infty, x])$ is a distribution function[2]*.*

Proof. Let F satisfy

$$F(x) = \lim_k F_{n_k}(x)$$

at all continuous points of F. Since each F_n is non-decreasing, so is F. In addition, it is easy to see that $F \in [0, 1]$ and the discontinuous points of F are at most countable. Suppose F is discontinuous at x_0, which means $F(x_0-) := \sup_{x<x_0} F(x) < F(x_0+) := \inf_{x>x_0} F(x)$. Note that we have the freedom of choosing the value of F at x_0 (it is not decided by F_{n_k}). Set $F(x_0) = F(x_0+)$ makes F right-continuous at x_0. Thus F is always right-continuous and non-decreasing.

Let $\varepsilon > 0$ be given. We can find $M > 0$ such that $\sup_n \mu_n\big([-M, M]^c\big) \leqslant \varepsilon$. W.L.O.G. we can assume that F is continuous at M. Thus

$$\lim_{x\to-\infty} F(x) \leqslant F(-M) = \lim_k F_{n_k}(-M) \leqslant \varepsilon$$

and

$$1 - \lim_{x\to\infty} F(x) \leqslant 1 - F(M) = 1 - \lim_k F_{n_k}(M) \leqslant \varepsilon\,.$$

This implies $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to\infty} F(x) = 1$. Hence F is a distribution function. ☐

Proof of Theorem 1.14 (ii). Let $\varphi := \lim_n \varphi_{X_n}$, and we use μ_n to denote the probability measure on $\mathbb{R}$ induced by F_n.

We first show that $\{\mu_n\}$ is tight. Let $\varepsilon > 0$. Since φ is continuous at 0, we can find a small $u > 0$ such that

$$\varepsilon \geqslant \frac{1}{u}\int_{-u}^{u}(1 - \varphi(t))\mathrm{d}t = \frac{1}{u}\lim_n \int_{-u}^{u}(1 - \varphi_{X_n}(t))\mathrm{d}t\,,$$

where in the second step we used Dominated Convergence Theorem. There exists $n_0 > 0$ such that

$$2\varepsilon \geqslant \lim_n \frac{1}{u}\int_{-u}^{u}(1 - \varphi_{X_n}(t))\mathrm{d}t$$

for all $n \geqslant n_0$. Note that

$$\begin{aligned}\frac{1}{u}\int_{-u}^{u}(1 - \varphi_{X_n}(t))\mathrm{d}t &= \int_{\mathbb{R}} \frac{1}{u}\int_{-u}^{u}(1 - \mathrm{e}^{\mathrm{i}tx})\mathrm{d}t\,\mathrm{d}F_n(x) = 2\int_{\mathbb{R}} 1 - \frac{\sin ux}{ux}\mathrm{d}F_n(x)\\ &\geqslant 2\int_{|x|\geqslant 2u^{-1}} 1 - \frac{1}{|ux|}\mathrm{d}F_n(x) \geqslant \mu_n\{x : |x| \geqslant 2u^{-1}\}\,.\end{aligned}$$

[2] As in (1.4), we only require convergence at points where the limit function is continuous.

Hence $2\varepsilon \geqslant \mu_n\{x : |x| \geqslant 2u^{-1}\}$ for all $n \geqslant n_0$. This shows $\{\mu_n\}$ is tight.

By Theorem 1.16 and Lemma 1.17, there exists a subsequence F_{n_k} that converges to a distribution function F. Thus

$$\varphi_F = \lim_k \varphi_{X_{n_k}} = \lim_n \varphi_{X_n} = \varphi\,.$$

As φ is unique, any subsequent weak limit of F_n must equal to F.

Finally, suppose F_n does not converge to F, then there exists a subsequence F_{n_k} such that at some continuous point x of F, we have

$$|F_{n_k}(x) - F(x)| \geqslant \varepsilon$$

for all $k \geqslant 1$. In other words, any subsequent weak limit of F_{n_k} cannot be F. But per Theorem 1.16 and Lemma 1.17, F_{n_k} possesses a subsequent weak limit. This contradicts to the previous paragraph. Hence we must have

$$F_n \xrightarrow{w} F \quad \text{and} \quad \lim_n \varphi_{F_n} = \varphi_F\,.$$

This finishes the proof of the first statement. The uniform convergence of φ_n on compact intervals is left to the readers as an exercise. □

Let us also mention the multi-dimensional case. Let $F_n(x_1, ..., x_k)$ be a sequence of distribution function on $\mathbb{R}^k$. If there exists a distribution F such that

$$\lim_n F_n(x_1, ..., x_k) = F(x_1, ..., x_k)$$

on all continuous points of F, then we say F_n weakly converges to F. In practice, we use the following characterization in high dimension.

Exercise 1.18. Let $\mathbf{X}_n = (X_{n1}, ..., X_{nk})$ be a sequence of random vectors in $\mathbb{R}^k$. If there exists a random vector $\mathbf{X} = (X_1, ..., X_k)$ such that

$$\sum_{i=1}^k a_i X_{ni} \xrightarrow{d} \sum_{i=1}^k a_i X_i$$

for all $(a_1, ..., a_k) \in \mathbb{R}^k$. Use (1.3) to show that $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.

1.5. Other forms of convergence

Let $X, X_1, X_2, ...$ be random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We say that X_n converges to X *almost surely* (or *a.s.*) if

$$\mathbb{P}\big(\{\omega : \lim_n X_n(\omega) = X(\omega)\}\big) = 1\,.$$

The next result gives a more accessible condition for almost sure convergence [11, Page 52, Theorem 5.2].

Theorem 1.19. *$X_n \to X$ almost surely if and only if for any $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big\{ \bigcup_{m\geqslant n} |X_m - X| \geqslant \varepsilon\Big\}\Big) = 0\,.$$

As a consequence, if $\sum_n \mathbb{P}(|X_n - X| \geqslant \varepsilon) < \infty$ for any $\varepsilon > 0$, or $\sum_n \mathbb{E}|X_n - X|^c < \infty$ for some $c > 0$, then $X_n \to X$ almost surely.

Let $X, X_1, X_2, ...$ be random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We say that X_n converges to X *in probability* (write $X_n \overset{\mathbb{P}}{\to} X$), if

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \varepsilon) = 0$$

for any $\varepsilon > 0$. For $p > 0$, we say that X_n converges to X in L_p (write $X_n \overset{L_p}{\to} X$), if

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0\,.$$

Theorem 1.20. *There are various relations between different types of convergence of random variables.*

(i) $X_n \to X$ *a.s. implies* $X_n \overset{\mathbb{P}}{\to} X$.

(ii) $X_n \overset{L_p}{\to} X$ *implies* $X_n \overset{\mathbb{P}}{\to} X$

(iii) $X_n \overset{\mathbb{P}}{\to} X$ *implies* $X_n \overset{d}{\to} X$.

(iv) If $X_n \overset{d}{\to} C$ *for some constant* C*, then* $X_n \overset{\mathbb{P}}{\to} C$.

Exercise 1.21. Give examples that the converses of Theorem 1.20 (i) – (iii) are not true. Give examples to show that there are in general no relation between $X_n \to X$ a.s. and $X_n \overset{L_p}{\to} X$.

2 Infinite divisible distributions and the universality of random variable sums

Think about a collection of random sequence $\mathbf{X}_1, \mathbf{X}_2, ...,$ where

$$\mathbf{X}_n = (X_{n1}, ..., X_{nk_n}) .$$

We assume that for each n, the random variables $X_1^{(n)}, ..., X_{k_n}^{(n)}$ are independent, and $\lim_n k_n = \infty$. In this section, we are interested in studying the limit of

$$S_n := \sum_{k=1}^{k_n} X_{nk} . \tag{2.1}$$

In Section 2.2, we shall see that under mild conditions, if S_n converges, the limit always lies in a special class of distributions that are "infinitely divisible". Thus in Section 2.1, we first introduce properties of infinitely divisible distributions. In Section 2.3, we consider the special case $k_n = n$, and completely characterize the possible limiting distributions of S_n.

2.1. Infinite divisible distributions

A characteristic function φ is *infinitely divisible*, if for any $n \geqslant 1$, there exists a characteristic function φ_n such that

$$\varphi(t) = (\varphi_n(t))^n \tag{2.2}$$

for all $t \in \mathbb{R}$. In this case, the corresponding distribution F of φ is called an infinitely divisible distribution. In other words, a random variable X is infinitely divisible, if for any $n \geqslant 1$, we can find i.i.d. random variables $X_{n1}, ..., X_{nn}$ such that $X \stackrel{d}{=} X_{n1} + \cdots X_{nn}$.

Example 2.1. The following class of random variables are infinitely divisible.

(i) A degenerate random variable X that only takes value at point c, i.e. $\mathbb{P}(X = c) = 1$. Then

$$\varphi_X(t) = \mathrm{e}^{\mathrm{i}ct} = \left(\mathrm{e}^{\mathrm{i}\frac{c}{n}t}\right)^n .$$

(ii) Let $X \stackrel{d}{=} \mathcal{N}(\mu, \sigma^2)$. Then

$$\varphi_X(t) = \exp(\mathrm{i}\mu t - \sigma^2 t^2/2) = \left(\exp\left(\mathrm{i}t\frac{\mu}{n} - \frac{\sigma^2}{n} \cdot \frac{t^2}{2}\right)\right)^n .$$

(iii) Let $X \stackrel{d}{=} \mathrm{Poi}(\lambda)$, i.e. Poisson distribution with parameter λ. Then

$$\varphi_X(t) = \exp(\lambda(\mathrm{e}^{\mathrm{i}t} - 1)) = \left(\exp\left(\frac{\lambda}{n}(\mathrm{e}^{\mathrm{i}t} - 1)\right)\right)^n$$

(iv) Let X be Cauchy-distributed with parameter $\gamma > 0$, i.e. X has density function $p_X(x) = \frac{1}{\pi} \cdot \frac{\gamma}{\gamma^2 + x^2}$, $x \in \mathbb{R}$. Then

$$\varphi_X(t) = \mathrm{e}^{-\gamma|t|} = \left(\mathrm{e}^{-\frac{\gamma}{n}|t|}\right)^n .$$

Theorem 2.2. *(i) If φ_1,...,φ_k are infinitely divisible characteristic functions (IDCF), so is $\varphi_1 \cdots \varphi_k$.*

(ii) If φ is an IDCF, so is $|\varphi|$.

Proof. (i) Let us prove the case when $k = 2$; the general case follows by induction.

Let φ_1 and φ_2 be the characteristic functions of X and Y respectively. W.L.O.G. assume X and Y are independent. Then $\varphi_1\varphi_2$ is the characteristic function of $X + Y$. In particular, we see that the multiplication of characteristic functions is again a characteristic functions.

Let $n \geqslant 1$. Since φ_1 and φ_2 are infinitely divisible, there exists characteristic functions ϕ_1, ϕ_2 such that $\varphi_1 = \phi_1^n$, $\varphi_2 = \phi_2^n$. Then $\phi_1\phi_2$ is also a characteristic function, with

$$\varphi_1\varphi_2 = (\phi_1\phi_2)^n \,.$$

This finishes the proof.

(ii) Let $n \geqslant 1$, then there exits a characteristic function ϕ such that $\varphi = \phi^{2n}$. If we write $\phi(t) = \mathbb{E}e^{itX}$, then $\overline{\phi(t)} = \mathbb{E}e^{-itX} = \mathbb{E}e^{it(-X)}$ is also a characteristic function. By (i), $\phi\overline{\phi}$ is a characteristic function, thus $|\varphi| = |\phi^{2n}| = (\phi\overline{\phi})^n$ is an IDCF. □

Theorem 2.3. *Let φ be an IDCF. Then $\varphi(t) \neq 0$ for all $t \in \mathbb{R}$.*

Proof. By Theorem 2.2 (ii), $|\varphi|$ is also an IDCF. Let ϕ_n be a characteristic function such that $|\varphi| = \phi_n^n$. If we examine the proof of Theorem 2.2 (ii), ϕ_n can be chosen to be the nonnegative nth root of $|\varphi|$. Thus

$$\phi(t) := \lim_{n\to\infty} \phi_n = \begin{cases} 1 & \text{if} \quad \varphi(t) \neq 0 \\ 0 & \text{if} \quad \varphi(t) = 0\,. \end{cases} \tag{2.3}$$

As $\varphi(0) = 1$ and φ is continuous, $\varphi(t)$ is non-zero in a neighborhood of 0. Thus $\phi(t) = 1$ in a neighborhood of 0. In particular, ϕ is continuous at 0. By Theorem 1.14 (ii), ϕ is a characteristic function, which implies ϕ is continuous. By (2.3), ϕ can only take value $0, 1$. Since $\phi(0) = 1$, we must have $\phi(t) = 1$ for all $t \in \mathbb{R}$. Using (2.3) again, we see that $\varphi(t) \neq 0$ for all $t \in \mathbb{R}$. □

Remark 2.4. In (2.2), it is not instantly clear which phase φ_n sits among the n possible ones. As φ is never 0, we can write $\varphi(t) = \exp(\log\varphi(t))$, where $\log\varphi(t)$ takes value in the Riemann surface and is continuous w.r.t. t. Let $\phi_n(t) := \exp(\frac{1}{n}\log\varphi(t))$, where we choose $\log\varphi(0) = \log 1 = 0$. Then φ_n and ϕ_n are both continuous, $(\varphi_n(t))^n = (\phi_n(t))^n$ for all t, and $\varphi_n(0) = \phi_n(0) = 1$. Define

$$\omega(t) := \varphi_n(t)/\phi_n(t)\,,$$

then $(\omega(t))^n = 1$ for all t. Thus for each t, $\omega(t) = \exp(2k\pi i/n)$ for some $k \in \{0, 1, ..., n-1\}$. As ω is continuous and $\omega(0) = 1$, we must have $\omega \equiv 1$. As a result

$$\varphi_n(t) = \phi_n(t) = \exp\left(\frac{1}{n}\log\varphi(t)\right).$$

In the sequel, for a characteristic function $\varphi = \exp(\log\varphi(t))$, we sometime use the notation $\varphi^{1/n}$ to denote $\varphi^{1/n}(t) := \exp(\frac{1}{n}\log\varphi(t))$, where $\log\varphi(0) = 0$.

Theorem 2.5. *If a sequence of IDCF $(\varphi_m)_{m\geqslant 1}$ converges to the characteristic function φ, then φ is also infinitely divisible.*

Proof. Let $n \geqslant 1$. As φ_m is infinitely divisible, $\varphi_m^{1/n} = \exp(\frac{1}{n}\log\varphi_m(t))$ is also a characteristic function. Since $\lim_m \varphi_m = \varphi$, $\lim_m \varphi_m^{1/n}(t) = \exp(\frac{1}{n}\log\varphi(t))$[3]. As φ is a characteristic function, φ is continuous near 0, hence $\exp(\frac{1}{n}\log\varphi(t))$ is also continuous near 0. By Theorem 1.14 (ii), $\exp(\frac{1}{n}\log\varphi(t))$ is a characteristic function. This finishes the proof. □

[3]The limit is clearly well-defined when $\varphi(t) \neq 0$; when $\varphi(t) = 0$, we simply define the limit to be 0

For the rest of Section 2.1, we shall prove the following result.

Theorem 2.6. *The function φ is an IDCF if and only if it can be represented as*

$$\varphi(t) = \exp\left(\mathrm{i}at + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x)\right),$$

where $a \in \mathbb{R}$ is a constant, and G is a bounded, non-decreasing, right-continuous function. In addition, for $g(x) := \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}$, define

$$g(0) := \lim_{x\to 0} g(x) = -\frac{t^2}{2}$$

so that g is continuous in $\mathbb{R}$.

The next section consists of some preparation work.

2.1.1. Poisson-type representation. Let $\lambda \geqslant 0$, $\alpha, \beta \in \mathbb{R}$. The random variable X is called *Poisson type* random variable, if

$$\mathbb{P}(X = \alpha + \beta k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for all $k \in \mathbb{N}$. We write $X := \mathrm{Poi}(\lambda, \alpha, \beta)$. It is not hard to see that

$$\varphi_X(t) = \exp\{\mathrm{i}\alpha t + \lambda(\mathrm{e}^{\mathrm{i}\beta t} - 1)\}\,.$$

Proposition 2.7. *Let φ be a characteristic function. It is infinitely divisible if and only if it is the limit of finite multilpplicaions of the characteristic functions of Poisson type random variables.*

Proof. (i) Suppose $\varphi = \lim_m \varphi_m$, where

$$\varphi_m(t) = \exp\{\mathrm{i}\alpha_1 t + \lambda_1(\mathrm{e}^{\mathrm{i}\beta_1 t} - 1)\}\cdots\exp\{\mathrm{i}\alpha_n t + \lambda_n(\mathrm{e}^{\mathrm{i}\beta_n t} - 1)\}\,,$$

and $\alpha_i, \beta_i, \lambda_i$ are parameters possibly depend on m, then by Theorem 2.2 (i), φ_m is infinitely divisible. Together with Theorem 2.5, $\lim_m \varphi_m$ is also infinitely divisible.

(ii) Let φ be an IDCF. By Theorem 2.3 and Remark 2.4,

$$\varphi^{1/n} = \exp\left(\frac{1}{n}\log\varphi\right) = 1 + \frac{1}{n}\log\varphi + O(n^{-2})\,.$$

Thus for any $t \in \mathbb{R}$,

$$\begin{aligned}\log\varphi(t) &= \lim_{n\to\infty} n(\varphi^{1/n} - 1) = \lim_{n\to\infty} n(\varphi_n - 1)\\ &= \lim_{n\to\infty} n\int_{\mathbb{R}}(\mathrm{e}^{\mathrm{i}tx} - 1)\mathrm{d}F_n(x) = \lim_{n,m\to\infty} n\int_{-m}^{m}(\mathrm{e}^{\mathrm{i}tx} - 1)\mathrm{d}F_n(x)\,,\end{aligned}$$

where $\varphi_n := \varphi^{1/n}$ is a characteristic function, and we denote the corresponding distribution by F_n. Let $-m = x_0 < x_1 < \cdots < x_N = m$, $\lim_N \max_j(x_{j+1} - x_j) = 0$. By the definition of Riemann-Stieltjes integral,

$$\begin{aligned}n\int_{-m}^{m}(\mathrm{e}^{\mathrm{i}tx} - 1)\mathrm{d}F_n(x) &= \lim_{N\to\infty} n\sum_{j=0}^{N-1}(\mathrm{e}^{\mathrm{i}tx_j} - 1)\big(F_n(x_{j+1}) - F_n(x_j)\big)\\ &= \lim_{N\to\infty}\sum_{j=0}^{N-1}\lambda_j(\mathrm{e}^{\mathrm{i}t\beta_j} - 1)\,,\end{aligned}$$

where $\lambda_j = n(F_n(x_{j+1}) - F_n(x_j))$, $\beta_j = x_j$. Thus

$$\varphi(t) = \lim_{n,m,N\to\infty} \prod_{j=1}^{N} \exp\left(\lambda_j(\mathrm{e}^{\mathrm{i}t\beta_j} - 1)\right).$$

This finises the proof. □

2.1.2. The Lévy-Khintchine representation for infinite divisible characteristic functions.

Let $a \in \mathbb{R}$, and $G(x)$ is a bounded, non-decreasing, right-continuous function on $\mathbb{R}$. Define

$$\xi(t) = \mathrm{i}at + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x). \tag{2.4}$$

The next result proves the "if" part of Theorem 2.6.

Proposition 2.8. *The function* $\mathrm{e}^{\xi(t)}$ *is an IDCF.*

Proof. By Riemann-Stieltjes integral, we have

$$\int_0^\infty \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x) = \lim_{m\to\infty}\int_{1/m}^{m} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x)$$
$$= \lim_{m,N\to\infty}\sum_{j=0}^{N-1}\left(\mathrm{e}^{\mathrm{i}tx_j} - 1 - \frac{\mathrm{i}tx_j}{x_j^2+1}\right)\frac{1+x_j^2}{x_j^2}\left(G(x_{j+1}) - G(x_j)\right),$$

where $1/m = x_0 < x_1 < \cdots < x_N = m$, $\lim_N \max_j(x_{j+1} - x_j) = 0$. Set

$$\lambda_j = \frac{1+x_j^2}{x_j^2}\left(G(x_{j+1}) - G(x_j)\right), \quad \beta_j = x_j, \quad \alpha_j = -\frac{\lambda_j x_j}{1+x_j^2},$$

then

$$I_1 := \int_0^\infty \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x) = \lim_{m,N\to\infty}\sum_{j=0}^{N-1}\left(\mathrm{i}\alpha_j t + \lambda_j(\mathrm{e}^{\mathrm{i}t\beta_j} - 1)\right).$$

The same can be done for $I_2 := \int_{-\infty}^0 g(x)\mathrm{d}G(x)$. Note that

$$\xi(t) = \mathrm{i}at + I_1 + I_2 - \frac{t^2}{2}[G(0) - G(0-)],$$

and we conclude the proof by Theorem 2.2 (i) and Proposition 2.7. □

To prove the "only if" part of Theorem 2.6, recall the definition of ξ from (2.4), and let us introduce

$$\Lambda(x) = \int_{-\infty}^{x}\left(1 - \frac{\sin y}{y}\right)\frac{1+y^2}{y^2}\mathrm{d}G(y) \tag{2.5}$$

and

$$\lambda(t) = \xi(t) - \int_0^1 \frac{\xi(t+h) + \xi(t-h)}{2}\mathrm{d}h, \tag{2.6}$$

where $A(y) := \big(1-\frac{\sin y}{y}\big)\frac{1+y^2}{y^2}$ is set to be $\lim_{y\to 0} A(y) = 1/6$ when $t=0$. Clearly, $A(t)$ is nonnegative and bounded, which makes $\Lambda(x)/\Lambda(\infty)$ a distribution function. In addition, by Fubini's Theorem,

$$\begin{aligned}\lambda(t) &= \int_0^1 \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx}(1-\cos hx)\frac{1+x^2}{x^2}\mathrm{d}G(x)\mathrm{d}h = \int_{\mathbb{R}}\int_0^1 \mathrm{e}^{\mathrm{i}tx}(1-\cos hx)\frac{1+x^2}{x^2}\mathrm{d}h\mathrm{d}G(x) \\ &= \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx}\Big(1-\frac{\sin x}{x}\Big)\frac{1+x^2}{x^2}\mathrm{d}G(x) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}\Lambda(x)\,.\end{aligned} \tag{2.7}$$

Thus $\lambda(t)/\Lambda(\infty)$ is the characteristic function of the distribution $\Lambda(x)/\Lambda(\infty)$.

Lemma 2.9. *Let $G(-\infty)=0$. Then there is a one-to-one correspondence between ξ and (a,G). For this reason, we write $\xi=(a,G)$.*

Proof. Obviously, (a,G) uniquely determines ξ. Let us prove the other direction. Given a function $\xi(t)$, by (2.6), it uniquely determines $\lambda(t)$. From (2.7) and the inversion formula (Theorem 1.10 (iii)), $\lambda(t)$ determines $\Lambda(x)$. Then we can deduce G from Λ using

$$G(x) := \int_{-\infty}^{x}\Big(1-\frac{\sin y}{y}\Big)^{-1}\frac{y^2}{1+y^2}\mathrm{d}\Lambda(y)\,.$$

Finally, (2.4) shows that G, ξ determines a. This finishes the proof. (Quick question: where did we use the condition $G(-\infty)=0$?) □

Lemma 2.10. *Let*

$$\xi_n(t) = \mathrm{i}a_n t + \int_{\mathbb{R}}\Big(\mathrm{e}^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{x^2+1}\Big)\frac{1+x^2}{x^2}\mathrm{d}G_n(x)\,, \tag{2.8}$$

where $a_n\in\mathbb{R}$, G_n is a bounded, non-decreasing, right-continuous function satisfying $G_n(-\infty)=0$.

(i) If $a_n\to a$ and $G_n \overset{w}{\to} G$, then $\xi_n(t)\to\xi(t)=(a,G)$.

(ii) If $\xi_n(t)\to\xi(t)$ and ξ is continuous, then there exists constant a and bounded, non-decreasing, right-continuous G such that $\xi=(a,G)$, $a_n\to a$, $G_n\overset{w}{\to} G$.

Proof. Part (i) is immediate from Theorem 1.13, and let us prove part (ii).

By Proposition 2.8, $\exp(\xi_n(t))$ is an IDCF. Since $\exp(\xi_n(t))$ is a characteristic function and $\exp(\xi(t))$ is continuous at 0, by Theorem 1.14 (ii), $\exp(\xi(t))$ is also a characteristic function. Since $\exp(\xi_n(t))$ is infinitely divisible, by Theorem 2.5, $\exp(\xi(t))$ is also infinitely divisible. Let I be a compact interval that contains the origin. By Theorem 2.3, $\exp(\xi(t))$ is never 0. Then there exists an $\varepsilon>0$ such that

$$\min_{t\in I}|\exp(\xi(t))|\geqslant\varepsilon\,. \tag{2.9}$$

Since the sequence of characteristic functions $\exp(\xi_n(t))$ converges to the characteristic function $\exp(\xi(t))$, by Theorem 1.14 (ii),

$$\exp(\xi_n(t))\to\exp(\xi(t))\quad\text{uniformly for}\quad t\in I\,. \tag{2.10}$$

Combining (2.9) and (2.10) we see that the function

$$h_n(t) := \exp(\xi_n(t)-\xi(t))$$

converges uniformly to 1 on I. Thus for n large enough, $|h_n(t) - 1| < 1/2$ for all $t \in I$. On the disk $|z - 1| < 1/2$, the logarithm is uniformly continuous and uniquely determined by continuity. As $\xi_n(0) = \xi(0) = 0$, we must have

$$\xi_n(t) \to \xi(t) \quad \text{uniformly for} \quad t \in I\,. \tag{2.11}$$

As I is arbitrary, by (2.6), we have

$$\lambda_n(t) \to \lambda(t) \tag{2.12}$$

for all $t \in \mathbb{R}$. Since ξ is also continuous, by (2.6), λ is also continuous in $\mathbb{R}$.

By (2.12) and the continuity theorem for Fourier-Stieltjes transform (see Remark 1.15 (ii)), there exists a finite measure Λ satisfying $\lambda(t) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}\Lambda(x)$ such that

$$\Lambda_n \xrightarrow{w} \Lambda\,. \tag{2.13}$$

Now G is defined from Λ via

$$G(x) := \int_{-\infty}^{x} \Big(1 - \frac{\sin y}{y}\Big)^{-1} \frac{y^2}{1+y^2} \mathrm{d}\Lambda(y)\,,$$

and thus (2.13) implies $G_n(x) \to G(x)$. Finally, (2.8) we get

$$a_n = (\mathrm{i}t)^{-1}\bigg(\xi_n(t) - \int_{\mathbb{R}} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\Big)\frac{1+x^2}{x^2}\mathrm{d}G_n(x)\bigg)$$

for all $t \neq 0$. As the RHS of the above is independent of t and converges, we define its limit to be a. This implies $a_n \to a$ and finishes the proof. □

Now we are ready to prove the "only if" part of Theorem 2.6. Let $\varphi(t)$ be an IDCF. As in the proof of Proposition 2.7 (ii),

$$\begin{aligned}\log \varphi(t) &= \lim_{n\to\infty} n(\varphi^{1/n}(t) - 1) = \lim_{n\to\infty} n(\varphi_n(t) - 1) = \lim_n \int_{\mathbb{R}} n(\mathrm{e}^{\mathrm{i}tx} - 1)\mathrm{d}F_n(x)\\ &= \lim_n \bigg(\mathrm{i}tn \int_{\mathbb{R}} \frac{x}{1+x^2}\mathrm{d}F_n(x) + \int_{\mathbb{R}} n\Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Big)\mathrm{d}F_n(x)\bigg) =: \lim_n \xi_n(t)\,,\end{aligned}$$

where

$$a_n = n\int_{\mathbb{R}} \frac{x}{1+x^2}\mathrm{d}F_n(x)\,, \quad G_n(x) = n\int_{-\infty}^{x} \frac{y^2}{1+y^2}\mathrm{d}F_n(y)\,.$$

As $\lim_n \xi_n(t) = \log\varphi(t)$ is continuous, by Lemma 2.10 (ii), there exits (a, G) such that $a_n \to a$, $G_n \to G$, and $\log\varphi(t) = (a, G)$. This finishes the proof of Theorem 2.6.

Example 2.11. Let us see the convergence of a_n and G_n through an example. Consider the Cauchy distribution with median α and scale 1, which has the distribution function

$$F(x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{1+(y-\alpha)^2}\mathrm{d}y\,,$$

which is infinitely divisible with characteristic function $\varphi(t) = \exp(-|t| - \mathrm{i}\alpha t)$. Clearly, $\varphi_n(t) = \varphi^{1/n}(t) = \exp(-(|t| + \mathrm{i}\alpha t)/n)$, and

$$F_n(x) = \frac{1}{n\pi}\int_{-\infty}^{x} \frac{1}{n^{-2} + (y - \alpha/n)^2}\mathrm{d}y\,.$$

In other words, it is a Cauchy distribution with median α/n and scale n^{-1}. Thus

$$G_n(x) = n\int_{-\infty}^x \frac{y^2}{1+y^2}\mathrm{d}F_n(y) = n\int_{-\infty}^x \frac{y^2}{1+y^2}\cdot\frac{1}{n\pi}\frac{1}{n^{-2}+(y-\alpha/n)^2}\mathrm{d}y \to \frac{1}{\pi}\int_{-\infty}^x \frac{1}{1+y^2}\mathrm{d}y =: G(x)\,,$$

and

$$a = \lim_n n\int_{\mathbb{R}} \frac{x}{1+x^2}\mathrm{d}F_n(x) = \lim_n n\int_{\mathbb{R}} \frac{x}{1+x^2}\frac{1}{n\pi}\frac{1}{n^{-2}+(x-\alpha/n)^2}\mathrm{d}x = \lim_n n\frac{\alpha n}{\alpha^2+(1+n)^2} = \alpha\,.$$

2.2. The sum of independent random variables.

Now let us come back to the original question at the beginning of Section 2, which is

$$S_n := \sum_{k=1}^{k_n} X_{nk}\,,$$

where $\{X_{nk}; k=1,2,...,k_n\}$ are independent and $\lim_n k_n = \infty$. Note that for an arbitrary random variable X, if we set $X_{n1} \overset{d}{=} X$ and $X_{nk}=0$ for all $k\geqslant 2$, then $S_n \overset{d}{\equiv} X$. So if we would like to produce relevant results, certain restrictions are needed on the random variables. This motivates the following definition.

Definition 2.12. We say the random variables $\{X_{nk}; k=1,2,...,k_n\}$ satisfy the *infinitesimal* condition, if for any $\varepsilon>0$, we have

$$\lim_{n\to\infty}\max_{1\leqslant k\leqslant k_n}\mathbb{P}(|X_{nk}|\geqslant\varepsilon) = 0\,.$$

There are some priori results where we leave the proofs to the reader.

Lemma 2.13. *The following conditions are equivalent.*

(i) $\{X_{nk}\}$ *satisfy the infinitesimal condition;*

(ii) $\max_{1\leqslant k\leqslant k_n}\int_{\mathbb{R}}\frac{x^2}{1+x^2}\mathrm{d}F_{nk}(x)\to 0$*;*

(iii) For any bounded set $S\subset\mathbb{R}$*, we have* $\max_{1\leqslant k\leqslant k_n}|\varphi_{nk}(t)-1|\to 0$ *uniformly for* $t\in S$*.*

Lemma 2.14. *If* $\{X_{nk}\}$ *satisfy the infinitesimal condition, then for any* $\tau, r>0$*, we have*

$$\max_k \int_{|x|<\tau}|x|^r\mathrm{d}F_{nk}(x)\to 0\,.$$

The main goal of Section 2.2 is to prove the following result.

Theorem 2.15. *Let* $\{X_{nk}\}$ *satisfy the infinitesimal condition, and* $S_n := \sum_{k=1}^{k_n} X_{nk}$ *converges. The limiting distribution class of* S_n *coincides with the set of infinitely divisible distributions.*

The proof consists of two steps, and similar to that of Theorem 2.6, one direction is easy.

Step 1. Let F be an infinitely divisible distribution, with characteristic function φ. Then $\varphi^{1/n}$ is also a characteristic function. Set $k_n=n$ and $\varphi_{nk}=\varphi^{1/n}$ for all $k=1,2...,n$, and we have $\varphi = \prod_k \varphi_{nk}$. Thus there exists independent random variables $\{X_{nk}\}_{k=1,2,\dots,n}$ such that $\mathbb{E}\exp(\mathrm{i}tX_{nk}) = \varphi_{nk}(t)$, and

$$S_n = \sum_{k=1}^{n} X_{nk}$$

has the distribution F. Recall from Theorem 2.3 that $\varphi(t) \neq 0$ for all $t \in \mathbb{R}$, thus for any bounded $S \subset \mathbb{R}$,

$$\max_k |\varphi_{nk}(t) - 1| = |\varphi_{n1}(t) - 1| = |\varphi^{1/n}(t) - 1| \to 0$$

uniformly for all $t \in S$. By Lemma 2.13, $\{X_{nk}\}$ is infinitesimal. Thus we have proved that any infinitely divisible distributions is the limiting distribution of some $S_n = \sum_{k=1}^{k_n} X_{nk}$.

Step 2. The proof of the other direction is substantially more complicated. Let $\{X_{nk}\}$ be infinitesimal, and $S_n := \sum_{k=1}^{k_n} X_{nk}$ converges to some distribution F. We would like to show that F is infinitely divisible. Let $\varphi_n(t) = \mathbb{E}\exp(\mathrm{i}tS_n)$, $\varphi_{nk}(t) = \mathbb{E}\exp(\mathrm{i}tX_{nk})$, and φ denotes the characteristic function of F. Clearly $\lim_n \varphi_n = \varphi$. Our goal here is to show that F is infinitely divisible. It suffices to show that

$$\log \varphi_n(t) = \log \prod_{k=1}^{k_n} \varphi_{nk}(t) = \mathrm{i}a_n t + \int_{\mathbb{R}} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\Big) \frac{1+x^2}{x^2} \mathrm{d}G_n(x) + \mathcal{E}_n(t) \tag{2.14}$$

for some $a_n \in \mathbb{R}$ and bounded, non-decreasing, right-continuous function G_n, and $\mathcal{E}_n(t) \to 0$. More precisely, Theorem 2.6 implies that $\xi_n = (a_n, G_n)$ is the logarithm of an IDCF. By (2.14), we have

$$\varphi(t) = \lim_n \varphi_n(t) = \lim_n \exp(\log \varphi_n(t)) = \lim_n \exp(\xi_n(t))\,. \tag{2.15}$$

Hence by Theorem 2.5, φ is also an IDCF.

Now let us prove (2.14). Let S be a bounded set in $\mathbb{R}$. By Lemma 2.13 (iii) we have (for n large enough)

$$\log \varphi_{nk}(t) = \log[1 + (\varphi_{nk}(t) - 1)] = \varphi_{nk}(t) - 1 + \theta_{nk}(\varphi_{nk}(t) - 1)^2$$

uniformly for $t \in S$. Here $\theta_{nk} \in \mathbb{C}$ satisfies $|\theta_{nk}| \in [0,1]$. Thus

$$\begin{aligned}
\log \varphi_n(t) &= \sum_{k=1}^{k_n} [\varphi_{nk}(t) - 1 + \theta_{nk}(\varphi_{nk}(t) - 1)^2] \\
&= \sum_{k=1}^{k_n} \int_{\mathbb{R}} (\mathrm{e}^{\mathrm{i}tx} - 1) \mathrm{d}F_{nk}(x) + \sum_{k=1}^{k_n} \theta_{nk}(\varphi_{nk}(t) - 1)^2 \\
&= \mathrm{i}t \sum_{k=1}^{k_n} \int_{\mathbb{R}} \frac{x}{1+x^2} \mathrm{d}F_{nk}(x) + \sum_{k=1}^{k_n} \int_{\mathbb{R}} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Big) \mathrm{d}F_{nk}(x) + \sum_{k=1}^{k_n} \theta_{nk}(\varphi_{nk}(t) - 1)^2\,.
\end{aligned} \tag{2.16}$$

Then it is tempting to define

$$a_n := \sum_{k=1}^{k_n} \int_{\mathbb{R}} \frac{x}{1+x^2} \mathrm{d}F_{nk}(x) \quad \text{and} \quad G_n(x) = \sum_{k=1}^{k_n} \int_{-\infty}^{x} \frac{y^2}{1+y^2} \mathrm{d}F_{nk}(y)\,.$$

However, it is not necessarily true that the last term on RHS of (2.16) is negligible. Let $k_n = n$. Inspired by Example 2.11, consider X_{nk} with Cauchy distribution having median $(-1)^k n^{-1/2}$ and scale $1/n$, i.e.

$$F_{nk}(x) = \frac{1}{n\pi} \int_{-\infty}^{x} \frac{1}{n^{-2} + (y - (-1)^k n^{-1/2})^2} \mathrm{d}y\,. \tag{2.17}$$

Clearly, $X_{n1} + \cdots + X_{nn}$ converges to the Cauchy distribution with median 0 and scale 1. Note that

$$\varphi_{nk}(t) = \exp\Big(-\frac{|t|}{n} - (-1)^k \frac{\mathrm{i}t}{\sqrt{n}}\Big)\,,$$

and thus $\varphi_{nk}(t) = 1 - (-1)^k \frac{\mathrm{i}t}{\sqrt{n}} + O(n^{-1})$. This leads to

$$\sum_{k=1}^{n} (\varphi_{nk}(t) - 1)^2 = \sum_{k=1}^{n} \Big((-1)^k \frac{\mathrm{i}t}{\sqrt{n}} + O(n^{-1}) \Big)^2 = -t^2 + O(n^{-1/2})\,,$$

which is nontrivial. This suggests that we need to shift the median of X_{nk} to 0 in our consideration. As X_{nk} is infinitesimal, the median mainly comes from the regime near 0.

In this spirit, fix a small $\tau > 0$, and define

$$\alpha_{nk} := \int_{-\tau}^{\tau} x \mathrm{d}F_{nk}(x)\,, \quad \underline{F}_{nk}(x) := F_{nk}(x + \alpha_{nk})\,. \tag{2.18}$$

Let $\underline{\varphi}_{nk}$ denotes the characteristic function of $\underline{F}_{nk}$. We have the following result.

Lemma 2.16. *Let $\{X_{nk}\}$ satisfy the infinitesimal condition. Then for any $t \in \mathbb{R}$, we have*

$$\sum_{k=1}^{k_n} (\log \underline{\varphi}_{nk}(t) - (\underline{\varphi}_{nk}(t) - 1)) \to 0\,.$$

The proof of Lemma 2.16 is technical and we omit here. Interested reader may refer to pages 38-43 of [11].

Exercise 2.17. Verify Lemma 2.16 for X_{nk} given in (2.17).

Now let us continue the proof of (2.14). Note that $\log \underline{\varphi}_{nk}(t) = \log \varphi_{nk}(t) - \mathrm{i}\alpha_{nk} t$. Thus $\log \varphi_n(t)$ can be written as

$$\begin{aligned}
\log \prod_{k=1}^{k_n} \varphi_{nk}(t) &= \sum_{k=1}^{k_n} \log \varphi_{nk}(t) = \sum_{k=1}^{k_n} \log \underline{\varphi}_{nk}(t) + \sum_{k=1}^{k_n} \mathrm{i}\alpha_{nk} t \\
&= \sum_{k=1}^{k_n} (\underline{\varphi}_{nk}(t) - 1) + \sum_{k=1}^{k_n} \mathrm{i}\alpha_{nk} t + \mathcal{E}_n(t) \\
&= \mathrm{i}t \sum_{k=1}^{k_n} \Big(\alpha_{nk} + \int_{\mathbb{R}} \frac{x}{x^2+1} \mathrm{d}\underline{F}_{nk}(x) \Big) + \sum_{k=1}^{k_n} \int_{\mathbb{R}} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2} \Big) \mathrm{d}\underline{F}_{nk}(x) + \mathcal{E}_n(t)\,.
\end{aligned}$$

Here $\mathcal{E}_n(t) := \sum_k (\log \underline{\varphi}_{nk}(t) - (\underline{\varphi}_{nk}(t) - 1))$, and Lemma 2.16 implies that $\mathcal{E}_n(t) \to 0$. Set

$$a_n := \sum_{k=1}^{k_n} \Big(\alpha_{nk} + \int_{\mathbb{R}} \frac{x}{x^2+1} \mathrm{d}\underline{F}_{nk}(x) \Big) \quad \text{and} \quad G_n(x) := \sum_{k=1}^{k_n} \int_{-\infty}^{x} \frac{y^2}{1+y^2} \mathrm{d}\underline{F}_{nk}(y)\,. \tag{2.19}$$

As each F_{nk} is bounded, non-decreasing, and right-continuous, so is G_n. This proves (2.14), and concludes the proof of Theorem 2.15.

The next result is a natural consequence of Theorem 2.15.

Lemma 2.18. *Let $\{X_{nk}\}$ satisfy the infinitesimal condition, with α_{nk}, $\underline{F}_{nk}$, a_n, $G_n(x)$ defined as in (2.18) and (2.19). Let X be a random variable having infinite divisible distribution F, and characteristic function*

$$\varphi(t) = \exp\left(\mathrm{i}at + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x)\right).$$

Then $S_n := \sum_k X_{nk} \xrightarrow{d} X$ if and only if $G_n \xrightarrow{w} G$ and $a_n \to a$.

Proof. Let φ_n be the characteristic function of S_n. In the proof of Theorem 2.15, one core results we proved is that if $\{X_{nk}\}$ are infinitesimal, then

$$\xi_n(t) := \mathrm{i}a_n t + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G_n(x) = \log\varphi_n - \mathcal{E}_n(t) \tag{2.20}$$

for some $\mathcal{E}_n(t) \to 0$.

(i) Suppose $S_n \xrightarrow{d} X$. By (2.20), we see that

$$\lim_n \exp(\xi_n(t)) = \lim_n \varphi_n = \varphi(t).$$

As $\xi_n(0) = \log\varphi(0) = 0$, by the continuity argument used in the proof of Lemma 2.10 (ii), we have $\xi_n \to \log\varphi$. By Lemma 2.10 (ii), we conclude that $G_n \xrightarrow{w} G$ and $a_n \to a$.

(ii) Suppose that $G_n \xrightarrow{w} G$ and $a_n \to a$. Using (2.20) we get

$$\lim_n \log\varphi_n(t) = \lim_n \left(\mathrm{i}a_n t + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G_n(x)\right) = \log\varphi(t),$$

which implies $\varphi_{S_n}(t) = \varphi_n(t) \to \varphi(t)$. This yields $S_n \xrightarrow{d} X$ as desired. $\square$

Remark 2.19. In the above arguments, as it was always assumed that $S_n := \sum_{k=1}^{k_n} X_{nk}$ converges, we (intentionally) ignore the "mean" of $\sum_k X_{nk}$. Consider the case when

$$k_n = n, \quad X_{nk} \overset{d}{=} \mathcal{N}(n^{-1/2}, \sigma^2 n^{-1}).$$

Then $\sum_k X_{nk} \overset{d}{=} \mathcal{N}(\sqrt{n}, \sigma^2)$, which does not converges. However, setting $b_n = n^{1/2}$ yields

$$S_n := \sum_k X_{nk} - b_n \overset{d}{=} \mathcal{N}(0, \sigma^2).$$

In general, let $\{X_{nk}\}$ satisfy the infinitesimal condition, and b_n be a sequence of numbers. Assume

$$S_n := \sum_{k=1}^{k_n} X_{nk} - b_n \tag{2.21}$$

converges. By going through the proofs of Theorem 2.15 and Lemma 2.18, we can get the following extensions.

(a) The limiting distribution class of S_n coincides with the set of infinitely divisible distributions.

(b) Let a_n, $G_n(x)$ defined as in (2.18) and (2.19). Let X be a random variable having infinite divisible distribution F, and characteristic function

$$\varphi(t) = \exp\left(\mathrm{i}at + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x)\right).$$

Then $S_n := \sum_k X_{nk} - b_n \xrightarrow{d} X$ if and only if $G_n \xrightarrow{w} G$ and $a_n - b_n \to a$.

Exercise 2.20. Prove statements (a) and (b) in Remark 2.19.

Theorems 2.6 and 2.15 imply that the characteristic function of the limiting distribution of $\sum_k X_{nk}$ is always of the form

$$\varphi(t) = \exp\left(\mathrm{i}at + \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x)\right).$$

Note that G is not necessarily continuous at 0. Define $\sigma^2 := G(0) - G(0-) \geqslant 0$. Since

$$\lim_{x\to 0}\left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2} = -t^2/2\,,$$

we can rewrite

$$\varphi(t) = \exp\left(\mathrm{i}at - \frac{\sigma^2 t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x)\right).$$

Sometimes, it is more convenient to use

$$L(x) := \begin{cases} \int_{-\infty}^{x} \frac{1+y^2}{y^2}\mathrm{d}G(y)\,, & x < 0\,, \\ -\int_{x}^{\infty} \frac{1+y^2}{y^2}\mathrm{d}G(y)\,, & x > 0\,. \end{cases}$$

Thus L is non-decreasing in $(-\infty, 0)$ and $(0, \infty)$, with $L(\infty) = L(-\infty) = 0$, and

$$\varphi(t) = \exp\left(\mathrm{i}at - \frac{\sigma^2 t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\right)\mathrm{d}L(x)\right). \tag{2.22}$$

By Lemma 2.9, there is a one-to-one correspondence between ξ and (a, σ^2, L). If $L(x)$ is constant in $(-\infty, 0)$ and $(0, \infty)$, the limiting distribution of $\sum_k X_{nk}$ is $\mathcal{N}(a, \sigma^2)$.

2.3. The triangular array case

Let us go back to what we are familiar of, i.e. the object

$$S_n := \frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n\,, \tag{2.23}$$

where $\{X_k\}_{k\geqslant 1}$ are independent, and $M_n > 0$. This can be thought as a special case of (2.21), by setting $k_n = n$, and $X_{nk} = X_k/M_n$. In this case, the infinitesimal condition reads

$$\lim_n \max_{1\leqslant k\leqslant n} \mathbb{P}\{|X_k| \geqslant \varepsilon M_n\} = 0\,. \tag{2.24}$$

for all $\varepsilon > 0$. In the sequel, when considering (2.23), we shall always work under the assumption (2.24).

Lemma 2.21. *If a random variable X satisfy $|\varphi_X(t)| = 1$ for $t \in (-\delta, \delta)$, where $\delta > 0$, then X is degenerate.*

Proof. $|\varphi_X(t)| = 1$ means $\varphi_X(t) = \exp(\mathrm{i}f(t))$ for some real function f. Then

$$\mathbb{E}[|\exp(\mathrm{i}tX) - \exp(\mathrm{i}f(t))|^2] = 2 - \mathbb{E}\exp(\mathrm{i}tX)\cdot\exp(-\mathrm{i}f(t)) - \mathbb{E}\exp(-\mathrm{i}tX)\cdot\exp(\mathrm{i}f(t)) = 0\,,$$

which implies

$$tX = f(t) \pmod{2\pi}$$

almost surely. Be careful that although $f(t)$ is deterministic, mod 2π might not be. We can find a rational number p and an irrational number s in $(0, \delta)$ such that $pX = f(p) + 2\pi k(p, X)$ and $sX = f(s) + 2\pi k(s, X)$. Note that

$$\left\{\frac{f(s) + 2\pi\mathbb{Z}}{s}\right\} \cap \left\{\frac{f(p) + 2\pi\mathbb{Z}}{p}\right\}$$

contains at most one point. Thus X must be degenerate. □

Lemma 2.22. *If (2.23) converges in distribution to a non-degenerate random variable X, then $M_n \to \infty$ and $M_{n+1}/M_n \to 1$. In addition, $b_n/n \to 0$.*

Proof. Suppose $M_n \not\to \infty$, then there exists a subsequence of M_n that converges, i.e. $\lim_{n'} M_{n'} = M \in \mathbb{R}$. For $t \in \mathbb{R}$, set $t_{n'} = M_{n'}t$. Since $\{X_k/M_n\}$ is infinitesimal, by Lemma 2.13, we have

$$\varphi_{X_k}(t) = \varphi_{X_k/M_{n'}}(M_{n'}t) \to 1\,.$$

As the LHS of the above is independent of n, we get $\varphi_{X_k} \equiv 1$ for all k. Thus

$$\varphi_{S_n}(t) = \mathrm{e}^{-\mathrm{i}tb_n}\prod_{k=1}^{n}\varphi_{X_k/M_n}(t) = \mathrm{e}^{-\mathrm{i}tb_n}\prod_{k=1}^{n}\varphi_{X_k}(t/M_n) = \mathrm{e}^{-\mathrm{i}tb_n}\,,$$

which contradicts to the fact that $\lim_n S_n$ is non-degenerate.

From (2.24), we see that $X_{n+1}/M_{n+1} \to 0$ in probability, thus

$$\widetilde{S}_n := \frac{1}{M_{n+1}}\sum_{k=1}^{n} X_k - b_{n+1} = S_{n+1} - X_{n+1}/M_{n+1} \xrightarrow{d} X\,.$$

Note that

$$\widetilde{S}_n = \frac{M_n}{M_{n+1}}S_n + \frac{b_nM_n}{M_{n+1}} - b_{n+1}\,.$$

Set $u_n := M_n/M_{n+1}$ and $v_n := b_nM_n/M_{n+1} - b_{n+1}$, then

$$S_n = \widetilde{S}_n/u_n - v_n/u_n\,.$$

There exists subsequences $u_{n'}, v_{n'}$ such that $\lim_{n'} u_{n'} = u \in \mathbb{R}_{\geqslant 0} \cup \{\infty\}$ and $\lim_{n'} v_{n'} = v \in \mathbb{R} \cup \{\pm\infty\}$. Suppose $u = \infty$, then

$$|\varphi_X(t)| = \left|\lim_{n'}\varphi_{S_{n'}}(t)\right| = \left|\lim_{n'}\varphi_{\widetilde{S}_{n'}/u_{n'} - v_{n'}/u_{n'}}(t)\right| = \left|\lim_{n'}\varphi_{\widetilde{S}_{n'}}(t/u_{n'})\exp(-\mathrm{i}tv_{n'}/u_{n'})\right| = |\varphi_X(0)| = 1$$

for all t. This contradicts to the fact that X is non-degenerate. Similarly, we can also show that $u \neq 0$. In addition, for t in a small neighborhood of 0,

$$\exp(-\mathrm{i}tv_{n'}/u_{n'}) = \frac{\varphi_{\widetilde{S}_{n'}}(t/u_{n'})\exp(-\mathrm{i}tv_{n'}/u_{n'})}{\varphi_{\widetilde{S}_{n'}}(t/u_{n'})} \to \frac{\varphi_X(t)}{\varphi_X(t/u)},$$

which implies $v = \lim_{n'} v_{n'} \in \mathbb{R}$. Then the above shows

$$1 = |\exp(-\mathrm{i}tv/u)| = \left|\frac{\varphi_X(t)}{\varphi_X(t/u)}\right|. \tag{2.25}$$

Suppose $u \neq 1$, then we can iterate (2.25) and show that $|\varphi_X| \equiv 1$ in a small neighborhood of 0, which contradicts to the fact that X is non-degenerate. Thus we must have $u = 1$. Note that the above argument shows that each converging subsequence of u_n must converges to 1. Thus $\lim_n u_n = 1$.

Note that the above argument also shows $v_n = b_n M_n/M_{n+1} - b_{n+1} \to 0$. Together with $M_n \to \infty$ and $M_{n+1}/M_n \to 1$, we can show that $b_n/n \to 0$ (exercise). This finishes the proof. □

2.3.1. The i.i.d. case. Now let us further assume in (2.23) that $\{X_n\}$ have identical distributions, and consider

$$S_n = \frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n. \tag{2.26}$$

We denote the class of limiting distributions of (2.26) by $\mathcal{S}$. We start with the following definition.

Definition 2.23. Let $X_1, X_2, \ldots$ be independent copies of a non-degenerate random variable X. We say that X (equivalently its distribution F) is stable, if for any $\alpha_1, \alpha_2 > 0$, there exists $\alpha > 0$, $\beta \in \mathbb{R}$ such that

$$\alpha_1 X_1 + \alpha_2 X_2 \stackrel{d}{=} \alpha X + \beta.$$

In other words, a distribution is said to be stable if a linear combination of two independent random variables with this distribution has the same distribution, up to location and scale parameters. It is easy to check that X is stable if and only if if for any $\alpha_1, \alpha_2 > 0$, there exists $\alpha > 0$, $\beta \in \mathbb{R}$ such that

$$\varphi_X(\alpha_1 t) \cdot \varphi_X(\alpha_2 t) = \mathrm{e}^{\mathrm{i}t\beta}\varphi_X(\alpha t).$$

If X is degenerate, then obviously X belongs to $\mathcal{S}$. If X is stable and non-degenerate, then let $X_1, X_2, \ldots$ be independent copies of X. By induction we have, for any $n \geqslant 1$,

$$X_1 + \cdots + X_n \stackrel{d}{=} M_n X + B_n$$

for some $M_n > 0, B_n \in \mathbb{R}$. Thus by defining $b_n := B_n/M_n$, we have

$$S_n = \frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n \stackrel{d}{=} X.$$

In particular, $S_n \to X$ in distribution. To show that X belongs to $\mathcal{S}$, we still need to prove the infinitesimal condition (2.24).

Note that we have

$$|\varphi_X(t)|^n = |\varphi_X(M_n t)|$$

for all $t \in \mathbb{R}$. Suppose $M_n \not\to \infty$, then for some subsequence M_{n_k}, we can find $M > 0$ such that $\max_k M_{n_k} < M$. As φ_X is continuous, let $[-\delta, \delta]$ be a small neighborhoods of 0 such that

$$\min_{t \in [-\delta,\delta]|} |\varphi_X(t)| > 0 \,. \tag{2.27}$$

Let $t_* \in [-\delta/M, \delta/M]$ such that $|\varphi_X(t_*)| < 1$. This is always possible as X is non-degenerate. Then

$$\lim_{n_k} |\varphi_X(M_{n_k} t_*)| = \lim_n |\varphi_X(t_*)|^{n_k} = 0 \,,$$

which contradicts (2.27). Thus we must have $M_n \to \infty$, which leads to $\lim_n \max_{1 \leqslant k \leqslant n} \mathbb{P}\{|X_k| \geqslant \varepsilon M_n\} = 0$.

To sum up, stable distributions always belongs to $\mathcal{S}$. By Theorem 2.15, stable distributions are infinitely divisible.

The above argument proves the "if" part of the next result.

Proposition 2.24. *A distribution F is in the class $\mathcal{S}$ if and only if it is stable.*

Proof of the "only if" part. Let $F \in \mathcal{S}$ be the limit of (2.26). As degenerate distributions are stable, we W.L.O.G. assume that F is non-degenerate. Let X be a random variable with distribution F, and let $X_1, X_2, \ldots$ be independent copies of X. Fix $\alpha_1 \geqslant \alpha_2 > 0$. Let

$$m_n = \max\{j : j \leqslant n, M_j \leqslant \alpha_2 M_n / \alpha_1\} \,.$$

As Lemma 2.22 implies $M_n \to \infty$ and $M_{n+1}/M_n \to 1$, we can show that

$$m_n \to \infty \quad \text{and} \quad M_{m_n}/M_n \to \alpha_2/\alpha_1 \,.$$

Let $\widetilde{M}_n := M_n/\alpha_1$, $\widetilde{b}_n := (M_n b_n + M_{m_n} b_{m_n})/\widetilde{M}_n$. Consider

$$\frac{M_n}{\widetilde{M}_n}\Big(\frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n\Big) + \frac{M_{m_n}}{\widetilde{M}_n}\Big(\frac{1}{M_{m_n}}\sum_{k=n+1}^{n+m_n} X_k - b_{m_n}\Big) = \frac{1}{\widetilde{M}_n}\sum_{k=1}^{n+m_n} X_k - \widetilde{b}_n \,. \tag{2.28}$$

It is not hard (exercise) to show that

$$\frac{M_n}{\widetilde{M}_n}\Big(\frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n\Big) \xrightarrow{d} \alpha_1 X \,, \quad \frac{M_{m_n}}{\widetilde{M}_n}\Big(\frac{1}{M_{m_n}}\sum_{k=n+1}^{n+m_n} X_k - b_{m_n}\Big) \to \alpha_2 X \,,$$

and

$$\begin{aligned}
\frac{1}{\widetilde{M}_n}\sum_{k=1}^{n+m_n} X_k - \widetilde{b}_n &= \frac{M_{n+m_n}}{\widetilde{M}_n} \cdot \Big(\frac{1}{M_{n+m_n}}\sum_{k=1}^{n+m_n} X_k - b_{n+m_n}\Big) + \frac{M_{n+m_n}}{\widetilde{M}_n} \cdot b_{n+m_n} - \widetilde{b}_n \\
&= \frac{M_{n+m_n}}{\widetilde{M}_n} \cdot S_{n+m_n} + \frac{M_{n+m_n}}{\widetilde{M}_n} \cdot b_{n+m_n} - \widetilde{b}_n \,.
\end{aligned} \tag{2.29}$$

So now we know that the RHS of (2.29) converges in law, and S_{n+m_n} converges to the non-degenerate random variable X. Thus the coefficients on RHS of (2.28) must also converge (you can prove this by the "subsequent argument" as in the proof of Lemma 2.22). Set

$$\alpha := \lim_n \frac{M_{n+m_n}}{\widetilde{M}_n} \quad \text{and} \quad \beta := \lim_n \frac{M_{n+m_n}}{\widetilde{M}_n} \cdot b_{n+m_n} - \widetilde{b}_n \,.$$

Taking limit in (2.28) yields $\alpha_1 X_1 + \alpha_2 X_2 \overset{d}{=} \alpha X + \beta$ as desired. This finishes the proof. □

Recall the representation (2.22) of the IDCF. The final goal of Section 2 is the following result.

Theorem 2.25. *An infinitely divisible random variable X (with characteristic function φ) is in $\mathcal{S}$ if and only if its L and σ^2 satisfy (i) $L(x) \equiv 0$; or (ii) $\sigma^2 = 0$ and*

$$L(x) = \begin{cases} c_1/|x^\alpha|, & x < 0, \\ -c_2/x^\alpha, & x > 0 \end{cases}$$

for some $\alpha \in (0, 2)$, $c_1, c_2 \geqslant 0$. In case (ii), if we in addition have $c_1 + c_2 > 0$, we say that L is α-stable.

Proof. Let us prove the "only if" part. The proof of the other direction is a simple verification. Let $X \in \mathcal{S}$ with the characteristic function e^ξ, where $\xi = (a, \sigma^2, L)$ according to Lemma 2.9 and (2.22). Let $X_1, X_2, ..., X_n$ be independent copies of X, and $Y_n := X_1 + \cdots + X_n$. Then

$$\log \varphi_{Y_n}(t) = \log \varphi_{X_1+\cdots+X_n}(t) = n\xi = \mathrm{i}ant - \frac{\sigma^2 n t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\Big) n \mathrm{d}L(x)\,.$$

Since $X \in S$, by Proposition 2.24 we know that $Y_n \stackrel{d}{=} \alpha_n X + \beta_n$, and thus

$$\begin{aligned}\log \varphi_{Y_n}(t) &= \mathrm{i}(a\alpha_n + \beta_n)t - \frac{\sigma^2\alpha_n^2 t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \Big(\mathrm{e}^{\mathrm{i}t\alpha_n x} - 1 - \frac{\mathrm{i}t\alpha_n x}{x^2+1}\Big)\mathrm{d}L(x) \\ &= \mathrm{i}(a\alpha_n + \beta_n + \gamma_n)t - \frac{\sigma^2\alpha_n^2 t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{x^2+1}\Big)\mathrm{d}L(x/\alpha_n)\,,\end{aligned}$$

where

$$\gamma_n := \int_{\mathbb{R}\setminus\{0\}} \frac{x}{x^2+1}\mathrm{d}L(x/\alpha_n) - \int_{\mathbb{R}\setminus\{0\}} \frac{\alpha_n x}{x^2+1}\mathrm{d}L(x)\,.$$

As the representation (2.23) for the characteristic function of Y_n is unique, we must have

$$\begin{cases} an = a\alpha_n + \beta_n + \gamma_n \\ \sigma^2 n = \sigma^2 \alpha_n^2 \\ n\mathrm{d}L(x) = \mathrm{d}L(x/\alpha_n)\,. \end{cases}$$

Case 1. If $\sigma^2 \neq 0$, then $\alpha_n = n^{1/2}$. By $n\mathrm{d}L(x) = \mathrm{d}L(x/\alpha_n)$ we have

$$L(x) = \begin{cases} c_1/x^2, & x < 0, \\ -c_2/x^2, & x > 0 \end{cases}$$

for some $c_1, c_2 \geqslant 0$. As

$$L(x) := \begin{cases} \int_{-\infty}^{x} \frac{1+y^2}{y^2}\mathrm{d}G(y)\,, & x < 0\,, \\ \\ -\int_{x}^{\infty} \frac{1+y^2}{y^2}\mathrm{d}G(y)\,, & x > 0\,. \end{cases} \tag{2.30}$$

and G is bounded, we must have $\lim_{x\to 0} x^2 L(x) = 0$ (since G is monotone and bounded, this can be proved via e.g. integration by parts). Thus $c_1 = c_2 = 0$, which makes $L(x) \equiv 0$.

Case 2. If $\sigma^2 = 0$, then by $n\mathrm{d}L(x) = \mathrm{d}L(x/\alpha_n)$ we have

$$L(x) = \begin{cases} c_1|x|^{-\log n/\log \alpha_n}, & x < 0, \\ -c_2 x^{-\log n/\log\alpha_n}, & x > 0 \end{cases}$$

By (2.30), and the fact that L is non-decreasing and independent of n, we must have $\log n/\log\alpha_n = \alpha \in (0, 2)$. This finishes the proof. □

2.3.2. Domain of contraction. The remaining question is, for which distributions of X_n does the i.i.d. sum $X_1 + ... + X_n$ converge?

To answer this question, we introduce the terminology of *slowly varying* function, which is a positive, measurable function $V : (0, \infty) \to (0, \infty)$ such that for any fixed $a > 0$,

$$\lim_{x\to\infty} \frac{V(ax)}{V(x)} = 1.$$

In other words, $V(x)$ changes so slowly at infinity that multiplying its argument by a constant factor a doesn't change its asymptotic behavior (up to lower-order terms). Examples of the slow varying function are any function that grows (or decays) slower than any power law X^ε, no matter how small $\varepsilon > 0$. For example, $\log x$, $\log\log x$, $(\log x)^p$ for any real p, or even a constant function.

We state the next result without giving the proof.

Theorem 2.26. *Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables with distribution function $F(x)$.*

(i) There exists $M_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n$$

converges to the limiting distribution $(a, \sigma^2, 0)$ with $\sigma^2 > 0$, if and only if

$$\lim_{z\to\infty} z^2 \int_{|x|\geqslant z} \mathrm{d}F(x) \Big/ \int_{|x|<z} x^2 \mathrm{d}F(x) = 0 \tag{2.31}$$

(ii) There exists $M_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\frac{1}{M_n}\sum_{k=1}^{n} X_k - b_n$$

converges to limiting distribution $(a, 0, L)$, where L is α-stable with $\alpha \in (0, 2)$, if and only if

$$F(x) = (c_1 + o(1))|x|^{-\alpha} V(|x|)$$

as $x \to -\infty$, and

$$1 - F(x) = (c_2 + o(1))x^{-\alpha} V(x)$$

as $x \to \infty$. Here $c_1, c_2 \geqslant 0$, and V is a slow-varying function.

Remark 2.27. The condition (2.31) can be met in two scenarios.

- $\mathbb{E}X_1^2 < \infty$. In this case we recover the classic central limit theorem, i.e.

$$\frac{X_1 + \cdots + X_n - n\mathbb{E}X_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(X_1))\,.$$

- $\mathbb{E}X_1^2 = \infty$, and $\mathbb{P}(|X_1| \geqslant x) = (1 + o(1))|x|^{-2} V(x)$ for some slow varying function V. Then we have

$$\frac{X_1 + \cdots + X_n - n\mathbb{E}X_1}{f(n)} \xrightarrow{d} \mathcal{N}(0, 1)\,.$$

Here $f(n)$ satisfies the implicit form

$$f(n) \sim \sqrt{2n \int_1^{f(n)} \frac{V(x)}{x} \mathrm{d}x}\,.$$

For instance, if $V \equiv C$, then $f(n) = \sqrt{Cn \log n}$; if $V(x) = \log x$, then $f(n) = \sqrt{n} \log n$; if $V(x) = 1/\log x$, then $f(n) = \sqrt{2n \log \log n}$.

Homework 1

Question 1. Are the following distribution infinitely divisible?

(i) $\mathbb{P}(X = k) = a^k(1+a)^{-k-1}, a > 0, k = 0, 1, 2, ...;$

(ii) $p_0 = \mathbb{P}(X = 0) = (1+\alpha\lambda)^{-1/\alpha}$, $\alpha, \lambda > 0$, and

$$\mathbb{P}(X = k) = \Big(\frac{\lambda}{1+\alpha\lambda}\Big)^k \frac{(1+\alpha)\cdots(1+(k-1)\alpha)}{k!} p_0$$

for $k = 1, 2, ...;$

(iii) The density function $f(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$.

Question 2. Let φ be a characteristic function. Suppose there is an integer sequence n_k satisfying $\lim_{k\to\infty} n_k = \infty$ such that for every $k \geqslant 1$, there exists a characteristic function φ_k satisfying $\varphi = (\varphi_k)^{n_k}$. Show that φ is infinitely divisible.

Question 3. Compute the Lévy-Khintchine representation for the follow characteristic functions.

(i) $\varphi(t) = (1 - \mathrm{i}t/\beta)^{-\alpha}, \alpha, \beta > 0;$

(ii) $\varphi(t) = e^{-\theta|t|}$, $\theta > 0$;

(iii) $\varphi(t) = (1+t^2)^{-1}$.

Question 4. Let X and Y be independent, infinitely divisible random variables. If $X+Y$ has Normal distribution, show that X and Y also have Normal distributions.

Question 5. Let $\{X_{nk}\}_{n\geqslant 1, 1\leqslant k\leqslant n}$ be independent random variables having distributions

$$\mathbb{P}(X_{nk} = k/n) = \mathbb{P}(X_{nk} = -k/n) = 1/n \quad \text{and} \quad \mathbb{P}(X_{nk} = 0) = 1 - 2/n\,.$$

Let $S_n := \sum_{k=1}^n X_{nk}$. Show that

(i) For any $n \geqslant 1$, S_n is not infinitely divisible.

(ii) There exists a non-Normal, infinitely divisible random variable S such that $S_n \xrightarrow{d} S$.

3 Around the central limit theorem

3.1. Three approaches to CLT

We shall give three different proofs for the following result.

Theorem 3.1 (Central limit theorem)**.** *Let $\{X_k\}_{k\geqslant 1}$ be a sequence of i.i.d. random variables, with $\mathbb{E}X_1 = \mu$ and $\operatorname{Var} X_1 = \sigma^2 > 0$. Let $S_n := X_1 + \cdots + X_n$. Then*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Note that by setting $X_i' = (X_i - \mu)/\sigma$ and $S_n' := X_1' + \cdots + X_n'$, we have $\mathbb{E}X_1' = 0$ and $\operatorname{Var} X_1' = 1$. In this case, Theorem 3.1 is equivalent to

$$\frac{S_n'}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Thus we shall W.L.O.G. assume that $\mu = 0$ and $\sigma^2 = 1$.

3.1.1. Proof 1: the classic method. As we already have the Levy-continuity theorem, Theorem 1.14, we can prove Theorem 3.1 rather simply.

Let $\varphi(t) := \mathbb{E}\exp(\mathrm{i}tX_1)$. As $\mathbb{E}X_1^2 = 1 < \infty$, the function φ is twice-differentiable at 0, and by Taylor expansion we have

$$\varphi(t) = 1 - \frac{t^2}{2} + o(t^2)$$

as $t \to 0$. Thus for any fixed $t \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}\exp(\mathrm{i}tS_n/\sqrt{n}) &= \prod_{i=1}^{n} \mathbb{E}\exp(\mathrm{i}tX_k/\sqrt{n}) = \varphi(t/\sqrt{n})^n = \Big(1 - \frac{t^2}{2n} + o(n^{-1})\Big)^n \\ &= \exp\Big(n\log\Big(1 - \frac{t^2}{2n} + o(n^{-1})\Big)\Big) = \exp\Big(n\Big(-\frac{t^2}{2n} + o(n^{-1}) + O(n^{-2})\Big)\Big)\,,\end{aligned}$$

which converges to $\mathrm{e}^{-t^2/2}$ as $n \to \infty$. By Theorem 1.14, we conclude $S_n/\sqrt{n} \xrightarrow{d} \mathcal{N}(0,1)$ as desired.

Remark 3.2. In the above argument, it is important that the variables $\{X_k\}_{k\geqslant 1}$ have identical distributions, so that the error $o(n^{-1})$ in $(1 - \frac{t^2}{2n} + o(n^{-1}))^n$ is uniform.

The next two proofs do not reply on Theorem 1.14, and as we shall see, the methods are also of independent interest. In fact, these ideas are still being used in current (as of January 2026) research of probability theory.

3.1.2. Proof 2: Lindberg replacement method. This approach starts with the observation that, for two independent random variables with distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, their sum has the distribution $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Let $\{Y_n\}$ be a sequence of i.i.d. random variables with distribution $\mathcal{N}(0,1)$, then

$$T_n := Y_1 + \cdots + Y_n \stackrel{d}{=} \mathcal{N}(0,n)\,. \tag{3.1}$$

In other words, $T_n/\sqrt{n} \stackrel{d}{=} \mathcal{N}(0,1)$. The central limit theorem simply state that, a sum of i.i.d. random variables is close to the sum of i.i.d. Gaussian, as long as the second moment exists.

Let us perform one modification before starting the proof. For $c > 0$, let $\widetilde{X}_k(c) := X_k \mathbf{1}_{|X_k| \leqslant c}$ for all k. Denote

$$\widetilde{\mu}(c) = \mathbb{E}\widetilde{X}_1(c)\,, \quad \widetilde{\sigma}^2(c) = \operatorname{Var}\widetilde{X}_1(c)\,, \quad \text{and} \quad \widetilde{S}_n(c) := \widetilde{X}_1(c) + \cdots + \widetilde{X}_n(c)\,.$$

Lemma 3.3. *If for any fixed $c > 0$ (implicitly, here we choose c large enough so that $\widetilde{X}_k(c)$ is non-degenerate, and thus $\widetilde{\sigma}^2(c) > 0$), we have*

$$\frac{\widetilde{S}_n(c) - n\widetilde{\mu}(c)}{\sqrt{n}\widetilde{\sigma}(c)} \xrightarrow{d} \mathcal{N}(0,1)\,, \tag{3.2}$$

then Theorem 3.1 holds.

Proof. Consider

$$\mathcal{E}(c) := \frac{S_n}{\sqrt{n}} - \frac{\widetilde{S}_n(c) - n\widetilde{\mu}(c)}{\sqrt{n}} = \frac{S_n - (\widetilde{S}_n(c) - n\widetilde{\mu}(c))}{\sqrt{n}}\,.$$

We have

$$\begin{aligned} \mathbb{E}\mathcal{E}^2(c) &= \mathbb{E}\bigg(\frac{1}{\sqrt{n}}\sum_{k=1}^n (X_k \mathbf{1}_{|X_k| \geqslant c} - \mathbb{E}[X_k \mathbf{1}_{|X_k| \geqslant c}])\bigg)^2 = \frac{1}{n}\sum_{k=1}^n \big(\mathbb{E}[X_k^2 \mathbf{1}_{|X_k| \geqslant c}] - \mathbb{E}[X_k \mathbf{1}_{|X_k| \geqslant c}]^2\big) \\ &= \mathbb{E}[X_1^2 \mathbf{1}_{|X_1| \geqslant c}] - \mathbb{E}[X_1 \mathbf{1}_{|X_1| \geqslant c}]^2 =: h(c)\,. \end{aligned}$$

By dominated convergence theorem, $h(c) \geqslant 0$ satisfies $\lim_{c \to +\infty} h(c) = 0$. Fix $\varepsilon > 0$ and let $x \in \mathbb{R}$. By Markov inequality we have $\mathbb{P}(|\mathcal{E}(c)| \geqslant \varepsilon) \leqslant h(c)\varepsilon^{-2}$, which implies

$$\mathbb{P}\Big(\frac{\widetilde{S}_n(c) - n\widetilde{\mu}(c)}{\sqrt{n}} \leqslant x - \varepsilon\Big) - h(c)\varepsilon^{-2} \leqslant \mathbb{P}(S_n/\sqrt{n} \leqslant x) \leqslant \mathbb{P}\Big(\frac{\widetilde{S}_n(c) - n\widetilde{\mu}(c)}{\sqrt{n}} \leqslant x + \varepsilon\Big) + h(c)\varepsilon^{-2}\,.$$

Thus

$$\mathbb{P}\bigg(\frac{\widetilde{S}_n(c) - n\widetilde{\mu}(c)}{\sqrt{n}\widetilde{\sigma}(c)} \leqslant \frac{x - \varepsilon}{\widetilde{\sigma}(c)}\bigg) - h(c)\varepsilon^{-2} \leqslant \mathbb{P}(S_n/\sqrt{n} \leqslant x) \leqslant \mathbb{P}\bigg(\frac{\widetilde{S}_n(c) - n\widetilde{\mu}(c)}{\sqrt{n}} \leqslant \frac{x + \varepsilon}{\widetilde{\sigma}(c)}\bigg) + h(c)\varepsilon^{-2}\,.$$

Let Φ be the distribution function of $\mathcal{N}(0,1)$. By (3.2), we have

$$\Phi\Big(\frac{x-\varepsilon}{\widetilde{\sigma}(c)}\Big) - h(c)\varepsilon^{-2} \leqslant \liminf_n \mathbb{P}(S_n/\sqrt{n} \leqslant x) \leqslant \limsup_n \mathbb{P}(S_n/\sqrt{n} \leqslant x) \leqslant \Phi\Big(\frac{x+\varepsilon}{\widetilde{\sigma}(c)}\Big) + h(c)\varepsilon^{-2}$$

As $\lim_{c \to \infty} \widetilde{\sigma}(c) = 1$, we get

$$\Phi(x - \varepsilon) \leqslant \liminf_n \mathbb{P}(S_n/\sqrt{n} \leqslant x) \leqslant \limsup_n \mathbb{P}(S_n/\sqrt{n} \leqslant x) \leqslant \Phi(x + \varepsilon)\,.$$

Hence $\lim_n \mathbb{P}(S_n/\sqrt{n} \leqslant x) = \Phi(x)$ for all $x \in \mathbb{R}$, which finishes the proof. □

Lemma 3.3 implies that we can W.L.O.G. assume the random variables $\{X_k\}_{k \geqslant 1}$ in Theorem 3.1 are bounded. Let f be a smooth function in $\mathbb{R}$ with bounded derivatives. By Theorem 1.13 and (3.1), it suffices to show that

$$\lim_n \mathbb{E}f(S_n/\sqrt{n}) = \lim_n \mathbb{E}f(T_n/\sqrt{n})\,. \tag{3.3}$$

For $k \in \{1, 2, ..., n\}$, define

$$R_n^{(k)} := \frac{1}{\sqrt{n}}(Y_1 + \cdots + Y_{k-1} + X_{k+1} + \cdots + X_n)\,.$$

Then

$$\begin{aligned}
\mathbb{E}f(S_n/\sqrt{n}) &= \mathbb{E}f\Big(R_n^{(1)} + \frac{X_1}{\sqrt{n}}\Big) = \mathbb{E}f\Big(R_n^{(1)} + \frac{Y_1}{\sqrt{n}}\Big) + \mathbb{E}f\Big(R_n^{(1)} + \frac{X_1}{\sqrt{n}}\Big) - \mathbb{E}f\Big(R_n^{(1)} + \frac{Y_1}{\sqrt{n}}\Big) \\
&= \mathbb{E}f\Big(R_n^{(2)} + \frac{X_2}{\sqrt{n}}\Big) + \mathbb{E}f\Big(R_n^{(1)} + \frac{X_1}{\sqrt{n}}\Big) - \mathbb{E}f\Big(R_n^{(1)} + \frac{Y_1}{\sqrt{n}}\Big) \\
&= \mathbb{E}f\Big(R_n^{(n)} + \frac{X_n}{\sqrt{n}}\Big) + \sum_{k=1}^{n}\bigg(\mathbb{E}f\Big(R_n^{(k)} + \frac{X_k}{\sqrt{n}}\Big) - \mathbb{E}f\Big(R_n^{(k)} + \frac{Y_k}{\sqrt{n}}\Big)\bigg) \\
&= \mathbb{E}f(T_n/\sqrt{n}) + \sum_{k=1}^{n}\bigg(\mathbb{E}f\Big(R_n^{(k)} + \frac{X_k}{\sqrt{n}}\Big) - \mathbb{E}f\Big(R_n^{(k)} + \frac{Y_k}{\sqrt{n}}\Big)\bigg)\,. \qquad (3.4)
\end{aligned}$$

As f is smooth and bounded,

$$f\Big(R_n^{(k)} + \frac{X_k}{\sqrt{n}}\Big) = f(R_n^{(k)}) + \Big(\frac{X_k}{\sqrt{n}}\Big)f'(R_n^{(k)}) + \frac{1}{2}\Big(\frac{X_k}{\sqrt{n}}\Big)^2 f^{(2)}(R_n^{(k)}) + \frac{1}{6}\Big(\frac{X_k}{\sqrt{n}}\Big)^3 f^{(3)}(R_n^{(k)} + \xi_k)\,,$$

where ξ_k takes value between 0 and X_k. As X_k and $f^{(3)}$ are bounded, taking expectation in the above yields

$$\mathbb{E}f\Big(R_n^{(k)} + \frac{X_k}{\sqrt{n}}\Big) = \mathbb{E}f(R_n^{(k)}) + \mathbb{E}\Big(\frac{X_k}{\sqrt{n}}\Big)\mathbb{E}f'(R_n^{(k)}) + \frac{1}{2}\mathbb{E}\Big[\Big(\frac{X_k}{\sqrt{n}}\Big)^2\Big]\mathbb{E}f^{(2)}(R_n^{(k)}) + O(n^{-3/2})\,.$$

Similarly,

$$\mathbb{E}f\Big(R_n^{(k)} + \frac{Y_k}{\sqrt{n}}\Big) = \mathbb{E}f(R_n^{(k)}) + \mathbb{E}\Big(\frac{Y_k}{\sqrt{n}}\Big)\mathbb{E}f'(R_n^{(k)}) + \frac{1}{2}\mathbb{E}\Big[\Big(\frac{Y_k}{\sqrt{n}}\Big)^2\Big]\mathbb{E}f^{(2)}(R_n^{(k)}) + O(n^{-3/2})\,.$$

As X_k and Y_k have identical first two moments, we have

$$\mathbb{E}f\Big(R_n^{(k)} + \frac{X_k}{\sqrt{n}}\Big) - \mathbb{E}f\Big(R_n^{(k)} + \frac{Y_k}{\sqrt{n}}\Big) = O(n^{-3/2})$$

uniformly for all k. Inserting the above into (3.4) yields

$$\mathbb{E}f(S_n/\sqrt{n}) = \mathbb{E}f(T_n/\sqrt{n}) + O(n^{-1/2})\,.$$

This proves (3.3) and finishes the (2nd) proof of Theorem 3.1.

3.1.3. Proof 3: Stein's method.

Lemma 3.4 (Stein's lemma)**.** *A random variable X is standard normal distribution, if and only if*

$$\mathbb{E}Xf(X) = \mathbb{E}f'(X)$$

for any $f \in C^1(\mathbb{R})$ satisfying $\lim_{x\to\infty} f(x)\mathrm{e}^{-x^2/2} = \lim_{x\to-\infty} f(x)\mathrm{e}^{-x^2/2} = 0$.

Proof. Let $X \overset{d}{=} \mathcal{N}(0,1)$, and $f \in C^1(\mathbb{R})$ satisfying $\lim_{x\to\infty} f(x)\mathrm{e}^{-x^2/2} = \lim_{x\to-\infty} f(x)\mathrm{e}^{-x^2/2} = 0$. Then the integration by parts formula implies

$$\begin{aligned}\mathbb{E}Xf(X) &= \int_{\mathbb{R}} xf(x) \cdot \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\mathrm{d}x = -\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(x)(e^{-x^2/2})'\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f'(x)e^{-x^2/2}\mathrm{d}x = \mathbb{E}f'(X)\,.\end{aligned}$$

This proves the "only if" part.

To prove the "if" part, let $Y \overset{d}{=} \mathcal{N}(0,1)$, and let g be a smooth function on $\mathbb{R}$ with compact support. It suffices to show that $\mathbb{E}g(X) = \mathbb{E}g(Y)$. Consider the function f satisfying the ODE

$$f'(x) = xf(x) + g(x) - \mathbb{E}g(Y)\,. \tag{3.5}$$

It is easy to see that the solution is of the form

$$f(x) = \mathrm{e}^{x^2/2}\bigg(\int_{-\infty}^{x} \mathrm{e}^{-y^2/2}(g(y) - \mathbb{E}g(Y))\mathrm{d}y + C\bigg)\,.$$

Setting $C = 0$ makes f a smooth function satisfying $\|f\|_\infty + \|f'\|_\infty + \|f''\|_\infty < \infty$. Setting $x = X$ in (3.5) and taking expectation, we have

$$0 = \mathbb{E}f'(X) - \mathbb{E}Xf(X) = \mathbb{E}g(X) - \mathbb{E}g(Y)$$

as desired. □

Lemma 3.5. *Let*

$$f(x) = \mathrm{e}^{x^2/2}\int_{-\infty}^{x} \mathrm{e}^{-y^2/2}(g(y) - \mathbb{E}g(Y))\mathrm{d}y\,,$$

where g is smooth with compact support, and $Y \overset{d}{=} \mathcal{N}(0,1)$. Then $\|f\|_\infty + \|f'\|_\infty + \|f''\|_\infty < \infty$.

Proof. Let $\Phi(x) = \mathbb{P}(Y \leqslant x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\mathrm{d}x$. For large x, we can show that

$$\frac{x\mathrm{e}^{-x^2/2}}{(1+x^2)\sqrt{2\pi}} \leqslant 1 - \Phi(x) \leqslant \frac{\mathrm{e}^{-x^2/2}}{x\sqrt{2\pi}}\,,$$

which implies

$$\frac{x}{1+x^2} \leqslant e^{x^2/2}\sqrt{2\pi}(1-\Phi(x)) \leqslant \frac{1}{x}\,. \tag{3.6}$$

As g has compact support, for x large enough, we have

$$f(x) = e^{x^2/2}\int_{\mathbb{R}} \mathrm{e}^{-y^2/2}g(y)\mathrm{d}y - \mathrm{e}^{x^2/2}\mathbb{E}g(Y)\int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\mathrm{d}y = \mathrm{e}^{x^2/2}\sqrt{2\pi}(1-\Phi(x))\mathbb{E}g(Y)\,,$$

and $\lim_{x\to\infty}|f(x)| = 0$ follows from (3.6). Similarly, $\lim_{x\to-\infty}|f(x)| = 0$. This proves $\|f\|_\infty < \infty$.

For x large enough, we have

$$f'(x) = xf(x) + g(x) - \mathbb{E}g(Y) = x\mathrm{e}^{x^2/2}\sqrt{2\pi}(1-\Phi(x))\mathbb{E}g(Y) - \mathbb{E}g(Y)\,,$$

and $\lim_{x\to\infty}|f'(x)| = 0$ follows from (3.6). Similarly, $\lim_{x\to-\infty}|f'(x)| = 0$. This proves $\|f'\|_\infty < \infty$.

In addition, note that

$$
\begin{aligned}
f''(x) &= (xf(x) + g(x) - \mathbb{E}g(Y))' = f(x) + xf'(x) + g'(x) \\
&= (1+x^2)f(x) + xg(x) - x\mathbb{E}g(Y) = (1+x^2)f(x) - x\mathbb{E}g(Y)\,.
\end{aligned}
$$

As $\|f\|_\infty < \infty$, it suffices to show that $x^2 f(x) - x\mathbb{E}g(Y)$ is bounded. We have, for x large enough that

$$
\begin{aligned}
x^2 f(x) - x\mathbb{E}g(Y) &= x^2 \mathrm{e}^{x^2/2}\sqrt{2\pi}(1-\Phi(x))\mathbb{E}g(Y) - x\mathbb{E}g(Y) \\
&= \mathbb{E}g(Y)\cdot \big(x^2\mathrm{e}^{x^2/2}\sqrt{2\pi}(1-\Phi(x)) - x\big) \to 0\,,
\end{aligned}
$$

where the last step again follows from (3.6). We can also prove the analogue for $x \to -\infty$. This finishes the proof. □

The *Stein operator* $\mathcal{A}$ is defined as

$$
(\mathcal{A}f)(x) = f'(x) - xf(x)\,.
$$

From Lemma 3.4, a random variable is standard Gaussian if and only if $\mathbb{E}(\mathcal{A}f)(X) = 0$.

Lemma 3.6. *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables, with $\mathbb{E}X_n = 0$ and $\sup_n \mathbb{E}X_n^2 < \infty$. Then $X_n \xrightarrow{d} \mathcal{N}(0,1)$ if and only if*

$$
\mathbb{E}[X_n f(X_n)] - \mathbb{E}f'(X_n) \to 0
$$

for any $f \in C^1(\mathbb{R})$ with $\|f\|_\infty + \|f'\|_\infty < \infty$.

Proof. (i) Suppose $X_n \xrightarrow{d} \mathcal{N}(0,1)$ and we denote $Y \overset{d}{=} \mathcal{N}(0,1)$. Let $f \in C^1(\mathbb{R})$ with $\|f\|_\infty + \|f'\|_\infty < \infty$. Then (why?)

$$
X_n f(X_n) \xrightarrow{d} Yf(Y)\,, \quad f'(X_n) \xrightarrow{d} f'(Y)\,.
$$

As $\sup_n \mathbb{E}X_n^2 < \infty$, then $\{X_n\}_{n\geqslant 1}$ is uniformly integrable (recall that this means for any $\varepsilon > 0$, there exits K such that $\sup_n \mathbb{E}|X_n \mathbf{1}_{|X_n|\geqslant K}| \leqslant \varepsilon$). As f is bounded, $\{X_n f(X_n)\}_{n\geqslant 1}$ is also uniformly integrable. Then (why?)

$$
\lim_n \mathbb{E}[X_n f(X_n)] = \mathbb{E}[Yf(Y)]\,, \quad \text{and} \quad \lim_n \mathbb{E}f'(X_n) = \mathbb{E}f'(Y)\,.
$$

As $\mathbb{E}[Yf(Y)] = \mathbb{E}f'(Y)$, we conclude the "only if" part of the proof.

(ii) Suppose

$$
\mathbb{E}[X_n f(X_n)] - \mathbb{E}f'(X_n) \to 0
$$

for any $f \in C^1(\mathbb{R})$ with $\|f\|_\infty + \|f'\|_\infty < \infty$. Let $Y \overset{d}{=} \mathcal{N}(0,1)$, and let g be a bounded smooth function on $\mathbb{R}$. As in the proof of Lemma 3.4, and consider the function f satisfying the ODE $f'(x) = xf(x) + g(x) - \mathbb{E}g(Y)$. Then

$$
\mathbb{E}g(X_n) - \mathbb{E}g(Y) = \mathbb{E}f'(X_n) - \mathbb{E}[X_n f(X_n)] \to 0\,.
$$

By Theorem 1.13, we have $X_n \xrightarrow{d} X \overset{d}{=} \mathcal{N}(0,1)$. This finishes the "if" part of the proof. □

Exercise 3.7. (i) Work out the two "(why?)" in the proof of Lemma 3.6.

(ii) Given a counter example that Lemma 3.6 fails if we drop the condition $\|f\|_\infty < \infty$.

Now we prove Theorem 3.1 using Lemma 3.6. By Lemma 3.3, it suffices to assume that $\{X_n\}_{\geqslant 1}$ are uniformly bounded. Let $S_n^{(k)} := S_n - X_k$. Note that X_k and $S_n^{(k)}$ are independent. Let f be a smooth function in $\mathbb{R}$ satisfying $\|f\|_\infty + \|f'\|_\infty + \|f''\|_\infty < \infty$. We have

$$
\begin{aligned}
\mathbb{E}\Big(\frac{S_n}{\sqrt{n}} f\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) &= \sum_{k=1}^n \mathbb{E}\Big(\frac{X_k}{\sqrt{n}} f\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) = \sum_{k=1}^n \mathbb{E}\Big(\frac{X_k}{\sqrt{n}} f\Big(\frac{S_n}{\sqrt{n}}\Big) - \frac{X_k}{\sqrt{n}} f\Big(\frac{S_n^{(k)}}{\sqrt{n}}\Big)\Big) \\
&= \sum_{k=1}^n \mathbb{E}\Big(\frac{X_k^2}{n}\int_0^1 f'\Big(\frac{S_n^{(k)}}{\sqrt{n}} + \frac{X_k}{\sqrt{n}}t\Big)\mathrm{d}t\Big) \\
&= \sum_{k=1}^n \mathbb{E}\Big(\frac{X_k^2}{n}\int_0^1 f'\Big(\frac{S_n^{(k)}}{\sqrt{n}}\Big)\mathrm{d}t\Big) + \sum_{k=1}^n \mathbb{E}\Big(\frac{X_k^2}{n}\int_0^1 f'\Big(\frac{S_n^{(k)}}{\sqrt{n}} + \frac{X_k}{\sqrt{n}}t\Big) - f'\Big(\frac{S_n^{(k)}}{\sqrt{n}}\Big)\mathrm{d}t\Big) \\
&= \frac{1}{n}\sum_{k=1}^n \mathbb{E}\Big(f'\Big(\frac{S_n^{(k)}}{\sqrt{n}}\Big)\Big) + O(n^{-1/2}), \qquad (3.7)
\end{aligned}
$$

where in the last step we used

$$
\Big|\sum_{k=1}^n \mathbb{E}\Big(\frac{X_k^2}{n}\int_0^1 f'\Big(\frac{S_n^{(k)}}{\sqrt{n}} + \frac{X_k}{\sqrt{n}}t\Big) - f'\Big(\frac{S_n^{(k)}}{\sqrt{n}}\Big)\mathrm{d}t\Big)\Big| \leqslant \sum_{k=1}^n \mathbb{E}\Big(\frac{X_k^3}{n^{3/2}}\|f^{(2)}\|_\infty\Big) = O(n^{-1/2}).
$$

Similarly, we can also show that

$$
\frac{1}{n}\sum_{k=1}^n \mathbb{E}\Big(f'\Big(\frac{S_n^{(k)}}{\sqrt{n}}\Big)\Big) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}\Big(f'\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) + O(n^{-1/2}) = \mathbb{E}\Big(f'\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) + O(n^{-1/2}).
$$

Combining the above equation with (3.7) yields

$$
\mathbb{E}\Big(\frac{S_n}{\sqrt{n}} f\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) = \mathbb{E}\Big(f'\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) + O(n^{-1/2}).
$$

Note that $\sup_n \mathbb{E}(S_n/\sqrt{n})^2 = \mathbb{E}X_1^2 = 1$. Together with Lemma 3.6 we conclude the (3rd) proof of Theorem 3.1.

Remark 3.8. Although Proof 3 is slightly more complicated than Proof 2, Stein's method has its own advantages and is applicable in general. For instance, we can take $g(x) = \mathbf{1}_{(-\infty,w]}(x)$, and in this case $\mathbb{E}g(X) - \mathbb{E}g(Y) = \mathbb{P}(X \leqslant w) - \mathbb{P}(Y \leqslant w)$. Although g is not differentiable, we have

$$
\begin{aligned}
f(x) &= \mathrm{e}^{x^2/2}\int_{-\infty}^x (g(y) - \mathbb{E}g(Y))\mathrm{e}^{-y^2/2}\mathrm{d}y \\
&= \mathrm{e}^{x^2/2}\int_{-\infty}^x [\mathbf{1}_{(-\infty,w]}(y) - \Phi(w)]\mathrm{e}^{-y^2/2}\mathrm{d}y = -\mathrm{e}^{x^2/2}\int_x^{\infty} [\mathbf{1}_{(-\infty,w]}(y) - \Phi(w)]\mathrm{e}^{-y^2/2}\mathrm{d}y \\
&= \begin{cases} \sqrt{2\pi}\mathrm{e}^{x^2/2}\Phi(x)(1-\Phi(w)) & x \leqslant w, \\ \sqrt{2\pi}\mathrm{e}^{x^2/2}\Phi(w)(1-\Phi(x)) & x \geqslant w, \end{cases}
\end{aligned}
$$

which is a smooth function. This can be used to study the convergence rate of CLT, which we will address in the next section, with a different method.

3.1.4. Necessary condition and the non-identical case. The following result asserts that the existence of the second moment is necessary for getting the CLT.

Theorem 3.9. *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, and $S_n := X_1 + \cdots + X_n$. Suppose*

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)\,, \tag{3.8}$$

then $\mathbb{E}X_1^2 < \infty$. In addition, we have $\mathbb{E}X_1 = 0$ and $\operatorname{Var} X_1 = 1$.

Proof. The proof of $\mathbb{E}X_1^2 < \infty$ is technical and we omit it here. The readers may refer to [11, pages 62-63].

Assuming that $\mathbb{E}X_1^2 < \infty$ is true, by Jensen's inequality, we get $(\mathbb{E}X_1)^2 \leqslant \mathbb{E}X_1^2 < \infty$. This implies $\mu := \mathbb{E}X_1$ and $\sigma^2 := \operatorname{Var} X_1 = \mathbb{E}X_1^2 - (\mathbb{E}X_1)^2$ exist. Moreover, as the limit of S_n is non-degenerate, so is X_1. This implies $\sigma^2 > 0$. Using Theorem 3.1, we get

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Comparing the above with (3.8), we must have $\mu = 1$ and $\sigma^2 = 1$. This finishes the proof. $\square$

With certain effort, Theorem 3.1 can be generated to the case where the distributions of X_n are not identical. In fact, we can even assume X_n have different means and variances. Here we state a simple version of this type of result; for the general case, one may refer to [11, page 64].

Exercise 3.10. Let $\{X_n\}_{n\geqslant 1}$ be a sequence of independent random variables, where $\inf_n \operatorname{Var} X_n \geqslant c$ and $\sup_n \mathbb{E}|X_n^3| \leqslant C$ for some constants $C, c > 0$. Denote $S_n := X_1 + \cdots + X_n$ and $\sigma_n^2 = \operatorname{Var} X_n$. Use the Lindeberg replacement method to show that

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

3.2. The convergence rate

A natural problem to study after Theorem 3.1 is, *how fast* does the convergence happen, in terms of n. To make the question mathematical rigorous, for two random variables X, Y (or equivalently two random variables), we define their Kolmogorov-Smirnov distance by

$$\Delta(X,Y) = \sup_{x\in\mathbb{R}} |\mathbb{P}(X \leqslant x) - \mathbb{P}(Y \leqslant x)|\,. \tag{3.9}$$

We shall prove the following result[4].

Theorem 3.11 (Berry-Esseen)**.** *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, with $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = 1$. Suppose that $L := \mathbb{E}|X_1^3| < \infty$. Denote $S_n := X_1 + \cdots + X_n$ and $Y \stackrel{d}{=} \mathcal{N}(0,1)$. Then*

$$\Delta(S_n/\sqrt{n}, Y) \leqslant CL/\sqrt{n}$$

for some universal constant $C > 0$.

[4]The proof we present here is based on notes by Jordan Bell [3]. You may also read the recent exposition by Vershynin [13]

Remark 3.12. The Berry-Esseen theorem actually holds under more general assumptions, in terms of the distribution and variance profile of $\{X_n\}$. For simplicity, we do not pursue it here.

The starting point is the following result.

Lemma 3.13. *Let F and G be two distribution functions. Assume G is differentiable with $M := \|G'\|_\infty < \infty$. Abbreviate $\delta := \frac{1}{2M}\Delta(F,G)$. Then there exists $a \in \mathbb{R}$ such that for all $T > 0$, we have*

$$TM\delta\pi - 6TM\delta \int_{T\delta}^{\infty} \frac{1-\cos x}{x^2}\mathrm{d}x \leqslant \left| \int_{\mathbb{R}} \frac{1-\cos Tx}{x^2}(F(x+a) - G(x+a))\mathrm{d}x \right| .$$

Proof. As $\lim_{x\to\pm\infty} |F(x) - G(x)| = 0$, there exists a compact interval K such that

$$2M\delta = \sup_{x\in\mathbb{R}} |F(x) - G(x)| = \sup_{x\in K} |F(x) - G(x))| .$$

By definition, there exists a sequence $\{x_n\} \subset K$ such that $\lim_n |F(x_n) - G(x_n)| = 2M\delta$. Since K is bounded, there exists a subsequence $\{u_n\}$ of $\{x_n\}$ that converges. Since K is compact, $u := \lim_n u_n \in K$. Then either there is a subsequence $\{\widehat{u}_n\}$ of $\{u_n\}$ with $\lim_n \widehat{u}_n \downarrow u$, or there is a subsequence $\{\widetilde{u}_n\}$ of $\{u_n\}$ with $\lim_n \widetilde{u}_n \uparrow u$. In the first case we have

$$F(u) - G(u) = 2M\delta \quad \text{or} \quad F(u) - G(u) = -2M\delta ,$$

and in the second case we have

$$F(u-) - G(u) = 2M\delta \quad \text{or} \quad F(u-) - G(u) = -2M\delta .$$

We shall proceed the proof under the assumption

$$F(u-) - G(u) = -2M\delta ; \tag{3.10}$$

other cases work in a similar fashion. Let $a = u - \delta$. For $|x| < \delta$, $x + a < \delta + a = u$. Then

$$0 \leqslant G(u) - G(x+a) = \int_{x+a}^{u} G'(y)\mathrm{d}y = \int_{u+x-\delta}^{u} G'(y)\mathrm{d}y \leqslant M(\delta - x) ,$$

which implies $G(x+a) \geqslant G(u) - M(\delta - x)$. Since $x + a < u$, we have

$$F(x+a) - G(x+a) \leqslant F(u-) - G(u) + M(\delta - x) = -M(x+\delta) .$$

Thus

$$\int_{-\delta}^{\delta} \frac{1-\cos Tx}{x^2}(F(x+a)-G(x+a))\mathrm{d}x \leqslant -M \int_{-\delta}^{\delta} \frac{1-\cos Tx}{x^2}(x+\delta)\mathrm{d}x = -2M\delta \int_0^{\delta} \frac{1-\cos Tx}{x^2}\mathrm{d}x .$$

On the other hand, as $|F(x+a) - G(x+a)| \leqslant 2M\delta$, we have

$$\left| \int_{\mathbb{R}\setminus[-\delta,\delta]} \frac{1-\cos Tx}{x^2}(F(x+a)-G(x+a))\mathrm{d}x \right| \leqslant 2M\delta \int_{\mathbb{R}\setminus[-\delta,\delta]} \frac{1-\cos Tx}{x^2}\mathrm{d}x = 4M\delta \int_{\delta}^{\infty} \frac{1-\cos Tx}{x^2}\mathrm{d}x .$$

Hence

$$\begin{aligned} \int_{\mathbb{R}} \frac{1-\cos Tx}{x^2}(F(x+a) - G(x+a))\mathrm{d}x &\leqslant -2M\delta \int_0^{\delta} \frac{1-\cos Tx}{x^2}\mathrm{d}x + 4M\delta \int_{\delta}^{\infty} \frac{1-\cos Tx}{x^2}\mathrm{d}x \\ = -2M\delta \int_0^{\infty} \frac{1-\cos Tx}{x^2}\mathrm{d}x + 6M\delta \int_{\delta}^{\infty} \frac{1-\cos Tx}{x^2}\mathrm{d}x &= -TM\delta\pi + 6TM\delta \int_{T\delta}^{\infty} \frac{1-\cos x}{x^2}\mathrm{d}x . \end{aligned}$$

This finishes the proof for the case (3.10). ∎

Exercise 3.14. Prove Lemma 3.13 for the three cases other than (3.10).

Lemma 3.15. *We adopt the assumptions of Lemma 3.13 and let F, G, M, δ be as in Lemma 3.13. In addition, assume*

$$\int_{\mathbb{R}} |F(x) - G(x)| \mathrm{d}x < \infty \,.$$

Denote

$$\varphi_F(t) := \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}F(x) \quad \textit{and} \quad \varphi_G(t) := \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}G(x) \,.$$

Then for any $T > 0$,

$$\Delta(F, G) \leqslant \frac{2}{\pi} \int_0^T \frac{|\varphi_F(t) - \varphi_G(t)|}{t} \mathrm{d}t + \frac{24M}{\pi T} \,.$$

Proof. As $F - G$ vanishes at $\pm\infty$, an integration by parts formula yields

$$\varphi_F(t) - \varphi_G(t) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}(F - G)(x) = -\mathrm{i}t \int_{\mathbb{R}} (F - G)(x) \mathrm{e}^{\mathrm{i}tx} \mathrm{d}x \,.$$

Let a be as in Lemma 3.13, then

$$\frac{\varphi_F(t) - \varphi_G(t)}{-\mathrm{i}t} e^{-\mathrm{i}ta} (T - t) = (T - t) \int_{\mathbb{R}} (F(x + a) - G(x + a)) \mathrm{e}^{\mathrm{i}tx} \mathrm{d}x \,.$$

As $\int_{\mathbb{R}} |F(x) - G(x)| \mathrm{d}x < \infty$, by Fubini's theorem,

$$\begin{aligned}
&\int_0^T \frac{\varphi_F(t) - \varphi_G(t)}{-\mathrm{i}t} e^{-\mathrm{i}ta} (T - t) \mathrm{d}t = \int_0^T (T - t) \int_{\mathbb{R}} (F(x + a) - G(x + a)) \mathrm{e}^{\mathrm{i}tx} \mathrm{d}x \, \mathrm{d}t \\
&= \int_{\mathbb{R}} (F(x + a) - G(x + a)) \int_0^T (T - t) \mathrm{e}^{\mathrm{i}tx} \mathrm{d}t \, \mathrm{d}x = \int_{\mathbb{R}} (F(x + a) - G(x + a)) \frac{1 - \cos Tx}{x^2} \mathrm{d}x \,.
\end{aligned}$$

Thus

$$\left| \int_{\mathbb{R}} (F(x + a) - G(x + a)) \frac{1 - \cos Tx}{x^2} \mathrm{d}x \right| \leqslant \int_0^T \frac{|\varphi_F(t) - \varphi_G(t)|}{t} (T - t) \mathrm{d}t \leqslant T \int_0^T \frac{|\varphi_F(t) - \varphi_G(t)|}{t} \mathrm{d}t \,.$$

Together with Lemma 3.13 we get

$$TM\delta\pi - 6TM\delta \int_{T\delta}^{\infty} \frac{1 - \cos x}{x^2} \mathrm{d}x \leqslant T \int_0^T \frac{|\varphi_F(t) - \varphi_G(t)|}{t} \mathrm{d}t \,.$$

As

$$6TM\delta \int_{T\delta}^{\infty} \frac{1 - \cos x}{x^2} \mathrm{d}x \leqslant 6TM\delta \int_{T\delta}^{\infty} \frac{2}{x^2} \mathrm{d}x = 12M$$

we have

$$TM\delta\pi - 12M \leqslant T \int_0^T \frac{|\varphi_F(t) - \varphi_G(t)|}{t} \mathrm{d}t \,.$$

This implies the desired result. ∎

To prove Theorem 3.11, let F_n be the distribution of $S_n/\sqrt{n}$, and let Φ be the distribution of Y. By Markov inequality, for $x<0$ we have

$$F_n(x)=\mathbb{P}(S_n/\sqrt{n}\leqslant x)\leqslant \frac{1}{x^2}\mathbb{E}\Big(\frac{S_n^2}{n}\Big)=\frac{1}{x^2}\,,\quad \Phi(x)=\mathbb{P}(Y\leqslant x)\leqslant\frac{1}{x^2}\mathbb{E}(Y^2)=\frac{1}{x^2}\,.$$

Similarly, for $x>0$

$$1-F_n(x)=\mathbb{P}(S_n/\sqrt{n}>x)\leqslant\frac{1}{x^2}\,,\quad 1-\Phi(x)\leqslant\frac{1}{x^2}\,.$$

Thus $|F_n(x)-\Phi(x)|\leqslant 1/x^2$ for all $x\in\mathbb{R}$ (why there is no factor 2 in the bound?). Hence

$$\int_{\mathbb{R}}|F_n(x)-\Phi(x)|\mathrm{d}x=\int_{-1}^{1}2\mathrm{d}x+\int_{|x|\geqslant 1}\frac{1}{x^2}\mathrm{d}x<\infty\,.$$

As a result, we are allowed to use Lemma 3.15 with $F=F_n$ and $G=\Phi$. Note that $M=\|\Phi'\|_\infty=\frac{1}{2\pi}$. Thus Lemma 3.15 implies

$$\begin{aligned}\Delta(S_n/\sqrt{n},Y)&\leqslant\frac{2}{\pi}\int_0^T\frac{|\varphi_{F_n}(t)-\varphi_\Phi(t)|}{t}\mathrm{d}t+\frac{24M}{\pi T}\\&=\frac{2}{\pi}\int_0^T\frac{|\varphi_{F_n}(t)-\mathrm{e}^{-t^2/2}|}{t}\mathrm{d}t+\frac{24}{\sqrt{2\pi}\pi T}\,.\end{aligned}\tag{3.11}$$

Abbreviate $\varphi_n=\varphi_{F_n}$. Our main task is estimating $|\varphi_n(t)-\mathrm{e}^{-t^2/2}|$ for $t\in[0,T]$. To get the rate $n^{-1/2}$ as in Theorem 3.11, we would have to set $T\sim\sqrt{n}$.

The following is an easy consequence from Taylor expansion.

Lemma 3.16. *For $n\geqslant 1$ and $|z|<1$, we have*

$$\Bigg|\log(1+z)-\sum_{i=1}^{n-1}\frac{(-1)^i z^i}{i}\Bigg|\leqslant\frac{|z|^n}{n(1-|z|)}\,.$$

In particular, when $|z|\leqslant 1/2$, we have

$$|\log(1+z)-z|\leqslant|z|^2\,.$$

Recall that $L:=\mathbb{E}|X_1^3|<\infty$. The next two results estimate the characteristic function on different scales of t.

Lemma 3.17. *If $|t|<\frac{n^{1/6}}{2L^{1/3}}$, we have*

$$|\varphi_n(t)-\mathrm{e}^{-t^2/2}|\leqslant Ln^{-1/2}|t|^3\mathrm{e}^{-t^2/2}\,.$$

Proof. Let $\varphi(t):=\mathbb{E}\exp(\mathrm{i}tX_1/\sqrt{n})$. Since $\mathbb{E}X_1=0$ and $\mathbb{E}X_1^2=1$, by Taylor expansion, we have

$$\varphi(t)=\varphi(0)+\varphi'(0)t+\frac{\varphi^{(2)}(0)}{2!}t^2+\frac{\varphi^{(3)}(s)}{3!}t^3=1-\frac{t^2}{2n}-\frac{\mathrm{i}\mathbb{E}(X_1^3\varphi(s))}{6n^{3/2}}t^3\tag{3.12}$$

for some s between 0 and t. By Jensen's inequality, $L\geqslant 1$, and thus

$$|\varphi(t)-1|\leqslant\frac{t^2}{2n}+\frac{L}{6n^{3/2}}|t^3|\leqslant\frac{1}{2n}\cdot\frac{n^{1/3}}{4L^{2/3}}+\frac{L}{6n^{3/2}}\cdot\frac{n^{1/2}}{8L}\leqslant\frac{1}{4n^{2/3}}\leqslant 1/4\,.$$

By Lemma 3.16, we have

$$|\log\varphi(t)-(\varphi(t)-1)|\leqslant|\varphi(t)-1|^2\leqslant\frac{t^4}{2n^2}+\frac{L^2t^6}{18n^3}\leqslant\frac{|t|^3}{4n^{11/6}}+\frac{L|t|^3}{124n^{3/2}}\leqslant\frac{L|t|^3}{2n^{11/6}}\,.$$

Note that (3.12) also implies $|\varphi(t)-1+\frac{t^2}{2n}|\leqslant\frac{L|t|^3}{6n^{3/2}}$. Thus $|\log\varphi(t)+\frac{t^2}{2n}|\leqslant\frac{L|t^3|}{1.5n^{3/2}}$, which implies

$$|\log(\varphi_n(t)\cdot e^{t^2/2})|=\left|\log\varphi_n(t)+\frac{t^2}{2}\right|\leqslant\frac{L|t^3|}{1.5n^{1/2}}\leqslant\frac{1}{8}\,,$$

where in the last step we used the assumption on t. Applying $|\mathrm{e}^z-1|\leqslant|z|\mathrm{e}^{|z|}$ with $z=\log(\varphi_n(t)\cdot e^{t^2/2})$, we have

$$|\varphi_n(t)\cdot e^{t^2/2}-1|\leqslant|\log(\varphi_n(t)\cdot e^{t^2/2})|\mathrm{e}^{|\log\varphi(t)+\frac{t^2}{2n}|}\leqslant\frac{L|t^3|}{1.5n^{1/2}}\cdot\mathrm{e}^{1/8}\leqslant\frac{L|t^3|}{n^{1/2}}$$

as desired. This finishes the proof. □

Lemma 3.18. *If $|t|\leqslant\frac{n^{1/2}}{4L}$, then $|\varphi_n(t)|\leqslant\mathrm{e}^{-t^2/3}$.*

Proof. Let $\varphi(t):=\mathbb{E}\exp(\mathrm{i}tX_1/\sqrt{n})$. Then

$$\begin{aligned}|\varphi(t)^2|&=\mathbb{E}\exp(\mathrm{i}tX_1/\sqrt{n})\cdot\mathbb{E}\exp(-\mathrm{i}tX_2/\sqrt{n})=\mathbb{E}\exp(\mathrm{i}t(X_1-X_2)/\sqrt{n})\\&=1+\frac{\mathrm{i}t}{\sqrt{n}}\mathbb{E}(X_1-X_2)+\frac{(\mathrm{i}t)^2}{2n}\mathbb{E}(X_1-X_2)^2+\frac{(\mathrm{i}t)^3}{6n^{3/2}}\mathbb{E}[(X_1-X_2)^3\exp(\mathrm{i}s(X_1-X_2)/\sqrt{n})]\end{aligned}$$

for some s between 0 and t. Thus

$$|\varphi(t)^2|\leqslant1-\frac{t^2}{n}+\frac{|t|^3}{6n^{3/2}}\cdot8L\leqslant\exp\Big(-\frac{t^2}{n}+\frac{4|t|^3L}{3n^{3/2}}\Big)\leqslant\exp\Big(-\frac{t^2}{n}+\frac{t^2}{3n}\Big)=\exp\Big(-\frac{2t^2}{3n}\Big)\,.$$

This implies $|\varphi_n(t)|=|\varphi(t)^n|\leqslant e^{-t^2/3}$ as desired. □

Combining Lemmas 3.17 and 3.18, it is not hard to deduce that

$$|\varphi_n(t)-\mathrm{e}^{-t^2/2}|\leqslant16Ln^{-1/2}|t|^3\mathrm{e}^{-t^2/3}\tag{3.13}$$

for all $|t|\leqslant\frac{n^{1/2}}{4L}$. Setting $T=\frac{n^{1/2}}{4L}$ in (3.11) and applying (3.13) yields

$$\begin{aligned}\Delta(S_n/\sqrt{n},Y)&\leqslant\frac{2}{\pi}\int_0^T\frac{|\varphi_n(t)-\mathrm{e}^{-t^2/2}|}{t}\mathrm{d}t+\frac{24}{\sqrt{2\pi}\pi T}\\&\leqslant\frac{2}{\pi}\int_0^{\frac{n^{1/2}}{4L}}16Ln^{-1/2}t^2\mathrm{e}^{-t^2/3}\mathrm{d}t+\frac{96L}{\sqrt{2\pi}\pi n^{1/2}}\\&\leqslant\frac{L}{\sqrt{n}}\cdot\left(\frac{32}{\pi}\int_0^\infty t^2\mathrm{e}^{-t^2/3}\mathrm{d}t+\frac{96}{\sqrt{2\pi}\pi}\right).\end{aligned}$$

This proves Theorem 3.11 with

$$C=\frac{32}{\pi}\int_0^\infty t^2\mathrm{e}^{-t^2/3}\mathrm{d}t+\frac{96}{\sqrt{2\pi}\pi}\approx35.64\,.$$

Remark 3.19. The order $O(n^{-1/2})$ in Theorem 3.11 is sharp. Consider the case of *Rademacher* random variable, namely $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Suppose n is even. Then

$$\begin{aligned}\mathbb{P}(S_n/\sqrt{n} = 0) = \mathbb{P}(S_n = 0) = \binom{n}{n/2} \cdot (1/2)^n &= \frac{n!}{((n/2)!)^2} \cdot \frac{1}{2^n} \\ &\approx \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n \Big(\sqrt{2\pi n/2}\Big(\frac{n}{2e}\Big)^{n/2}\Big)^{-2} \cdot \frac{1}{2^n} \approx \sqrt{\frac{2}{\pi n}}\,.\end{aligned}$$

Here $A \approx B$ is defined as $A = B(1+o(1))$. This implies a jump of $S_n/\sqrt{n}$ at 0, thus it is impossible to approximate $S_n\sqrt{n}$ by any continuous distribution with a precision better than $\sqrt{\frac{1}{2\pi n}}$. In this case,

$$C = \sqrt{\frac{1}{2\pi}} \approx 0.389\,.$$

There was in fact an effort of getting the smallest possible C in Theorem 3.11. Currently the best result is 0.4748 [14]. It is believed that the constant cannot be smaller than 0.4097.

3.3. Local limit theorem (good news: no proofs)

Let $\{X_n\}_{n\geqslant 1}$ be i.i.d. with $\mathbb{E}X_1 = 0$ and $\operatorname{Var} X_1 = 1$. Let $S_n := X_1 + \cdots + X_n$. The Central Limit Theorem, Theorem 3.1, is a *qualitative* result. It implies that, for any fixed (independent of n) interval $[a, b] \subset \mathbb{R}$, we have

$$\mathbb{P}(S_n/\sqrt{n} \in [a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \mathrm{d}x + o(1)\,.$$

Here the small o notation means $o(1) \to 0$ as $n \to \infty$. On the other hand, the Berry-Esseen bound, Theorem 3.11, is a *quantitative* result. It implies that for any interval $[a, b] \in \mathbb{R}$,

$$\mathbb{P}(S_n/\sqrt{n} \in [a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \mathrm{d}x + O(n^{-1/2})\,. \tag{3.14}$$

Here the big O notation means, there exists a constant $C > 0$, independent of n, such that $|O(n^{-1/2})| \leqslant Cn^{-1/2}$ for all $n \geqslant 1$. In (3.14), when a is bounded, the estimate is meaningful as long as $a - b \gg n^{-1/2}$.

The quantitative result does not involve taking the limit of the fundamental large parameter (in our case, it is n), and it is generally stronger than the asymptotic, qualitative results.

Now, can we push Theorem 3.11 and (3.14) a bit further, and obtain a meaningful result when $a - b = O(n^{-1/2})$? If so, the most ideal guess would be

$$\mathbb{P}(S_n/\sqrt{n} \in [a, a + h/\sqrt{n}]) = \frac{h}{\sqrt{2\pi n}} \mathrm{e}^{-a^2/2} + o(n^{-1/2}) \tag{3.15}$$

for any bounded $a, h \in \mathbb{R}$.

However, (3.15) cannot always hold. Consider the example as in Remark 3.19, where $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Let n be even. Then

$$\mathbb{P}(S_n/\sqrt{n} = k/\sqrt{n}) = \binom{n}{(n-k)/2} \cdot (1/2)^n = \frac{n!}{((n-k)/2)! \cdot ((n+k)/2)!} \cdot \frac{1}{2^n} \approx \sqrt{\frac{2}{\pi n}}$$

when k is bounded and even. $\mathbb{P}(S_n/\sqrt{n} = k/\sqrt{n}) = 0$ if k is odd. Thus

$$\mathbb{P}(S_n/\sqrt{n} \in [0, 0 + 0.1/\sqrt{n}]) = \mathbb{P}(S_n/\sqrt{n} = 0) = \sqrt{\frac{2}{\pi n}} \neq \frac{0.1}{\sqrt{2\pi n}},$$

which disproves (3.15). The reason of the above is the special structure of X_1, which makes $S_n/\sqrt{n}$ only takes values in $2\mathbb{Z}/\sqrt{n}$.

To distinguish this case, let $\mathcal{L}(a, \lambda) := a + \lambda\mathbb{Z}$. A random variable X is said to have *lattice distribution* if there exists constants $a \in \mathbb{R}$ and $\lambda > 0$ such that $\mathbb{P}(X \in \mathcal{L}(a, \lambda)) = 1$. Obviously, $\mathbb{P}(X \in \mathcal{L}(a, \lambda)) = 1$ implies $\mathbb{P}(X \in \mathcal{L}(a, \lambda/m)) = 1$ for any $m \in \mathbb{N}_+$. Let

$$\Lambda := \max\{\lambda : \exists a \in \mathbb{R}, \mathbb{P}(X_1 \in \mathcal{L}(a, \lambda)) = 1\}.$$

The next results links the concept of lattice distribution to the behaviour of the characteristic function.

Proposition 3.20. *Let*

$$t_* := \inf\{t_0 : t_0 > 0, |\varphi_X(t_0)| = 1, |\varphi_X(t)| < 1 \text{ for all } t \in (0, t_0)\}.$$

There are three possibilities.

(i) $t_* \in (0, \infty)$. *Then* $|\varphi_X(t_*)| = 1$ *and* $|\varphi_X(t)| < 1$ *for all* $t \in (0, t_*)$. *In this case* X *has a lattice distribution.*

(ii) $t_* = \infty$. *Then* $|\varphi_X(t)| < 1$ *for all* $t > 0$. *In this case,* X *has non-lattice distribution.*

(iii) $t_* = 0$. *Then* $|\varphi_X(t)| = 1$ *for all* $t \in \mathbb{R}$. *In this case,* X *is degenerate.*

Example 3.21. Let X_1 satisfy $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Then $\varphi_{X_1}(t) = \frac{1}{2}(\mathrm{e}^{\mathrm{i}t} + \mathrm{e}^{-\mathrm{i}t}) = \cos t$. Then $|\varphi_{X_1}(t)| = 1$ if and only if $t \in \pi Z$. Thus $t_* = \pi$ and $\Lambda = 2$.

Theorem 3.22. *Let* $\{X_n\}_{n\geqslant 1}$ *be i.i.d. with* $\mathbb{E}X_1 = 0$ *and* $\operatorname{Var} X_1 = 1$. *Let* $S_n := X_1 + \cdots + X_n$. *Suppose there exists constants* $a \in \mathbb{R}$ *and* $\Lambda > 0$ *such that* $\mathbb{P}(X_1 \in \mathcal{L}(a, \Lambda)) = 1$. *Then*

$$\mathbb{P}(S_n/\sqrt{n} = x) = \begin{cases} \frac{\Lambda}{\sqrt{2\pi n}} \mathrm{e}^{-x^2/2} + o(n^{-1/2}) & \text{if} \quad x \in \mathcal{L}(\sqrt{n}a, \Lambda) \\ 0 & \text{otherwise}. \end{cases}$$

The error term $o(n^{-1/2})$ *is uniform in* x.

Theorem 3.23. *Let* $\{X_n\}_{n\geqslant 1}$ *be i.i.d. with* $\mathbb{E}X_1 = 0$ *and* $\operatorname{Var} X_1 = 1$. *Let* $S_n := X_1 + \cdots + X_n$. *Suppose* X_1 *has non-lattice distribution. Then*

$$\mathbb{P}(S_n/\sqrt{n} \in [a, a + h/\sqrt{n}]) = \frac{h}{\sqrt{2\pi n}} \mathrm{e}^{-a^2/2} + o(n^{-1/2}) \tag{3.16}$$

for any bounded $a, h \in \mathbb{R}$. *Here the error term* $o(n^{-1/2})$ *is independent of* a*, but it may depend on* h.

We will not prove Proposition 3.20, Theorems 3.22 and 3.23 here. The readers may refer to [8].

3.4. Large and moderate deviations

Let us go back to the setting of the central limit theorem. We have a sequence of i.i.d. random variables $\{X_n\}_{n\geqslant 1}$, with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}X_1^2 = \sigma^2$. $S_n := X_1 + \cdots + X_n$. We have seen from Theorem 3.1 that

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}} \to \mathcal{N}(0,1)\,.$$

In other words, $S_n - n\mu$ is *most likely* to be on the scale $O(\sqrt{n})$. On the other hand, it is still *possible* for S_n to take larger values, even to the scale $O(n)$. *Large deviation* theory studies the quantity

$$\mathbb{P}(S_n - n\mu \geqslant xn)$$

for fixed $x > 0$, while *moderate deviation* studies the behavior of

$$\mathbb{P}(S_n - n\mu \geqslant x\sqrt{n})$$

for any $1 \ll x \ll n^{1/2}$.

3.4.1. Large deviation. In this section, we prove the following result.

Theorem 3.24 (Cramér's theorem)**.** *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, and set $S_n := X_1 + \cdots + X_n$. Suppose $\mathbb{E}(\exp(tX_1)) < \infty$ for t in a neighborhood of 0. For $t \in \mathbb{R}$, denote the cumulant generating function by*

$$K(t) := \log \mathbb{E}(\exp(tX_1))\,.$$

Then

$$\frac{1}{n}\log \mathbb{P}(S_n \geqslant xn) = -\sup_{t\geqslant 0}(tx - K(t)) + o(1)\,.$$

for all fixed $x > \mathbb{E}X_1$. Equivalently,

$$\mathbb{P}(S_n \geqslant xn)^{1/n} = \exp(-\sup_{t\geqslant 0}(tx - K(t)))(1 + o(1)) = \inf_{t>0} \mathbb{E}\mathrm{e}^{tX_1 - tx}(1 + o(1))\,.$$

For a random variable X_1, its *rate function* is defined via $I(x) := \sup_{t\in\mathbb{R}}(tx - K(t))$. Theorem 3.24 implies that

$$\mathbb{P}(S_n \geqslant xn) \approx \exp(-nI(x))\,.$$

Exercise 3.25. (i) Let X_1 be a Bernoulli random variable with parameter p, i.e. $\mathbb{P}(X_1 = 1) = p$, $\mathbb{P}(X_1 = 0) = 1 - p$, $p \in (0,1)$. Show that the rate function is

$$I(x) = \begin{cases} x\log\left(\dfrac{x}{p}\right) + (1-x)\log\left(\dfrac{1-x}{1-p}\right), & x \in [0,1] \\ +\infty, & \text{otherwise.} \end{cases}$$

(ii) Let $X_1 \overset{d}{=} \mathcal{N}(0,1)$. Show that $I(x) = x^2/2$.

We present the proof of Theorem 3.24 given in [5]. For $x \in \mathbb{R}$, define

$$s(x) := \sup_{n\geqslant 1} \frac{1}{n}\log \mathbb{P}(S_n \geqslant nx)\,.$$

We have the following duality.

Proposition 3.26. *For all $t \geqslant 0$, we have*

$$K(t) = \sup_{u\in\mathbb{R}} \big(tu + s(u)\big)\,.$$

Proof. (i) One side of the equality is easy to prove. From the Markov inequality, we have

$$K(t) = \log \mathbb{E}\exp(tX_1) = \frac{1}{n}\log \mathbb{E}\exp(tS_n) \geqslant \frac{1}{n}\log\Big(e^{ntu}\mathbb{P}(\mathrm{e}^{tS_n} \geqslant \mathrm{e}^{ntu})\Big) = tu + \frac{1}{n}\log \mathbb{P}(S_n \geqslant nu)\,.$$

Hence, taking the supremum over n and then over u, we get

$$K(t) \geqslant \sup_{u\in\mathbb{R}} \big(tu + s(u)\big)\,.$$

(ii) In this step we prove the equality for $t = 0$. Note that for al $u \in \mathbb{R}$, we have

$$s(u) = \sup_{n\geqslant 1}\frac{1}{n}\log \mathbb{P}(S_n \geqslant nu) \geqslant \log \mathbb{P}(S_1 \geqslant u) = \log \mathbb{P}(X_1 \geqslant u)\,.$$

Letting $u \to -\infty$ in the above yields

$$\sup_u s(u) \geqslant \lim_{u\to-\infty} s(u) \geqslant \lim_{u\to-\infty} \log \mathbb{P}(X_1 \geqslant u) = 0 = K(0)\,.$$

Together with part (i) we proved the proposition for $t = 0$.

(iii) Now we prove the general result. Let $M > 0$. Since $\{X_n\}_{n\geqslant 1}$ are i.i.d. we have

$$\begin{aligned}\log \mathbb{E}(\mathrm{e}^{tX_1}\mathbf{1}_{|X_1|\leqslant M}) &= \frac{1}{n}\log \mathbb{E}(\mathrm{e}^{t(X_1+\cdots+X_n)}\mathbf{1}_{|X_1|\leqslant M}\cdots \mathbf{1}_{|X_n|\leqslant M}) \leqslant \frac{1}{n}\log \mathbb{E}(\mathrm{e}^{tS_n}\mathbf{1}_{|S_n|\leqslant nM})\\ &= \frac{1}{n}\log \mathbb{E}\Big(\Big(\mathrm{e}^{-ntM} + \int_{-M}^{S_n/n} nt\mathrm{e}^{ntu}\mathrm{d}u\Big)\mathbf{1}_{|S_n|\leqslant nM}\Big)\Big)\\ &\leqslant \frac{1}{n}\log\Big(\mathrm{e}^{-ntM} + \int_{\mathbb{R}} \mathbb{E}\big(\mathbf{1}_{-M\leqslant u\leqslant S_n/n}\mathbf{1}_{|S_n|\leqslant nM}\big)nt\mathrm{e}^{ntu}\mathrm{d}u\Big)\,.\end{aligned}$$

Here in the last step we used the Fubini theorem (it is possible since the integrand is nonnegative). Since

$$\mathbb{E}\big(\mathbf{1}_{-M\leqslant u\leqslant S_n/n}\mathbf{1}_{|S_n|\leqslant nM}\big) \leqslant \mathbb{E}\big(\mathbf{1}_{S_n\geqslant un}\mathbf{1}_{-M\leqslant u\leqslant M}\big) = \mathbb{P}(S_n \geqslant un)\mathbf{1}_{|u|\leqslant M} \leqslant \mathrm{e}^{ns(u)}\mathbf{1}_{|u|\leqslant M}\,,$$

we get

$$\begin{aligned}\log \mathbb{E}(\mathrm{e}^{tX_1}\mathbf{1}_{|X_1|\leqslant M}) &\leqslant \frac{1}{n}\log\Big(\mathrm{e}^{-ntM} + \int_{-M}^{M} \mathrm{e}^{ns(u)}nt\mathrm{e}^{ntu}\mathrm{d}u\Big)\\ &\leqslant \frac{1}{n}\log\Big(\mathrm{e}^{-ntM} + 2Mnt\exp(n\sup_{u\in\mathbb{R}}(s(u)+tu))\Big)\,.\end{aligned}$$

By part (ii), we can choose M large enough so that

$$-tM \leqslant \sup_{u\in\mathbb{R}}(s(u)+tu)\,.$$

Then

$$\log \mathbb{E}(\mathrm{e}^{tX_1}\mathbf{1}_{|X_1|\leqslant M}) \leqslant \frac{1}{n}\log\Big((1+2Mnt)\exp(n\sup_{u\in\mathbb{R}}(s(u)+tu))\Big) = \frac{1}{n}\log(1+2Mnt) + \sup_{u\in\mathbb{R}}(s(u)+tu)\,.$$

Taking $n \to \infty$ yields

$$\log \mathbb{E}(\mathrm{e}^{tX_1}\mathbf{1}_{|X_1|\leqslant M}) \leqslant \sup_{u\in\mathbb{R}}(s(u)+tu)\,.$$

Finally, taking $M \to \infty$ yields $K(t) \leqslant \sup_{u\in\mathbb{R}}\big(tu+s(u)\big)$ as desired. □

To use the above duality, we need to show that s is a concave function.

Lemma 3.27. *We have*

$$\lim_n \frac{1}{n} \log \mathbb{P}(S_n \geqslant nx) = s(x) \in [-\infty, 0]$$

for all $x \in \mathbb{R}$.

Proof. If $\mathbb{P}(X_1 \geqslant x) = 0$, then the result is trivially true. Thus can we assume that $\mathbb{P}(X_1 \geqslant x) > 0$. Fix $m \geqslant 1$. For $n \geqslant m$, write $n = mp + q$, where $p \geqslant 1$, $q \in \{0, 1, 2, ..., m-1\}$. Then

$$\mathbb{P}(S_n \geqslant nx) \geqslant \mathbb{P}(S_m \geqslant mx)^p \cdot \mathbb{P}(S_q \geqslant qx)\,,$$

which implies

$$\frac{1}{n} \log \mathbb{P}(S_n \geqslant nx) \geqslant \frac{p}{n} \log(\mathbb{P}(S_m \geqslant mx) + \frac{1}{n} \log \mathbb{P}(S_q \geqslant qx)\,.$$

As $\lim_n \frac{p}{n} = \frac{1}{m}$ and $\lim_n \frac{1}{n} \log \mathbb{P}(S_q \geqslant qx) = 0$, we have

$$\liminf_n \frac{1}{n} \log \mathbb{P}(S_n \geqslant nx) \geqslant \frac{1}{m} \log(\mathbb{P}(S_m \geqslant mx)\,.$$

As the above holds for all m, taking supremum over m yields

$$\liminf_n \frac{1}{n} \log \mathbb{P}(S_n \geqslant nx) \geqslant s(x) \geqslant \limsup_n \frac{1}{n} \log \mathbb{P}(S_n \geqslant nx)\,,$$

where the last inequality comes from the definition of $s(x)$. This finishes the proof. □

Armed with Lemma 3.27, we can prove the following.

Corollary 3.28. *The function $s : \mathbb{R} \to [-\infty, 0]$ is concave.*

Exercise 3.29. Prove Corollary 3.28.

Now we can finish the proof of Theorem 3.24. From Proposition 3.26 we know that

$$\inf_{t \geqslant 0} (K(t) - tx) = \inf_{t \geqslant 0} \sup_{u \in \mathbb{R}} (t(u - x) + s(u)) \tag{3.17}$$

It remains to prove that the (3.17) equals $s(x)$. The result is standard in convex function theory and known as Fenchel-Legendre duality. Let us give an elementary proof in our setting. The right-hand side of (3.17) is clearly greater than or equal to $s(x)$: take $u = x$. To prove the converse inequality we set

$$x_* := \inf\{x : \mathbb{P}(X_1 \geqslant x) = 0\} \in (-\infty, \infty]\,.$$

If $x < x_*$, then $s(x) > -\infty$. Since s is concave, the function

$$g(u) := \frac{s(x) - s(u)}{x - u}$$

is non-increasing in $\mathbb{R} \backslash \{x\}$. Set

$$-t_* := \lim_{u \to x_-} g(u) \in [-\infty, 0]\,.$$

Then we have $s(u) + t_*(u - x) \leqslant s(x)$ for all $u \in \mathbb{R}$. From this the result follows.

If $x \geqslant x_*$, then $\mathbb{P}(X_1 > x_*) = \lim_{\varepsilon \to 0_+} \mathbb{P}(X_1 \geqslant x_* + \varepsilon) = 0$. For all $t \geqslant 0$ and $\varepsilon > 0$,

$$K(t) - tx = \log \mathbb{E} e^{t(X_1 - x)} = \log \mathbb{E}\Big(e^{t(X_1 - x)}(\mathbf{1}_{X_1 < x - \varepsilon} + \mathbf{1}_{x - \varepsilon \leqslant X_1 \leqslant x_*})\Big) \leqslant \log\Big(\mathrm{e}^{-t\varepsilon} + \mathbb{P}(X_1 \geqslant x - \varepsilon)\Big)\,.$$

Taking minimum over t and sending $\varepsilon \to 0$, we get

$$\inf_{t \geqslant 0} (K(t) - tx) \leqslant \log \mathbb{P}(X_1 \geqslant x) = s(x)\,.$$

This finishes the proof.

3.4.2. Moderate deviation. Recall from Theorem 1.10 (iv) that for a random variable X with all finite moments, its kth cumulant is defined via

$$\mathcal{C}_k(X) := (-\mathrm{i})^k \cdot \big(\partial_t^k \log \mathbb{E}[\mathrm{e}^{\mathrm{i}tX}]\big)\big|_{t=0}$$

for all $k \geqslant 1$. If $\mathbb{E}[\mathrm{e}^{tX}] < \infty$ when t is near 0, we can also write

$$\mathcal{C}_k(X) := \big(\partial_t^k \log \mathbb{E}[\mathrm{e}^{tX}]\big)\big|_{t=0}.$$

We have the following result.

Theorem 3.30. *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables with $\mathbb{E}X_1 = 0$ and $\operatorname{Var} X_1 = 1$. Set $S_n = X_1 + \cdots + X_n$, and $Y \overset{d}{=} \mathcal{N}(0,1)$. Suppose $\mathbb{E}e^{tX_1} < \infty$ for t in a small neighborhood of 0. Then for any $0 < x \ll n^{1/2}$, we have*

$$\frac{\mathbb{P}(S_n > \sqrt{n}x)}{\mathbb{P}(Y > x)} = \exp\bigg(\frac{x^3}{\sqrt{n}}\lambda\Big(\frac{x}{\sqrt{n}}\Big)\bigg)\Big(1 + O\Big(\frac{1+x}{\sqrt{n}}\Big)\Big)\,.$$

Here

$$\lambda(z) := \sum_{k=0}^{\infty} \frac{\mathcal{C}_{k+3}(X_1)}{(k+3)!} z^k\,,$$

and

The proof of Theorem 3.30 is complicated and we omit here. Interested readers may refer to [11] for the case that X_1 has a density. We will interpret what the result tells us.

(i) When $x > 0$ does not grow with n, Theorem 3.30 can be reduced to

$$\frac{\mathbb{P}(S_n > \sqrt{n}x)}{\mathbb{P}(Z > x)} = 1 + O(n^{-1/2})\,.$$

This is a result that agrees with Theorem 3.11, the Berry-Esseen bound.

(ii) When $1 \leqslant x \ll n^{1/6}$, we have

$$\frac{\mathbb{P}(S_n > \sqrt{n}x)}{\mathbb{P}(Z > x)} = 1 + O(x^3 n^{-1/2})\,.$$

Thus the leading contribution of $\mathbb{P}(S_n > \sqrt{n}x)$ is still given by the standard Gaussian random variable.

(iii) When $n^{1/6} \ll x \ll n^{1/2}$ and $\mathcal{C}_3(X_1) = \mathbb{E}X_1^3 \neq 0$, we have

$$\frac{\mathbb{P}(S_n > \sqrt{n}x)}{\mathbb{P}(Z > x)} = \mathrm{e}^{\mathcal{C}_3(X_1)x^3/\sqrt{n}}(1 + O(xn^{-1/2}))\,.$$

In this case, the behavior of $\mathbb{P}(S_n > \sqrt{n}x)$ is governed by the third cumulant of X_1.

(iv) When $X_1 \overset{d}{=} \mathcal{N}(0,1)$, we have $\mathbb{E}[\mathrm{e}^{tX}] = e^{t^2/2}$. Thus

$$\mathcal{C}_k(X) := \big(\partial_t^k \log \mathbb{E}[\mathrm{e}^{tX}]\big)\big|_{t=0} = \begin{cases} 1 & \text{if } k = 2 \\ 0 & \text{otherwise}\,. \end{cases}$$

This makes $\lambda(x/\sqrt{n}) \equiv 0$.

(v) From Theorems 3.1, 3.24 and 3.30, we see that when $0 \leqslant x \ll n^{1/6}$, the (leading) behavior of $\mathbb{P}(S_n \geqslant \sqrt{n}x)$ agree with that of the standard Gaussian, regardless of the distribution of X_1. We can call this a *universality* result. When $x \geqslant n^{1/6}$, $\mathbb{P}(S_n \geqslant \sqrt{n}x)$ depends on the distribution of X_1, and thus the result is no longer universal.

Homework 2

Question 1. Let $\{X_n\}_{n\geqslant 1}$ be i.i.d. random variables having the uniform distribution on $[-1,1]$. Show that

$$\frac{X_1+\cdots+X_n}{\sqrt{X_1^2+\cdots+X_n^2}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Question 2. Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables with $\mathbb{E}X_1 = \mu > 0$ and $\operatorname{Var} X_1 = \sigma^2$. For $t > 0$, define the random variable

$$N_t := \sup\{n : X_1+\cdots+X_n \leqslant t\}\,.$$

Show that

$$\frac{\mu^{3/2}(N_t - t/\mu)}{\sigma\sqrt{t}} \xrightarrow{d} \mathcal{N}(0,1)$$

as $t\to\infty$.

Question 3. Use the central limit theorem to show that

$$\mathrm{e}^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} \to \frac{1}{2}\,.$$

Question 4. Let $Z \overset{d}{=} \mathcal{N}(0,1)$. Show that

$$\frac{1}{\sqrt{2\pi}}(x^{-1}-x^{-3})\mathrm{e}^{-x^2/2} \leqslant \mathbb{P}(Z\geqslant x) \leqslant \frac{1}{x\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$$

for all $x > 0$.

Question 5. Solve Exercise 3.25.

Question 6. Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, with $\mathbb{E}X_1 = 0$ and $\mathbb{E}\exp(tX_1) = \infty$ for all $t > 0$. Show that

$$\lim_n \frac{1}{n}\log\mathbb{P}(X_1+\cdots+X_n \geqslant nx) = 0$$

for all fixed $x > 0$.

4 Law of large numbers

4.1. The strong law of large numbers

I trust that everyone at this stage can prove the following result.

Exercise 4.1. Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables, with $\mathbb{E}X_n = \mu$ for all n, and $\sup_n \mathbb{E}X_n^2 < \infty$. We assume that the random variables are uncorrelated, i.e.

$$\mathrm{Cov}(X_i, X_j) := \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) = 0$$

for all $i \neq j$. Show that

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{L_2} \mu \tag{4.1}$$

as $n \to \infty$. In particular, the convergence also holds in probability.

The relation (4.1) is called the *weak law of large numbers*. In fact, even only given the existence of the first moment, we can prove the convergence in probability quickly.

Proposition 4.2. *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, and assume $\mathbb{E}X_1 = \mu$. Let $S_n := X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$$

as $n \to \infty$.

Proof. Let $\varphi_1(t) := \mathbb{E}\exp(\mathrm{i}tX_1)$. By Taylor expansion,

$$\varphi_1(t/n) = 1 + \mathrm{i}\mu t/n + o(1/n)\,.$$

Then

$$\begin{aligned}\mathbb{E}\exp(\mathrm{i}tS_n/n) &= \varphi_1(t/n)^n = \left(1 + \mathrm{i}\mu t n^{-1} + o(n^{-1})\right)^n = \exp\left(n\log\left(1 + \mathrm{i}\mu t n^{-1} + o(n^{-1})\right)\right)\\ &= \exp\left(n\left(\mathrm{i}\mu t n^{-1} + o(n^{-1})\right)\right) = \exp(\mathrm{i}\mu t) + o(1)\,.\end{aligned}$$

By Theorem 1.14, we have $S_n/n \xrightarrow{d} \mu$, which is equivalent to the desired result. ☐

Our main goal in Section 4.1 is to prove the following *strong law of large numbers* (SLLN).

Theorem 4.3. *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, and assume $\mathbb{E}X_1 = \mu$. Let $S_n := X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu$$

as $n \to \infty$.

Remark 4.4. The SLLN, in the strongest possible form, only requires the sequence $\{X_n\}_{n\geqslant 1}$ to be pairwise independent. It is easy to construct pairwise independent random variables that are not *mutually* independent. (Think about independent X, Y both having the Bernoulli distribution with parameter $1/2$, and let $Z = X + Y \pmod 2$.)

The classic proof of Theorem 4.3 (you can find it in [6] for instance) is lengthy and it does not tell us too much beyond the proof itself. Here we present one shorter alternative.

For a sequence of random variables $\{X_n\}_{n\geqslant 1}$, we define

$$\mathcal{F}_n := \sigma(X_n, X_{n+1}, ...)$$

for all $n \geqslant 1$. The definition means, $\mathcal{F}_n$ is the smallest σ-algebra that contains all $\sigma(X_n)$[5], $\sigma(X_{n+1}), \sigma(X_{n+2}), ...$ The *asymptotic σ-algebra* of $\{X_n\}$ is defined as

$$\mathcal{F}_\infty := \bigcap_{n\geqslant 1}^{\infty} \mathcal{F}_n = \bigcap_{n\geqslant 1}^{\infty} \sigma(X_k; k \geqslant n)\,.$$

Elements of $\mathcal{F}_\infty$ are called *tail events*. For example, it contains the set of convergence (in $[-\infty,\infty]$) of the series $\sum_{k\geqslant 1} X_k$:

$$\Omega_1 = \left\{\omega \in \Omega : \liminf_n \sum_{k=1}^{n} X_k(\omega) = \limsup_n \sum_{k=1}^{n} X_k(\omega)\right\},$$

since for each $m \geqslant 1$,

$$\Omega_1 = \left\{\omega \in \Omega : \liminf_n \sum_{k=m}^{n} X_k(\omega) = \limsup_n \sum_{k=m}^{n} X_k(\omega)\right\} \subset \mathcal{F}_m\,,$$

which implies $\Omega_1 \in \cap_m \mathcal{F}_m = \mathcal{F}_\infty$.

Lemma 4.5 (Kolmogorov 0-1 law)**.** *Let $\{X_n\}_{n\geqslant 1}$ be independent random variables, and let A be a tail event of $\{X_n\}$. Then*

$$\mathbb{P}(A) \in \{0,1\}\,.$$

Proof. Obviously, for any $n \geqslant 1$ the sets $\sigma(X_1, X_2, ...X_n)$ and $\mathcal{F}_\infty$ are *independent*, meaning

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

for any $A \in \sigma(X_1, X_2, ...X_n)$ and $B \in \mathcal{F}_\infty$. As a result, $\bigcup_{n=1}^{\infty} \sigma(X_1, X_2, ...X_n)$ and $\mathcal{F}_\infty$ are also independent. Fix $B \in \mathcal{F}_\infty$, and define

$$\mathcal{L} := \{A \subset \Omega : \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\}\,.$$

It is easy to verify that $\mathcal{L}$ is a *λ-system*, meaning

(i) $\Omega \in \mathcal{L}$;

(ii) $X \in \mathcal{L} \implies X^c \in \mathcal{L}$;

(iii) For pairwise disjoint $\{X_n\}$ in $\mathcal{L}$, we have $\cup_n X_n \in \mathcal{L}$.

[5] The σ-algebra generated by a random variable X is the smallest σ-algebra containing the set $\{X^{-1}(U) : U \subset \mathbb{R}$ *is open*$\}$.

In addition, $\bigcup_{n=1}^{\infty} \sigma(X_1, X_2, ...X_n)$ is a π-*system*, meaning it is closed under finite intersections. Recall that the $\pi - \lambda$ *theorem*[6] states

"If $\mathcal{A}$ is a π-system, $\mathcal{B}$ is a λ-system, and $\mathcal{A} \subset \mathcal{B}$, then $\sigma(\mathcal{A}) \subset \mathcal{B}$."

Thus applying the theorem with $\mathcal{A} = \bigcup_{n=1}^{\infty} \sigma(X_1, X_2, ...X_n)$ and $\mathcal{B} = \mathcal{L}$ yields

$$\sigma\Big(\bigcup_{n=1}^{\infty} \sigma(X_1, X_2, ...X_n)\Big) \subset \mathcal{L}\,.$$

Obviously,

$$\mathcal{F}_1 = \sigma\{\sigma(X_1), \sigma(X_2), ...\} \subset \sigma\Big(\bigcup_{n=1}^{\infty} \sigma(X_1, X_2, ...X_n)\Big) \subset \mathcal{L}\,.$$

Hence $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{F}_1$. As B is arbitrary, $\mathcal{F}_1$ and $\mathcal{F}_\infty$ are independent. Since $\mathcal{F}_\infty \subset \mathcal{F}_1$, $\mathcal{F}_\infty$ is independent with itself. Thus for any $A \in \mathcal{F}_\infty$ we have

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$$

This finishes the proof. □

Corollary 4.6. *Let $\{X_n\}_{n \geqslant 1}$ be a sequence of independent random variables. Then the probability that $\sum_{n=1}^{\infty} X_n$ converges is either 0 or 1. Similarly, the probability that*

$$\frac{1}{n}\sum_{k=1}^{n} X_k$$

converges is either 0 or 1.

Having Lemma 4.5 at hand, we now present a short proof of Theorem 4.3 found in [7]. The main step is the following result.

Proposition 4.7. *Let $\{X_n\}_{\geqslant 1}$ be i.i.d. random variables with $\mathbb{E}X_1 > 0$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\inf_{n \geqslant 1} S_n > -\infty$$

almost surely. In particular,

$$\liminf_n \frac{S_n}{n} \geqslant 0$$

almost surely.

Let $\{X_n\}$ and S_n be as in Theorem 4.3, with $\mathbb{E}X_1 = \mu$. Let $c < \mu$. Applying Proposition 4.7 for $\{X_n - c\}_{n \geqslant 1}$, we get

$$\liminf_n \frac{S_n - nc}{n} \geqslant 0$$

almost surely. Thus $\liminf_n S_n/n \geqslant c$ almost surely. Since c is any number smaller than μ, we have

$$\liminf_n S_n/n \geqslant \mu \quad a.s.$$

Applying the same augment to $\{-X_n\}$ to see that $\liminf_n -S_n/n \geqslant -\mu$ a.s. Thus

$$\limsup_n S_n/n \leqslant \mu \quad a.s.$$

This finishes the proof of Theorem 4.3.

[6] The proof is rather abstract and we omit here.

Proof of Proposition 4.7. Let $\mu = \mathbb{E}X_1 > 0$, and $J_n := \min\{S_1, ..., S_n\}$. Note that each J_n is integrable and $n \mapsto J_n$ is non-increasing. We have

$$J_n = X_1 + \min\{0, S_2 - S_1, S_3 - S_1, ..., S_n - S_1\} =: X_1 + \min\{0, J'_{n-1}\},$$

where J'_{n-1} has the same distribution as J_{n-1}. Therefore

$$\mathbb{E}J_n = \mu + \mathbb{E}\min\{0, J_{n-1}\}.$$

Subtracting $\mathbb{E}J_{n-1}$ from both sides yields

$$\mathbb{E}[J_n - J_{n-1}] = \mu - \mathbb{E}J_{n-1}^+.$$

Telescoping:

$$\sum_{k=1}^n \mathbb{E}J_k^+ = n\mu + \mathbb{E}[J_1 - J_{n+1}] \geqslant n\mu.$$

Thus

$$\liminf_n \frac{1}{n}\sum_{k=1}^n \mathbb{E}J_k^+ \geqslant \mu. \tag{4.2}$$

Since $n \mapsto \mathbb{E}J_n^+$ is non-increasing, we get $\mathbb{E}J_n^+ \geqslant \mu$ for all $n \geqslant 1$. Write $J_\infty := \lim J_n = \inf\{S_1, S_2, ...\}$. By the monotone convergence theorem,

$$\mathbb{E}J_\infty^+ \geqslant \mu.$$

In particular, $\mathbb{P}(J_\infty > -\infty) \geqslant \mathbb{P}(J_\infty > 0) > 0$. By Lemma 4.5, we get $\mathbb{P}(J_\infty > -\infty) = 1$ as desired. □

Theorem 4.3 can be extended to cases where $\mathbb{E}|X_1|^p < \infty$. We present the following result without proof.

Theorem 4.8 (Marcinkiewicz SLLN)**.** *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of i.i.d. random variables, such that $\mathbb{E}|X_1|^p < \infty$ for some $p \in (0, 2)$. Then there exists $a \in \mathbb{R}$ such that*

$$n^{-1/p}\sum_{k=1}^n (X_k - a) \xrightarrow{a.s.} 0. \tag{4.3}$$

Remark 4.9. (i) For $p \geqslant 2$, (4.3) can no longer hold: it would violate Theorem 3.1 (CLT). Nevertheless, we have

$$n^{-1/\alpha}\sum_{k=1}^n (X_k - a) \xrightarrow{a.s.} 0$$

for any $\alpha \in (0, 2)$.

(ii) For $p \in [1, 2)$, $a = \mathbb{E}X_1$, and Theorem 4.8 generalizes Theorem 4.3 to smaller scales: $\sum_{k=1}^n (X_k - a)$ is a.s. a point when viewed at the scale $n^{1/p}$. For $p \in (0, 1)$, a can be any real number.

4.2. The convergence of independent sums

The goal of Section 4.2 is to prove the following result.

Theorem 4.10 (Kolmogorov's Three-Series Theorem)**.** *Let $\{X_n\}_{n\geqslant 1}$ be independent random variables. Let $c>0$ and set $X_n^c = X_n \mathbf{1}_{|X_n|\leqslant c}$. Then $\sum_{n=1}^\infty X_n$ converges a.s. if the following conditions hold for some $c>0$, only if the following conditions hold for any $c>0$.*

(i) $\sum_{n=1}^\infty \mathbb{P}(|X_n|>c)<\infty$;

(ii) $\sum_{n=1}^\infty \mathbb{E}X_n^c$ *converges;*

(iii) $\sum_{n=1}^\infty \operatorname{Var} X_n^c < \infty$.

Remark 4.11. By Lemma 4.5, the probability that $\sum_{n=1}^\infty X_n$ converges is 0 or 1. Thus if at least one of (i) – (iii) in Theorem 4.10 does not hold, $\sum_{n=1}^\infty X_n$ diverges a.s.

Example 4.12. Let $\{Z_n\}_{\geqslant 1}$ be i.i.d. random variables, with $\mathbb{P}(Z_1=1)=\mathbb{P}(Z_1=-1)=1/2$. Let $X_n = Z_n/n^\alpha$ for some $\alpha>0$. Set $c=1$. Then

(i) $\sum_{n=1}^\infty \mathbb{P}(|X_n|>1)=0$;

(ii) $\sum_{n=1}^\infty \mathbb{E}X_n^c = \sum_{n=1}^\infty \mathbb{E}X_n = 0$;

(iii) $\sum_{n=1}^\infty \operatorname{Var}(X_n^c) = \sum_{n=1}^\infty \operatorname{Var}(X_n) = \sum_{n=1}^\infty n^{-2\alpha}$.

Thus by Theorem 4.10, $\sum_{n=1}^\infty X_n$ converges a.s. if and only of $\alpha > 1/2$.

Let $\{A_n\}_{n\geqslant 1}$ be a sequence of events. Define

$$\limsup_n A_n = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k \quad \text{and} \quad \liminf_n A_n = \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k\,.$$

Obviously, $\liminf A_n \subset \limsup A_n$. If these two sets are equal, we call it the limit of $\{A_n\}$, denoted by $\lim_n A_n$.

Lemma 4.13 (Borel-Cantelli)**.** *(i) If $\sum_n \mathbb{P}(A_n)<\infty$, then $\mathbb{P}(\limsup A_n)=0$.*
(ii) If $\{A_n\}$ are independent and $\sum_n \mathbb{P}(A_n)=\infty$, then $\mathbb{P}(\limsup A_n)=1$.

Proof. (i) We have

$$\mathbb{P}(\limsup A_n) = \mathbb{P}\Big(\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k\Big) \leqslant \mathbb{P}\Big(\bigcup_{k=n}^\infty A_k\Big) \leqslant \sum_{k\geqslant n} \mathbb{P}(A_k)$$

for all $n\geqslant 1$. As the RHS of the above goes to 0 as $n\to\infty$, we have $\mathbb{P}(\limsup A_n)\leqslant 0$, which implies the desired result.

(ii) We have for every $n\geqslant 1$ that

$$\begin{aligned} 1-\mathbb{P}\Big(\bigcup_{k=n}^\infty A_k\Big) &= \mathbb{P}\Big(\Big(\bigcup_{k=n}^\infty A_k\Big)^c\Big) = \mathbb{P}\Big(\bigcap_{k=n}^\infty A_k^c\Big) = \prod_{k\geqslant n}(1-\mathbb{P}(A_k)) \\ &\leqslant \prod_{k\geqslant n} \mathrm{e}^{-\mathbb{P}(A_k)} = \mathrm{e}^{-\sum_{k\geqslant n}\mathbb{P}(A_k)} = 0\,. \end{aligned}$$

Here in the third step we used independence, and the last step comes from $\sum_n \mathbb{P}(A_n) = \infty$. Thus $\mathbb{P}(\cup_{k=n}^{\infty} A_k) = 1$ for all $n \geqslant 1$, which implies

$$\mathbb{P}(\limsup A_n) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\Big) = \lim_n \mathbb{P}\Big(\bigcup_{k=n}^{\infty} A_k\Big) = 1$$

as desired. □

Lemma 4.14 (Kolmogorov inequality)**.** *Let $\{X_n\}_{n\geqslant 1}$ be independent random variables, with $\mathbb{E}X_n = 0$ and $\operatorname{Var} X_n < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then for any $\varepsilon > 0$, we have*

$$\mathbb{P}\{\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\} \leqslant \frac{1}{\varepsilon^2}\sum_{k=1}^{n} \mathbb{E}X_k^2 \,.$$

Proof. Let us first decompose

$$\bigcup_{k=1}^{n} A_k := \bigcup_{k=1}^{n} \{|S_k| \geqslant \varepsilon, |S_j| < \varepsilon \textit{ for all } j < k\} = \{\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\}.$$

Then

$$\begin{aligned} \mathbb{E}S_n^2 &\geqslant \sum_{k=1}^{n} \mathbb{E}[S_n^2 \mathbf{1}_{A_k}] = \sum_{k=1}^{n} \mathbb{E}[(S_k + S_n - S_k)^2 \mathbf{1}_{A_k}] \\ &\geqslant \sum_{k=1}^{n} \mathbb{E}[S_k^2 \mathbf{1}_{A_k}] + \sum_{k=1}^{n} 2\mathbb{E}[S_k(S_n - S_k)\mathbf{1}_{A_k}] = \sum_{k=1}^{n} \mathbb{E}[S_k^2 \mathbf{1}_{A_k}] \\ &\geqslant \sum_{k=1}^{n} \varepsilon^2 \mathbb{P}(A_k) = \varepsilon^2 \mathbb{P}\{\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\}\,. \end{aligned} \tag{4.4}$$

Here in the fourth step we used $\mathbb{E}[S_k(S_n-S_k)\mathbf{1}_{A_k}] = \mathbb{E}[S_k\mathbf{1}_{A_k}]\mathbb{E}[S_n-S_k] = 0$. As $\mathbb{E}S_n^2 = \sum_{k=1}^{n} \mathbb{E}X_k^2$, (4.4) yields desired result. □

The Kolmogorov inequality implies the following.

Theorem 4.15. *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of independent random variables, with $\mathbb{E}X_n = 0$ and $\sum_{n=1}^{\infty} \mathbb{E}X_n^2 < \infty$. Then $\sum_{n=1}^{\infty} X_n$ converges a.s.*

Proof. We will show that S_n is a Cauchy sequence a.s. For $N \geqslant 1$, define

$$W_N := \sup_{m,n\geqslant N} |S_m - S_n|\,.$$

As W_N is nonnegative and non-increasing, we have $W_N \downarrow W_\infty$ as $N \to \infty$. It suffice to show that $W_\infty = 0$ a.s. For $\varepsilon > 0$, since

$$\sup_{m\geqslant N} |S_m - S_N| \leqslant \varepsilon \implies \sup_{m,n\geqslant N} |S_m - S_n| \leqslant 2\varepsilon\,,$$

we have

$$\mathbb{P}(W_N > 2\varepsilon) \leqslant \mathbb{P}(\sup_{m\geqslant N} |S_m - S_N| > \varepsilon) = \lim_{M\to\infty} \mathbb{P}(\sup_{N\leqslant m\leqslant M} |S_m - S_N| > \varepsilon)\,.$$

Now by Lemma 4.14 (think about why)

$$\mathbb{P}(\sup_{N\leqslant m\leqslant M}|S_m-S_N|>\varepsilon)\leqslant\frac{1}{\varepsilon^2}\sum_{k=N+1}^{M}\mathbb{E}X_k^2\leqslant\frac{1}{\varepsilon^2}\sum_{k=N+1}^{\infty}\mathbb{E}X_k^2\,.$$

Combing the above two relations yields

$$\mathbb{P}(W_\infty>2\varepsilon)\leqslant\mathbb{P}(W_N>2\varepsilon)\leqslant\frac{1}{\varepsilon^2}\sum_{k=N+1}^{\infty}\mathbb{E}X_k^2$$

for any $\varepsilon>0$ and $N\geqslant 1$. Taking $N\to\infty$ and then take $\varepsilon\to 0$ we get $\mathbb{P}(W_\infty>0)\leqslant 0$. This finishes the proof. □

Now let us prove the "if" part of Theorem 4.10; the proof of the "only if" part is more technical and we omit here. You may refer to [11] if interested.

Suppose conditions (i) – (iii) in Theorem 4.10 hold. Define $Y_n:=X_n^c-\mathbb{E}X_n^c$. Then $\{Y_n\}_{n\geqslant 1}$ are independent, centered, and $\sum_{n=1}^\infty\mathbb{E}Y_n^2=\sum_{n=1}^\infty\operatorname{Var}X_n^c<\infty$. Thus Theorem 4.15 implies $\sum_{n=1}^\infty Y_n$ converges a.s. Together with (ii) we know that

$$\sum_{n=1}^{\infty}X_n^c=\sum_{n=1}^{\infty}Y_n+\sum_{n=1}^{\infty}\mathbb{E}X_n^c \tag{4.5}$$

converges a.s. Thanks to the first lemma of Borel-Cantelli and Condition (i), we have

$$\mathbb{P}(\liminf_n\{|X_n|\leqslant c\})=1-\mathbb{P}(\limsup_n\{|X_n|>c\})=1\,.$$

On the set of $\liminf_n\{|X_n|\leqslant c\}$, it holds that

$$\sum_{n=1}^{\infty}X_n(\omega)\ \textit{converges}\quad\Longleftrightarrow\quad\sum_{n=1}^{\infty}X_n^c\ \textit{converges}\,. \tag{4.6}$$

Hence (4.5) and (4.6) imply $\sum_{n=1}^\infty X_n$ converges a.s. This finishes the proof.

4.3. The law of iterated logarithm

Once again, let us go back to our classic model. Let $\{X_n\}$ be a sequence of i.i.d. random variables, with $\mathbb{E}X_1=0$ and $\mathbb{E}X_1^2=1$. Let $S_n=X_1+\cdots+X_n$. By Theorems 3.1 and 4.3, we know that

$$\frac{S_n}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,1)\quad\text{and}\quad\frac{S_n}{n}\xrightarrow{a.s.}0\,. \tag{4.7}$$

The LLN tells us that n grows too quickly relative to S_n to retain any useful information about the deviation as $n\to\infty$. CLT does a better job, by showing that $S_n/\sqrt{n}$ converges to a non-trivial distribution. However, CLT does not tells us what happens for any *particular outcome* ω. In fact, for any $M>0$, we have

$$\mathbb{P}\Big(\limsup_n\frac{S_n}{\sqrt{n}}>M\Big)=\lim_n\mathbb{P}\Big(\sup_{k\geqslant n}\frac{S_k}{\sqrt{k}}>M\Big)\geqslant\lim_n\sup_{k\geqslant n}\mathbb{P}\Big(\frac{S_k}{\sqrt{k}}>M\Big)=\lim_n\mathbb{P}\Big(\frac{S_n}{\sqrt{n}}>M\Big)>0\,,$$

where in the last step we used the CLT. By Lemma 4.5, we have

$$\mathbb{P}\Big(\limsup_n\frac{S_n}{\sqrt{n}}>M\Big)=1$$

for any $M > 0$. Thus

$$\mathbb{P}\Big(\limsup_n \frac{S_n}{\sqrt{n}} = \infty\Big) = 1\,, \quad \text{and analogusly} \quad \mathbb{P}\Big(\liminf_n \frac{S_n}{\sqrt{n}} = -\infty\Big) = 1\,.$$

In particular, $S_n/\sqrt{n}$ diverges a.s.

We hope to find a function $f(n)$ with $\sqrt{n} \ll f(n) \ll n$ such that we can say something stronger about the convergence of $S_n/f(n)$. By Theorem 4.8, we have

$$\frac{S_n}{n^{1/2+\varepsilon}} \xrightarrow{a.s.} 0$$

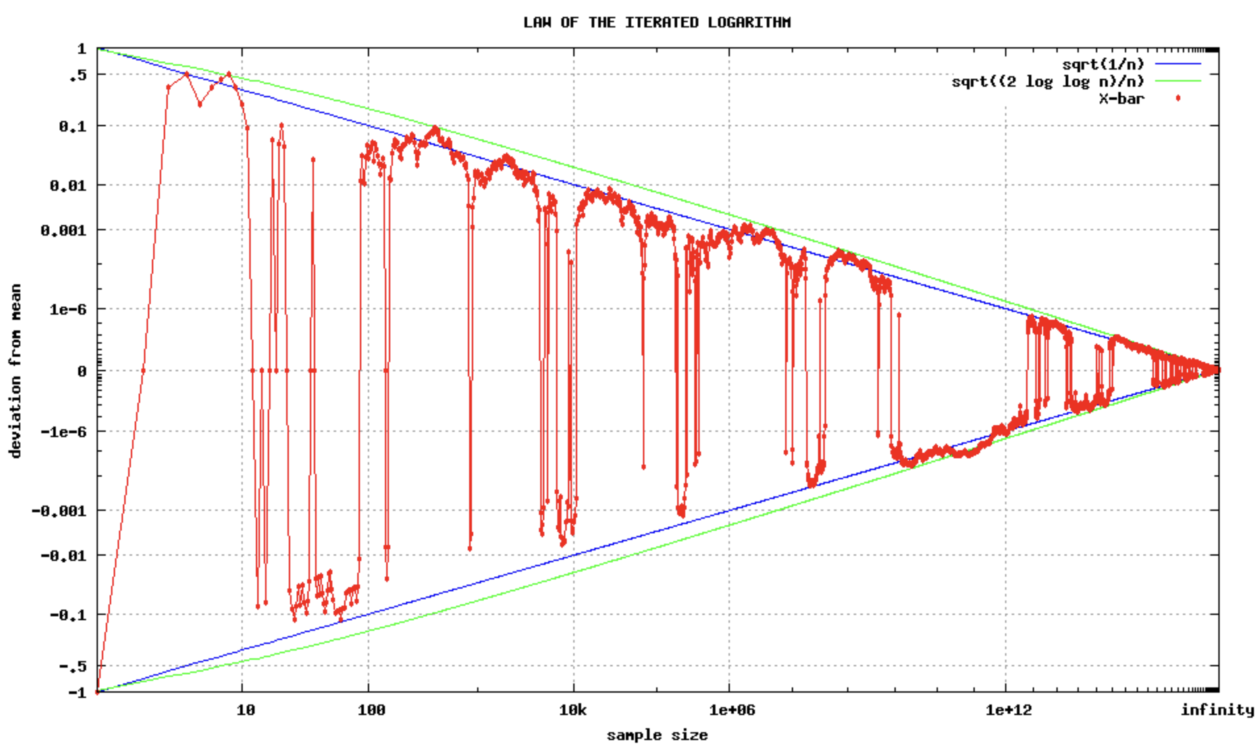
for any fixed $\varepsilon > 0$. The following law of iterated logarithm (LRL) gives a precise scaling for S_n.

Theorem 4.16 (Hartman-Winter)**.** *Let $\{X_n\}$ be a sequence of i.i.d. random variables, with $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = 1$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\limsup_n \frac{S_n}{\sqrt{2n \log\log n}} = 1$$

almost surely.

The following figure is a plot of S_n/n (red), its standard deviation $1/\sqrt{n}$ (blue), and its bound $\sqrt{2\log\log n/n}$ given by LIL (green). Notice the way it randomly switches from the upper bound to the lower bound. Both axes are log-transformed.



Remark 4.17. (i) By symmetry, we also have

$$\liminf_n \frac{S_n}{\sqrt{2n\log\log n}} = -1$$

almost surely.

(ii) Comparing Theorems 3.1 and 4.16, we see that LRL provides the scaling factor where the two limits become different:

$$\frac{S_n}{\sqrt{2n\log\log n}} \xrightarrow{\mathbb{P}} 0\,, \quad \text{while} \quad \frac{S_n}{\sqrt{2n\log\log n}} \overset{a.s.}{\not\to} 0\,.$$

In other words, for any given large n, $|S_n/\sqrt{2n\log\log n}\,|$ is less than any predefined $\varepsilon > 0$ for almost all outcomes ω; on the other hand, for almost all ω, $S_n/\sqrt{2n\log\log n}$ will be visiting the neighborhoods of any point in the interval $(-1,1)$.

Proof of Theorem 4.16: the Gaussian case. Let us further assume that $X_i \overset{d}{=} \mathcal{N}(0,1)$. In this case, $S_n/\sqrt{n} \overset{d}{=} \mathcal{N}(0,1)$ for all n. By Question 4 in Homework 2, we know that

$$\frac{1}{\sqrt{2\pi}}(x^{-1} - x^{-3})\mathrm{e}^{-x^2/2} \leqslant \mathbb{P}(S_n/\sqrt{n} \geqslant x) \leqslant \frac{1}{x\sqrt{2\pi}}\mathrm{e}^{-x^2/2} \tag{4.8}$$

for all $x > 0$.

(i) Fix $\varepsilon > 0$. Then (4.8) implies

$$\begin{aligned} &\mathbb{P}\Big(\frac{S_n}{\sqrt{2n\log\log n}} > 1+\varepsilon\Big) = \mathbb{P}\big(S_n/\sqrt{n} > (1+\varepsilon)\sqrt{2\log\log n}\,\big) \\ \leqslant\ & \frac{1}{(1+\varepsilon)\sqrt{4\pi\log\log n}}\exp(-(1+\varepsilon)^2\log\log n) \leqslant (\log n)^{-(1+\varepsilon)^2} \end{aligned} \tag{4.9}$$

for $n \geqslant 2$. Choose $q > 1$ such that

$$\varepsilon^2/(q-1) > 2\,. \tag{4.10}$$

Choose the subsequence $n_k = \lfloor q^k \rfloor$. We can find an $M \equiv M(\varepsilon) \in \mathbb{N}$ such that $n_M \geqslant 2$. This leads to

$$\begin{aligned} &\sum_{k=M}^{\infty} \mathbb{P}\Big(\frac{S_{n_k}}{\sqrt{2n_k\log\log n_k}} > 1+\varepsilon\Big) \leqslant \sum_{k=M}^{\infty}(\log n_k)^{-(1+\varepsilon)^2} \\ =\ & \sum_{l=1}^{\infty}\sum_{k=lM}^{(l+1)M-1}(\log n_k)^{-(1+\varepsilon)^2} \leqslant \sum_{l=1}^{\infty} M(l\log 2)^{-(1+\varepsilon)^2} < \infty. \end{aligned}$$

Then Lemma 4.13 (i) implies

$$\mathbb{P}\Big(\limsup_k \Big\{\frac{S_{n_k}}{\sqrt{2n_k\log\log n_k}} > 1+\varepsilon\Big\}\Big) = 0\,,$$

which further shows (why?)

$$\mathbb{P}\Big(\limsup_k \frac{S_{n_k}}{\sqrt{2n_k\log\log n_k}} > 1+\varepsilon\Big) = 0\,. \tag{4.11}$$

To handle the terms between each n_k, note that we have the reflection principle

$$\mathbb{P}\Big(\max_{1\leqslant m\leqslant n} S_n > a\Big) \leqslant 2\mathbb{P}(S_m > a) \tag{4.12}$$

for all $a \in \mathbb{R}$ and $n \geqslant 1$. To prove (4.12), define the stopping time

$$\tau := \min\{m : S_m > a\}.$$

Thus $1 \geqslant \sum_{m=1}^n \mathbf{1}_{\tau=m}$, which leads to

$$\begin{aligned}
\mathbb{P}(S_n > a) &\geqslant \sum_{m=1}^n \mathbb{P}(S_n > a, \tau = m) = \sum_{m=1}^n \mathbb{P}(S_n > a | \tau = m) \cdot \mathbb{P}(\tau = m) \\
&\geqslant \sum_{m=1}^n \frac{1}{2}\mathbb{P}(\tau = m) = \frac{1}{2}\mathbb{P}(\tau \leqslant n) = \frac{1}{2}\mathbb{P}\Big(\max_{1\leqslant m\leqslant n} S_m > a\Big),
\end{aligned}$$

where in the third step we used $\{X_n\}_{n\geqslant 1}$ are independent and symmetric (w.r.t. 0). Applying (4.12) yields

$$\begin{aligned}
&\mathbb{P}\Big(\max_{n_k\leqslant m\leqslant n_{k+1}} (S_m - S_{n_k}) > \varepsilon\sqrt{2n_k \log\log n_k}\Big) \\
&= \mathbb{P}\Big(\max_{m\leqslant n_{k+1}-n_k} S_m > \varepsilon\sqrt{2n_k \log\log n_k}\Big) \leqslant 2\,\mathbb{P}\Big(S_{n_{k+1}-n_k} > \varepsilon\sqrt{2n_k \log\log n_k}\Big) \\
&\leqslant 2\,\mathbb{P}\bigg(\frac{S_{n_{k+1}-n_k}}{\sqrt{n_{k+1}-n_k}} > \varepsilon\sqrt{2\log\log n_k/(q-1)}\bigg) \leqslant 2(\log n_k)^{-\varepsilon^2/(q-1)},
\end{aligned}$$

where the last step is similar to (4.9). Note that (4.10) guarantees

$$\sum_{k=M}^{\infty} \mathbb{P}\Big(\max_{n_k\leqslant m\leqslant n_{k+1}} (S_m - S_{n_k}) > \varepsilon\sqrt{2n_k \log\log n_k}\Big) < \infty,$$

and together with Lemma 4.13 (i) we get

$$\mathbb{P}\bigg(\limsup_k \max_{n_k\leqslant m\leqslant n_{k+1}} \frac{S_m - S_{n_k}}{\sqrt{2n_k \log\log n_k}} > \varepsilon\bigg) = 0. \tag{4.13}$$

Combining (4.11) and (4.13) yields

$$\mathbb{P}\bigg(\limsup_n \frac{S_n}{\sqrt{2n\log\log n}} > 1 + 2\varepsilon\bigg) = 0 \tag{4.14}$$

for any fixed $\varepsilon > 0$.

(ii) Fix $\delta > 0$, and let $p \in \{2, 3, ...\}$ such that

$$\frac{(1-\delta)p}{p-1} \leqslant 1 \quad \text{and} \quad \frac{2}{\sqrt{p}} \leqslant \delta. \tag{4.15}$$

Consider the subsequence $n_k = p^k$ and define the event

$$A_k := \{S_{n_{k+1}} - S_{n_k} > (1-\delta)\sqrt{2n_{k+1}\log\log n_{k+1}}\} = \bigg\{\frac{S_{n_{k+1}} - S_{n_k}}{\sqrt{n_{k+1}-n_k}} > (1-\delta)\sqrt{2p\log\log n_{k+1}/(p-1)}\bigg\}.$$

By the first relation of (4.8), we have

$$\mathbb{P}(A_k) \geqslant \frac{c_{p,\delta}}{\sqrt{\log\log n_{k+1}}}(\log n_{k+1})^{-(1-\delta)^2 p/(1-p)} \geqslant \frac{c_{p,\delta}}{\sqrt{\log\log n_{k+1}}}(\log n_{k+1})^{-(1-\delta)} \geqslant \frac{c'_{p,\delta}}{\log k} k^{-(1-\delta)}$$

for $k \geqslant 1$. Here $c_{p,\delta}, c'_{p,\delta} > 0$ are constants. This leads to

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \infty\,.$$

As $\{A_k\}$ are independent events, Lemma 4.13 (ii) yields

$$\mathbb{P}(\limsup_k A_k) = 1\,,$$

which yields

$$\mathbb{P}\Big(\limsup_k \frac{S_{n_{k+1}} - S_{n_k}}{\sqrt{2n_{k+1}\log\log n_{k+1}}} > 1-\delta\Big) = 1 \tag{4.16}$$

Note that we can repeat the proof of (4.14) for $\{-X_n\}$ and $\{-S_n\}$ to show that

$$\mathbb{P}\Big(\limsup_n \frac{-S_n}{\sqrt{2n\log\log n}} \geqslant 2\Big) = 0\,,$$

and thus

$$\mathbb{P}\Big(\limsup_n \frac{-S_n}{\sqrt{2n\log\log n}} < 2\Big) = 1\,,$$

which implies

$$1 = \mathbb{P}\Big(\limsup_k \frac{-S_{n_k}}{\sqrt{2n_k\log\log n_k}} < 2\Big) \leqslant \mathbb{P}\Big(\limsup_k \frac{-S_{n_k}}{\sqrt{2n_{k+1}\log\log n_{k+1}}} < \frac{2}{\sqrt{p}}\Big)\,.$$

Together with the second relation of (4.15) we have

$$\mathbb{P}\Big(\liminf_k \frac{S_{n_k}}{\sqrt{2n_{k+1}\log\log n_{k+1}}} > -\delta\Big) = 1\,. \tag{4.17}$$

Combining (4.16) and (4.17) yields

$$\mathbb{P}\Big(\limsup_n \frac{S_n}{\sqrt{2n\log\log n}} > 1-2\delta\Big) \geqslant \mathbb{P}\Big(\limsup_k \frac{S_{n_{k+1}}}{\sqrt{2n_{k+1}\log\log n_{k+1}}} > 1-2\delta\Big) = 1\,.$$

Together with (4.14) we conclude the proof. □

Under the assumption of Theorem 4.16, we can also show that

$$\limsup_n \frac{\max_{1\leqslant k\leqslant n}|S_k|}{\sqrt{2n\log\log n}} = 1 \tag{4.18}$$

almost surely. Comparing Theorem 4.16 and (4.18), we see that S_n and $\max_{1\leqslant k\leqslant n}|S_k|$ have the same asymptotic behavior in terms of lim sup. For lim inf, we have the following result.

Theorem 4.18 (Chung)**.** *Let $\{X_n\}$ be a sequence of i.i.d. random variables, with $\mathbb{E}X_1 = 0$, $\mathbb{E}X_1^2 = 1$ and $\mathbb{E}|X_1^3| < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\liminf_n \sqrt{\frac{8\log\log n}{n\pi^2}} \max_{1\leqslant k\leqslant n}|S_k| = 1$$

almost surely.

Homework 3

Question 1. Let $\{X_n\}n \geqslant 1$ be i.i.d. random variables, with the uniform distribute on $[0, 1]$. Let $Y_n := (\prod_{i=1}^n X_i)^{1/n} \geqslant 0$. Find a constant a such that

$$Y_n \xrightarrow{\mathbb{P}} a\,.$$

Question 2. Let f, g be L_1 functions from $[0, 1]$ to $(0, \infty)$. Suppose there exists $C > 0$ such that $f(x) < Cg(x)$ for all x. Show that

$$\lim_n \int_0^1 \cdots \int_0^1 \frac{f(x_1) + \cdots + f(x_n)}{g(x_1) + \cdots + g(x_n)} \mathrm{d}x_1 \cdots \mathrm{d}x_n = \frac{\int_0^1 f(x)\mathrm{d}x}{\int_0^1 g(x)\mathrm{d}x}\,.$$

Question 3. Let $\{X_n\}_{n\geqslant 1}$ be i.i.d. standard Gaussian random variables. Find a function $f(n)$ such that

$$\limsup_n \frac{X_n}{f(n)} = 1$$

almost surely.

Question 4. Let $\{X_n\}_{n\geqslant 1}$ be pairwise independent random variables, having symmetric distributions and satisfy $\sup_n \mathbb{E}X_n^2 < \infty$. Show that

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{a.s.} 0\,.$$

Question 5. Let $\{X_n\}_{n\geqslant 1}$ be i.i.d. random variables with $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. Show that

$$\liminf_n \frac{|S_n|}{\sqrt{n}} = 0$$

almost surely.

Question 6. Let $\{X_n\}_{n\geqslant 1}$ be i.i.d. symmetric random variables with $\mathbb{E}X_1^2 = \infty$. Let $S_n = X_1 + \cdots + X_n$. Show that

$$\limsup_n \frac{|S_n|}{\sqrt{n \log\log n}} = \infty$$

almost surely.

5 Some stories

From now on, let us move on to the topic of random matrix theory. We start with some motivations.

In 1999, a Czech physicist named Petr Šeba took his first trip to Mexico. While riding a bus in a small town, he noticed something very interesting: at every bus stop, a boy would come up to the driver and hand him a piece of paper. The driver would look at the paper and then give the boy some money.

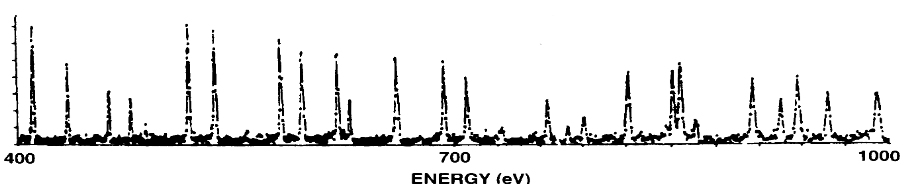
At first, Šeba thought this might be related to the mafia or drug dealers. But then he realized that what was written on the paper was simply the departure time of the previous bus. Because the buses in Mexico were independently operated for profit, the drivers needed to know the status of other buses in order to maximize their earnings. If the previous bus had just left, the driver would slow down to allow more passengers to accumulate at the next stop. But if the previous bus had left a long time ago, the driver would speed up, since there would already be enough people waiting at the next stop. By doing this, he could also avoid being caught by the bus behind him.

Because of the presence of these boys, behind the simple event of bus departures was actually the interaction of all the drivers in the town. Šeba found this very interesting and thought it might relate to something he had seen in physics. So he paid one of the boys for the data he had collected on bus departure times, and Šeba plotted them on a computer.

Now, please look at the image in this link (I cannot include it in this file for copyright reasons), and think about which of the three patterns corresponds to the bus departure times.

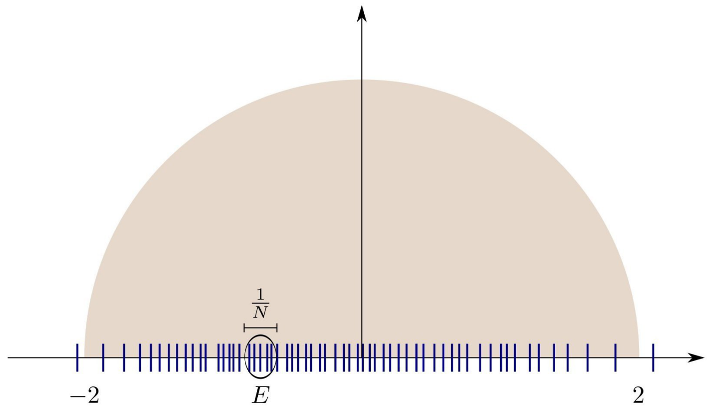
Clearly, the last pattern is deterministic, so it is not our answer. The first pattern is a typical realization of a Poisson process, which means that the events are independent of each other. More generally, this also corresponds to what we call a *weakly correlated system*, meaning that if there are correlations, they are relatively weak, this is what we would see. For instance, if we look at the statistics of earthquakes in an area or the floods of a river, this is what we observe.

The second pattern corresponds to what happened with the buses in Mexico. We see that it walks the fine line between being completely fixed and completely chaotic. This is what we observe when we look at *strongly correlated systems*.



Šeba thought this was very interesting because it related to something he had seen in physics before. So let us go back a bit in time. In 1955, physicists had experimental data on heavy nuclei and wanted to understand why the pattern they observed looked the way it did. But theoretically, this was a very difficult question: the structure of heavy nuclei is extremely complicated, and there are many interactions occurring within a single atom, including certain relativistic effects. Thus it was impossible to determine the entire structure of a heavy nucleus and compute its energy levels explicitly (as could be done for hydrogen, for instance).

Then the physicist Eugene Wigner came forward and suggested that the question could be approached from another perspective: while it was not possible to compute the energy levels directly, statistically, this pattern could be modeled by the eigenvalues of random matrices—that is, matrices with random coefficients.



The model Wigner considered back then is now known as the Wigner matrix.

Definition 5.1. Let $H = H^* = (H_{ij} : 1 \leqslant i, j \leqslant N) \in \mathbb{C}^{N \times N}$ be a $N \times N$ Hermitian matrix. We say H is a Wigner matrix, if $H_{ij} (i \leqslant j)$ are independent random variables satisfying

$$\mathbb{E} H_{ij} = 0 \,, \quad \mathbb{E}|H_{ij}^2| = N^{-1}(1 + O(\delta_{ij})) \,, \quad \text{and} \quad \mathbb{E}|H_{ij}^k| \leqslant C_k N^{-k/2}$$

for all $k \geqslant 3$. Here $C_k > 0$ are constants.

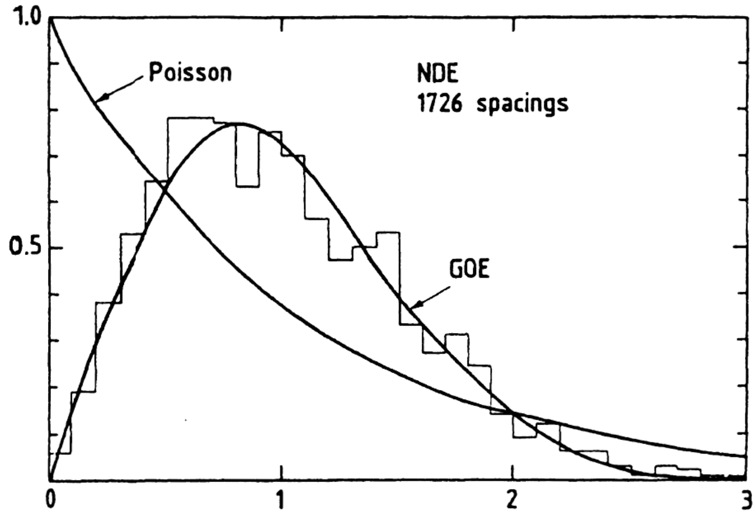
If H_{ij} are real Gaussian random variables, H is the Gaussian Orthogonal Ensemble (GOE). If $H_{ij} (i < j)$ are complex Gaussian random variables and H_{ii} are real Gaussian random variables, H is the Gaussian Unitary Ensemble (GUE).

In the random matrix setting, N is our fundamental large parameter, and we are interested in the properties of the model when N gets large. The famous Wigner's semicircle law asserts that the empirical eigenvalue density converges to the semicircle law when $N \to \infty$.

As far as our heavy nuclei (and the buses in Mexico) are concerned, the more interesting object is the gap distribution between individual eigenvalues. For GOE and GUE, the joint eigenvalue density is known, and the gap distribution can be explicitly computed. Roughly speaking, for GOE, it is given by the Wigner surmise

$$\frac{\pi s}{2} \mathrm{e}^{-\pi s^2/4} \,.$$

People in 1955 could already see that the patterns of the heavy nuclei look very similar to the gap distribution of the GOE, but they could not verify this statistically: every nucleus only has a few dozens of energy levels, and this number is not enough in the statistical sense.



In 1982, Bohigas, Haq and Pandey collected the experimental data of 27 different nuclei, and got more than 1000 data points. This is the nuclear date ensemble (NDE). If we compare the gap histogram of NDE and the gap distribution of GOE (which is computable), we can see that they match quite well, even towards the tail. As a comparison, if we look at the gap of the Poisson process, which has exponential decay, it does not match the gap histogram of the NDE at all.

Let me tell one more example that relates to random matrices. The Riemann Zeta function is defined as

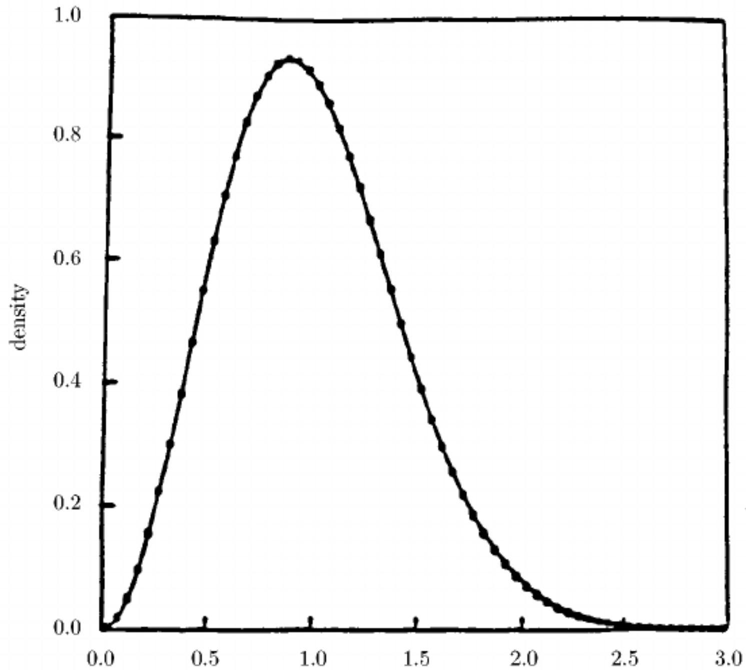
$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

for all $\mathrm{Re}\, s > 1$, and it can be analytically extended to $\mathbb{C}\backslash 0$. We know that ζ has trivial zeros at $-2, -4, -6, \ldots$. The most famous and important open problem in mathematics is the Riemann Hypothesis, which conjectures that all the non-trivial zeros of the zeta function lies on the critical line $s = 1/2 + \mathrm{i}\gamma$. Apparently, no one knows how to prove the Riemann Hypothesis, but it has been checked numerically, that for the first 10^{13} zeros, the RH is true.

So what does this have to do with random matrix theory? The thing is, we can consider the imaginary part of the nth nontrivial zeros of the zeta function, scale it by a elementary function, and consider the gap of the rescaled zeros.

$$\gamma(1/2 + \mathrm{i}\gamma_n) = 0\,, \quad \widehat{\gamma}_n = \frac{1}{2\pi}\gamma_n \log \gamma_n\,, \quad \delta_n = \widehat{\gamma}_{n+1} - \widehat{\gamma}_n\,.$$

What people found out in 1999 is, if one goes very high up in the complex plane, then the density of the rescaled gaps matches perfectly with the gap density of GUE.



δ_n among 70 million zeroes beyond the 10^{20}th zero of ζ, verses GUE.

So now we see that things from totally different worlds, the buses in Mexico, the heavy nuclei, the Riemann Zeta function, all have something to do with random matrix theory. The reason is quite simple: systems with enough complexity and randomness always show the same pattern. We study random matrix theory, partially because it is one of the mathematically simplest complex systems.

6 Local semicircle law for Wigner matrices

6.1. Global and local law

In Section 6, we consider the $N \times N$ Wigner matrix H as in Definition 5.1.

Convention. In random matrix theory, we use N as our fundamental large parameter. We are interested in the properties of H when N gets large. Most quantities we study shall depend on N, thus we almost always omit the explicit argument N from our notation. Quantities are independent of N will be called a constant (or sometimes fixed), and they are usually denoted by C, c. Every quantity that is not explicit a constant is in fact a sequence indexed by N.

From Definition 5.1, we see that for fixed $k \geqslant 3$,

$$\|H_{ij}\|_2 \leqslant \|H_{ij}\|_k \leqslant (C_k)^{1/k} N^{-1/2} \leqslant \widetilde{C}_k \|H_{ij}\|_2 \,.$$

This means the entries of a Wigner matrix are light-tailed.

Let $\lambda_1 \geqslant \cdots \geqslant \lambda_N$ be the eigenvalues of H, and we denote the empirical eigenvalue density by

$$\mu(x) := \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i}(x) \,.$$

Here δ_{λ_i} is the Dirac measure, meaning

$$\delta_{\lambda_i}(x) = \begin{cases} 0 & \text{if} \quad x \neq \lambda_i \\ \infty & \text{if} \quad x = \lambda_i \,, \end{cases} \quad \text{and} \quad \int_{\mathbb{R}} \delta_{\lambda_i}(x) \mathrm{d}x = 1 \,.$$

Let $\varrho_{sc} := \frac{1}{2\pi}\sqrt{(4-X^2)_+}$ denotes the semicircle density, and the famous Wigner's semicircle law asserts that

$$\mu \xrightarrow{w} \varrho_{sc} \tag{6.1}$$

almost surely. The above convergence holds in two layers: the almost sure convergence is w.r.t. the randomness of μ, and "$\xrightarrow{w}$" denotes the weak convergence of probability measures. Equivalently, we can write that for any bounded continuous function f,

$$\int_{\mathbb{R}} f(x)\mu(x)\mathrm{d}x \to \int_{\mathbb{R}} f(x)\varrho_{sc}(x)\mathrm{d}x$$

almost surely.

Another way to formulate the convergence (6.1) is through Stieltjes transforms. For the rest of Section 6, we write

$$z = E + \mathrm{i}\eta \,, \quad \eta > 0 \,.$$

The Stieltjes transforms of μ and ϱ_{sc} are given by

$$s(z) := \int_{\mathbb{R}} \frac{\mu(x)}{x-z} \mathrm{d}x \quad \text{and} \quad m_{sc}(z) := \int_{\mathbb{R}} \frac{\varrho_{sc}(x)}{x-z} \mathrm{d}x \,.$$

They are well-defined as $\eta > 0$. The following is a standard result on Stieltjes transforms.

Lemma 6.1. *We have $\mu \xrightarrow{w} \varrho_{sc}$ almost surely, if and only if $s(x) \to m_{sc}(z)$ almost surely for any fixed $z \in \mathbb{C}_+$.*

Note that

$$s(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i - z} = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i - E + \mathrm{i}\eta}{(\lambda_i - E)^2 + \eta^2}\,, \quad \operatorname{Im} s(z) = \frac{1}{N\eta}\sum_{i=1}^{N}\frac{1}{\left(\frac{\lambda_i - E}{\eta}\right)^2 + 1}\,.$$

Thus $\operatorname{Im} s(z)$ is a control of the empirical distribution of μ smoothed on the scale η. For this reason, η is called the *spectral resolution*.

Theorem 6.2. *We have $s(x) \to m_{sc}(z)$ almost surely for any fixed $z \in \mathbb{C}_+$.*

In Theorem 6.2, z is N-independent, which means we are on the spectral resolution of order 1. Theorem 6.2 is therefore called the *global semicircle law*.

A *local law* is a result that controls $s(z) - m_{sc}(z)$ for all η satisfying $\eta \gg N^{-1}$. In other words, η depends on N. The restriction $\eta \gg N^{-1}$ is obvious, since the N eigenvalues lies roughly on $[-2, 2]$, and the typical separation of the eigenvalues is of order N^{-1}.

6.2. Statements and consequences of the local law

In order to state the local law, we shall need the following notion.

Definition 6.3 (Stochastic domination)**.** Let

$$X = (X^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)})\,, \quad Y = (Y^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)})$$

be two families of random variables, where $U^{(N)}$ is a possibly N-dependent parameter set, and $Y \geqslant 0$. We say that X is stochastically dominated by Y , uniformly in u, if for any fixed $D, \varepsilon > 0$,

$$\sup_{u \in U^{(N)}} \mathbb{P}(|X| \geqslant Y N^{\varepsilon}) = O_{\varepsilon,D}(N^{-D})\,.$$

If X is stochastically dominated by Y , we use the notation $X = O_{\prec}(Y)$, or equivalently $X \prec Y$. We say an event Ω holds with very high probability if for any fixed $D > 0$, $1 - \mathbb{P}(\Omega) = O_D(N^{-D})$.

Remark 6.4. (i) Stochastic domination provides a good measurement on the scale of a random variable. For instance, consider the random variable $Y \overset{d}{=} \mathcal{N}(0,1)$, whose typical size is of order 1. But we cannot find C such that $|Y| \leqslant C$ almost surely. However, we can write $Y \prec 1$.

(ii) In practice to prove $X \prec Y$, we show that $\mathbb{E}|X^p| \leqslant C_p Y^p$ for any fixed p. Then given $\varepsilon, D > 0$, choose p large enough such that

$$\mathbb{P}(|X| \geqslant N^{\varepsilon} Y) \leqslant C_p N^{-\varepsilon p} \leqslant N^{-D}\,.$$

As the local law is stated via Stieltjes transform, we introduce the *Green function*

$$G(z) := (H - z)^{-1}\,.$$

We denote the normalized trace $\underline{G}(z) := \frac{1}{N}\operatorname{Tr} G(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i - z} = s(z)$. The local law controls $\underline{G}$ as well as G_{ij}. Fix $\tau > 0$, and define the spectral domain

$$\mathbf{S} \equiv \mathbf{S}_{\tau}(N) := \{z = E + \mathrm{i}\eta : |E| \leqslant 10, N^{-1+\tau} \leqslant \eta \leqslant 10\}\,. \tag{6.2}$$

The next result is our main target in Section 6.

Theorem 6.5 (Local semicircle law)**.** *We have*

$$\underline{G}(z) - m_{sc}(z) \prec \frac{1}{N\eta} \tag{6.3}$$

and

$$\max_{i,j} |G_{ij}(z) - \delta_{ij} m_{sc}(z)| \prec \frac{1}{N\eta} + \sqrt{\frac{\operatorname{Im} m_{sc}(z)}{N\eta}} \tag{6.4}$$

uniformly for all $z \in \mathbf{S}$.

The local law has several consequences, which we will prove in later parts of Section 6.

Corollary 6.6. *We have*

$$\text{number of eigenvalues in } [a,b] = N \int_a^b \mu(x)\mathrm{d}x = N \int_a^b \varrho_{sc}(x)\mathrm{d}x + O_{\prec}(1)$$

uniformly for all $[a,b] \subset \mathbb{R}$.

For the eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_N$ of H, we denote the typical location of λ_i by γ_i, which is defined via

$$\int_{\gamma_i}^2 \varrho_{sc}(x)\mathrm{d}x = \frac{i}{N}\,.$$

The next result is known as the rigidity of eigenvalues.

Corollary 6.7. *We have*

$$|\lambda_i - \gamma_i| \prec N^{-2/3}(i \wedge N+1-i)^{-1/3}$$

uniformly for all $i = 1, 2, ..., N$.

Note that

$$\gamma_k - \gamma_{k+1} \sim N^{-2/3}$$

for fixed k, and

$$\gamma_i - \gamma_{i+1} \sim N^{-1}$$

for $i \in [cN, (1-c)N]$. In other words, the result of Corollary 6.7 is optimal, subject to the N^{ε} hidden in the definition of "$\prec$".

The next result is about the eigenvectors.

Corollary 6.8 (Delocalization)**.** *Let* $\mathbf{u}_1, ..., \mathbf{u}_N$ *be the eigenvectors of* H *associated with* $\lambda_1 \geqslant \cdots \geqslant \lambda_N$. *Assume the eigenvectors a normalized, i.e.* $\|\mathbf{u}_i\|_2 = 1$ *for all* i. *Then*

$$\max_{1 \leqslant i,k \leqslant N} |\mathbf{u}_i(k)| \prec N^{-1/2}\,.$$

Proof. This is a simple consequence of Theorem 6.5. As H is Hermitian, it is also normal ($HH^* = H^*H$). Thus H and G admit the *spectral decomposition*

$$H = U\Lambda U = \sum_{i=1}^N \lambda_i \mathbf{u}_i \mathbf{u}_i^*\,, \quad \text{and} \quad G(z) = \sum_{i=1}^N \frac{\mathbf{u}_i \mathbf{u}_i^*}{\lambda_i - z}\,.$$

Choosing $z = \lambda_j + \mathrm{i}\eta$, we get

$$\operatorname{Im} G_{kk}(\lambda_j + \mathrm{i}\eta) = \operatorname{Im} \sum_{i=1}^{N} \frac{|u_i(k)|^2}{\lambda_i - (\lambda_j + \mathrm{i}\eta)} = \sum_{i=1}^{N} \frac{\eta}{(\lambda_i - \lambda_j)^2 + \eta^2} |\mathbf{u}_i(k)|^2 \geqslant \frac{1}{\eta} |\mathbf{u}_i(k)|^2 .$$

As $|m_{sc}(z)| < 1$ and

$$\operatorname{Im} G_{kk} \leqslant |G_{kk}| \leqslant |G_{kk} - m_{sc}| + m_{sc} \leqslant O_{\prec}\Big(\sqrt{\frac{1}{N\eta}}\Big) + 1 \prec 1 ,$$

we get

$$|u_i(k)|^2 \prec \eta \leqslant N^{-1+\tau}$$

for any fixed $\tau > 0$. This finishes the proof. □

6.2.1. Outline of Proof. We first give an outline of how to prove Theorem 6.5. It is easy to see that

$$m_{sc}(z) = \int_{\mathbb{R}} \frac{1}{2\pi} \frac{\sqrt{(4 - x^2)_+}}{x - z} \mathrm{d}x = \frac{-z + \sqrt{z^2 - 4}}{2} .$$

Here the branch cut of $\sqrt{\cdot}$ is taken at the positive real axis, meaning

$$\operatorname{Im} \sqrt{z} > 0 \quad \text{for all} \quad z \in \mathbb{C} \backslash \mathbb{R} .$$

We often use that $m_{sc}(z)$ is the unique solution of

$$1 + zm + m^2 = 0$$

satisfying $\operatorname{Im} m(z) > 0$. We show $\underline{G}$ is close to m_{sc} by proving $1 + z\underline{G} + \underline{G}^2 \approx 0$.

Step 1. The probabilistic estimates. Suppose that for some $z \in \mathbf{S}_\tau$, we have $\max_{ij} |G_{ij}(z)| \prec 1 + \phi$ for some $\phi \in [N^{-1}, N^{\tau/10}]$. For any fixed $n \in \mathbb{N}_+$, prove that

$$\mathbb{E}|1 + z\underline{G} + \underline{G}^2|^n = O(\mathcal{E}_1(\phi)^n) \quad \text{and} \quad \mathbb{E}|\delta_{ij} + zG_{ij} + \underline{G}G_{ij}|^n = O(\mathcal{E}_2(\phi)^n) ,$$

and conclude that

$$1 + z\underline{G} + \underline{G}^2 \prec \mathcal{E}_1(\phi) \quad \text{and} \quad \delta_{ij} + zG_{ij} + \underline{G}G_{ij} \prec \mathcal{E}_2(\phi) .$$

Step 1.5. Reflection. The error terms $\mathcal{E}_1(\phi)$ and $\mathcal{E}_2(\phi)$ in Step 1 are small enough if $\phi \leqslant N^{\tau/10}$. However, it is already non-trivial to establish $\max_{ij} |G_{ij}| \prec N^{\tau/10}$. By spectral decomposition, we can only get

$$|G_{ij}(z)| = \Big| \sum_{i=1}^{N} \frac{\mathbf{u}_i(k)\mathbf{u}_i^*(k)}{\lambda_k - (E + \mathrm{i}\eta)} \Big| \leqslant \eta^{-1} \sum_{i=1}^{N} |\mathbf{u}_i(k)\mathbf{u}_i^*(k)| \leqslant \eta^{-1} , \tag{6.5}$$

which is terribly bad if $\eta \ll N^{-\tau/10}$.

Step 2. Initial bound. In our analysis, we consider one particular E, and go down in the spectral resolution η. We first treat the case when $\eta \geqslant 1$, where (6.5) gives $\max_{ij} |G_{ij}(z)| \prec 1$. The estimates $1 + z\underline{G} + \underline{G}^2 \prec \mathcal{E}_1(1)$ and $\delta_{ij} + zG_{ij} + \underline{G}G_{ij} \prec \mathcal{E}_2(1)$ then gives

$$\underline{G} - m_{sc} \ll 1 \quad \text{and} \quad G_{ij} - \delta_{ij} m_{sc} \ll 1 ,$$

which is the global law at $\eta \geqslant 1$.

Step 3. Finally we perform a "bootstrap" argument and extend our results for all $\eta \geqslant N^{-1+\tau}$. Suppose we have the local law at $z_0 = E + \mathrm{i}\eta_0$. Then we know that

$$|G_{ij}(z_0)| \leqslant |m_{sc}(z_0)\delta_{ij}| + |G_{ij}(z_0) - \delta_{ij} m_{sc}(z_0)| \prec 1\,.$$

To reach smaller η, we have the following result.

Lemma 6.9. *Let* $\Gamma(z) := \max_{ij} |G_{ij}(z)|$*. For any* $M > 1$*, we have*

$$\Gamma(E + \mathrm{i}\eta/M) \leqslant M\Gamma(E + \mathrm{i}\eta)\,.$$

From Lemma 6.9, we know that for all $\eta_1 \in [\eta_0 N^{-\tau/10}, \eta_0)$,

$$\Gamma(E + \mathrm{i}\eta_1) \leqslant \frac{\eta_0}{\eta_1}\Gamma(E + \mathrm{i}\eta_0 \prec N^{\varepsilon} \cdot \Gamma(E + \mathrm{i}\eta_0) \prec N^{\tau/10}\,.$$

Thus from Step 1 we know that

$$1 + z\underline{G} + \underline{G}^2 \prec \mathcal{E}_1(N^{\tau/10}) \quad \text{and} \quad \delta_{ij} + zG_{ij} + \underline{G}G_{ij} \prec \mathcal{E}_2(N^{\tau/10})\,.$$

at $z = E + \mathrm{i}\eta_1$. This will imply local law at $E + \mathrm{i}\eta_1$, and allows us to repeat step 3.

6.3. Proof of Theorem 6.5: probabilistic part

We shall work on the case when $H = H^T$ is real and symmetric, which makes H_{ij} real random variables. The complex Hermitian case works in a similar fashion. The main probabilistic estimate is the following result.

Proposition 6.10. *Let us define*

$$\Pi(G) := 1 + zG + \underline{G}G \in \mathbb{C}^{N\times N}\,.$$

Let $z \in \mathbf{S}_\tau$ *and suppose that* $\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \phi$ *for some deterministic* $\phi \in [N^{-1}, N^{\tau/10}]$ *at* z*. Then*

$$\max_{ij} |\Pi(G)_{ij}| \prec (1+\phi)^3 \sqrt{\frac{\operatorname{Im} m_{sc} + \phi}{N\eta}} \tag{6.6}$$

and

$$\underline{\Pi(G)} \prec (1+\phi)^6 \frac{\operatorname{Im} m_{sc} + \phi}{N\eta} \tag{6.7}$$

at z*.*

6.3.1. Cumulant expansion. For a real random variable h, all of whose moments are finite, the kth-cumulant of h is

$$\mathcal{C}_k(h) := (-\mathrm{i})^k \left(\frac{\mathrm{d}^k}{\mathrm{d}t^k} \log \mathbb{E}[e^{\mathrm{i}th}]\right)\bigg|_{t=0}\,.$$

One easily checks that $\mathcal{C}_1(h) = \mathbb{E}h$, $\mathcal{C}_2(h) = \operatorname{Var} h$, $\mathcal{C}_3(h) = \mathbb{E}(h - \mathbb{E}h)^3$. The cumulants have the following properties.

(i) Let $a \in \mathbb{R}$. Since $\log \mathbb{E}\mathrm{e}^{\mathrm{i}t(h+a)} = \mathrm{i}ta + \log \mathbb{E}\mathrm{e}^{\mathrm{i}th}$, we have

$$\mathcal{C}_n(h+a) = (-\mathrm{i})^k \left(\frac{\mathrm{d}^k}{\mathrm{d}t^k} \log \mathbb{E}[e^{\mathrm{i}t(h+a)}]\right)\bigg|_{t=0} = \begin{cases} \mathcal{C}_1(h) + a & \text{if} \quad n = 1 \\ \mathcal{C}_n(h) & \text{if} \quad n \geqslant 2\,. \end{cases}$$

(ii) Let h_1, h_2 be two independent random variables with $\mathbb{E}|h_1^n|, \mathbb{E}|h_2^n| < \infty$. Since

$$\log \mathbb{E}\mathrm{e}^{\mathrm{i}t(h_1+h_2)} = \log[\mathbb{E}\mathrm{e}^{\mathrm{i}th_1} \cdot \mathbb{E}\mathrm{e}^{\mathrm{i}th_2}] = \log \mathbb{E}\mathrm{e}^{\mathrm{i}th_1} + \log \mathbb{E}\mathrm{e}^{\mathrm{i}th_2} \,,$$

we have $\mathcal{C}_n(h_1 + h_2) = \mathcal{C}_n(h_1) + \mathcal{C}_(h_2)$.

(iii) Let $h \overset{d}{=} \mathcal{N}(\mu, \sigma^2)$, then

$$\log \mathbb{E}\mathrm{e}^{\mathrm{i}th} = \log \mathrm{e}^{\mathrm{i}t\mu - \frac{1}{2}\sigma^2 t^2} = \mathrm{i}t\mu - \frac{1}{2}\sigma^2 t^2 \,.$$

Thus $\mathcal{C}_n(h) = 0$ for all $n \geqslant 3$.

(iv) $\mathcal{C}_n(ah) = a^n \mathcal{C}_n(h)$.

Note that (i)–(iii) in the above are in general not true for moments. The main tool that we use in our computation is the following formula by Andrew Barbour [1]. It is a generalization of Lemma 3.4.

Lemma 6.11 (Cumulant expansion)**.** *Let $F : \mathbb{R} \to \mathbb{C}$ be a smooth function, and denote by $F^{(n)}$ its nth derivative. Then, for every fixed $\ell \in \mathbb{N}$, we have*

$$\mathbb{E}\big[h \cdot F(h)\big] = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}[F^{(k)}(h)] + \mathcal{R}_{\ell+1}, \tag{6.8}$$

assuming that all expectations in (6.8) exist, where $\mathcal{R}_{\ell+1}$ is a remainder term (depending on f and h), such that for any $t > 0$,

$$\mathcal{R}_{\ell+1} = O(1) \cdot \bigg(\mathbb{E} \sup_{|x| \leqslant |h|} \big|F^{(\ell+1)}(x)\big|^2 \cdot \mathbb{E}\,\big|h^{2\ell+4} \mathbf{1}_{|h|>t}\big|\bigg)^{1/2} + O(1) \cdot \mathbb{E}|h|^{\ell+2} \cdot \sup_{|x| \leqslant t} \big|F^{(\ell+1)}(x)\big| \,.$$

We start with an elementary inequality.

Lemma 6.12. *Let X be a nonnegative random variable with finite moments. Then for any $a, b, t \geqslant 0$, we have*

$$\mathbb{E}X^a \mathbb{E}\big[X^b \mathbf{1}_{X>t}\big] \leqslant \mathbb{E}\big[X^{a+b} \mathbf{1}_{X>t}\big] \,.$$

Proof. It suffices to assume $a > 0$. Let us abbreviate $\|X\|_a := (\mathbb{E}X^a)^{1/a}$. For $t \geqslant \|X\|_a$, we have

$$\mathbb{E}X^a \mathbb{E}\big[X^b \mathbf{1}_{X>t}\big] \leqslant \mathbb{E}\big[t^a X^b \mathbf{1}_{X>t}\big] \leqslant \mathbb{E}\big[X^{a+b} \mathbf{1}_{X>t}\big] \,, \tag{6.9}$$

which is the desired result. For $t < \|X\|_a$, we have

$$\mathbb{E}X^a \mathbb{E}\big[X^b \mathbf{1}_{X\leqslant t}\big] > \mathbb{E}\big[t^a X^b \mathbf{1}_{X\leqslant t}\big] \geqslant \mathbb{E}\big[X^{a+b} \mathbf{1}_{X\leqslant t}\big] \,. \tag{6.10}$$

Jensen's (or Hölder's) inequality yields

$$\mathbb{E}X^a \, \mathbb{E}X^b \leqslant \mathbb{E}X^{a+b} \,, \tag{6.11}$$

and the claim follows from (6.10) and (6.11), using $1 = \mathbf{1}_{X\leqslant t} + \mathbf{1}_{X>t}$. ☐

Proof of Lemma 6.11. Let $\chi(t) := \log \mathbb{E} e^{\mathrm{i}th}$. For $n \geqslant 1$, we have

$$\partial_t^n \big(e^{\chi(t)}\big) = \partial_t^{n-1} \big(\chi'(t) e^{\chi(t)}\big) = \sum_{k=1}^{n} \binom{n-1}{k-1} \partial_t^k \big(\chi(t)\big) \partial_t^{n-k} \big(e^{\chi(t)}\big)\,,$$

hence

$$\mathbb{E} h^n = (-\mathrm{i})^n \partial_t^n e^{\chi(t)}\big|_{t=0} = \sum_{k=1}^{n} \binom{n-1}{k-1} \mathcal{C}_k(h) \mathbb{E} h^{n-k}\,.$$

For $g(h) = h^\ell$, we have

$$\mathbb{E}\big[h \cdot g(h)\big] = \mathbb{E} h^{\ell+1} = \sum_{k=1}^{\ell+1} \binom{\ell}{k-1} \mathcal{C}_k(h) \mathbb{E} h^{\ell+1-k} = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}\big[g^{(k)}(h)\big]\,,$$

and by linearity the same relation holds for any polynomial P of degree $\leqslant \ell$:

$$\mathbb{E}\big[h \cdot P(h)\big] = \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}\big[P^{(k)}(h)\big]\,. \tag{6.12}$$

Next, let f be as in the statement of Lemma 6.11, and fix $\ell \in \mathbb{N}$. By Taylor expansion we can find a polynomial P of degree at most ℓ, such that for any $0 \leqslant k \leqslant \ell$,

$$f^{(k)}(h) = P^{(k)}(h) + \frac{1}{(\ell+1-k)!} f^{(\ell+1)}(\xi_k) h^{\ell+1-k}\,, \tag{6.13}$$

where $\xi_k \equiv \xi_k(h)$ is a random variable taking values between 0 and h.

By (6.12), (6.13), homogeneity of the cumulants, and Jensen's inequality we find that the error term in (6.8) satisfies

$$\begin{aligned}
\mathcal{R}_{\ell+1} &= \mathbb{E}\big[h \cdot f(h)\big] - \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}\big[f^{(k)}(h)\big] \\
&= \mathbb{E}\big[h \cdot (f(h) - P(h))\big] - \sum_{k=0}^{\ell} \frac{1}{k!} \mathcal{C}_{k+1}(h) \mathbb{E}\big[f^{(k)}(h) - P^{(k)}(h)\big] \\
&= \frac{1}{(\ell+1)!} \mathbb{E}\big[f^{(\ell+1)}(\xi_0) \cdot h^{\ell+2}\big] - \sum_{k=0}^{\ell} \frac{1}{k!(\ell+1-k)!} \mathcal{C}_{k+1}(h) \mathbb{E}\big[f^{(\ell+1)}(\xi_k) \cdot h^{\ell+1-k}\big] \\
&\leqslant O(1) \cdot \sum_{k=0}^{\ell+1} \mathbb{E}|h|^k \cdot \mathbb{E}\Big[\sup_{|x| \leqslant |h|} \big|f^{(\ell+1)}(x)\big| \cdot h^{\ell+2-k}\Big] \\
&\leqslant O(1) \cdot \mathbb{E}|h|^{\ell+2} \cdot \sup_{|x| \leqslant t} \big|f^{(\ell+1)}(x)\big| + O(1) \cdot \sum_{k=0}^{\ell+1} \mathbb{E}|h|^k \cdot \mathbb{E}\Big[\sup_{|x| \leqslant |h|} \big|f^{(\ell+1)}(x)\big| \cdot h^{\ell+2-k} \cdot \mathbf{1}_{|h|>t}\Big]\,.
\end{aligned}$$

The desired result then follows from estimating the last term of the above by Cauchy-Schwarz inequality and Lemma 6.12. □

6.3.2. Proof of Proposition 6.10 (i). A main tool that we use in our estimate is the following identity.

Lemma 6.13 (Ward)**.** *Let $G = G(z)$. Then*

$$\sum_j |G_{ij}|^2 = \frac{\operatorname{Im} G_{ii}}{\eta}\,.$$

Proof. We have $\sum_j |G_{ij}|^2 = \sum_j G_{ij}G^*_{ji} = (GG^*)_{ii}$. Here $G^* = (H^* - z^*)^{-1} = (H - \bar{z})^{-1}$. For any convertible matrices $A, B \in \mathbb{C}^{N\times N}$, we have

$$A^{-1}(B-A)B^{-1} = A^{-1} - B^{-1}\,. \tag{6.14}$$

Setting $A = H - z$ and $B = H - \bar{z}$ yields

$$G(H - \bar{z} - H + z)G^* = 2\mathrm{i}\eta GG^* = G - G^*\,.$$

Hence

$$\sum_j |G_{ij}|^2 = (GG^*)_{ii} = \frac{G_{ii} - G^*_{ii}}{2\mathrm{i}\eta} = \frac{\operatorname{Im} G_{ii}}{\eta}$$

as desired. □

The Ward identity sums up N many $O(1)$ quantities, and get a result of order η^{-1}. This gives an improvement of factor $(N\eta)^{-1}$. Note that our assumption gives us

$$\max_{ij} |G_{ij}| \prec 1 + \phi\,, \tag{6.15}$$

which will be used to estimate the additional Green functions in our computation. Now we are ready to prove Proposition 6.10.

Let $i, j \in \{1, 2, ..., N\}$, and we use the abbreviation

$$\Pi_{ij} = \Pi(G)_{ij} = \delta_{ij} + zG_{ij} + \underline{G}G_{ij} = (HG)_{ij} + \underline{G}G_{ij}\,.$$

Fix $n \geqslant 1$. We have

$$\begin{aligned}\mathbb{E}|\Pi_{ij}|^{2n} &= \mathbb{E}((HG)_{ij} + \underline{G}G_{ij})\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n} = \mathbb{E}\Big(\sum_k H_{ik}G_{kj} + \underline{G}G_{ij}\Big)\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}\\ &= \sum_k \mathbb{E}H_{ik}G_{kj}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n} + \mathbb{E}\underline{G}G_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n} =: (A) + (B)\,.\end{aligned}$$

Now for each k, we apply Lemma 6.11 with $h = H_{ik}$ and $f(H) = G_{kj}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}$. As the upper triangular entries of H are independent, here we first consider the randomness of H_{ik} and H_{ki} in $f(H)$. Thus

$$(A) = \sum_k \sum_{s=0}^{\ell} \frac{1}{s!}\mathcal{C}_{s+1}(H_{ik})\mathbb{E}\partial_{ik}^{s}(G_{kj}\Pi_{ij}^{s-1}\overline{\Pi}_{ij}^{s}) + \sum_k \mathcal{R}_{\ell+1}^{(k)} =: \sum_{s=0}^{\ell} L_s + \mathcal{R}_{\ell+1}\,,$$

where we abbreviate $\partial_{ik} := \frac{\partial}{\partial H_{ik}}$. The Green function is easy to differentiate, i.e.

$$0 = \frac{\partial G(H-z)}{\partial H_{ij}} = \frac{\partial G}{\partial H_{ij}}(H-z) + G\frac{\partial (H-z)}{\partial H_{ij}} = \frac{\partial G}{\partial H_{ij}}(H-z) + G(\Delta^{ij} + \Delta^{ji})(1+\delta_{ij})^{-1}\,.$$

Here $\Delta^{ij} \in \mathbb{R}^{N\times N}$ satisfy $\Delta^{ij}_{kl} = \delta_{ik}\delta_{jl}$. Thus

$$\frac{\partial G}{\partial H_{ij}} = -G(\Delta^{ij} + \Delta^{ji})(1+\delta_{ij})^{-1}G$$

in the matrix sense, which implies

$$\frac{\partial G_{kl}}{\partial H_{ij}} = -(G_{ki}G_{jl} + G_{kj}G_{il})(1+\delta_{ij})^{-1}\,. \tag{6.16}$$

Recall that $\mathcal{C}_1(H_{ij}) = \mathbb{E}H_{ik} = 0$, and $\mathcal{C}_2(H_{ik}) = \mathbb{E}H_{ik}^2 = N^{-1}(1+O(\delta_{ik}))$. For simplicity, here we assume $\mathbb{E}H_{ik}^2 = N^{-1}(1+\delta_{ik})^{-1}$, i.e. the diagonal entries of H has variance $2/N$. Thus $L_0 = 0$, and

$$\begin{aligned}
L_1 &= \sum_k N^{-1}(1+\delta_{ik})\mathbb{E}\partial_{ik}(G_{kj}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}) \\
&= \sum_k N^{-1}(1+\delta_{ik})\mathbb{E}(-(G_{ki}G_{kj} + G_{kk}G_{ij})(1+\delta_{ik})^{-1}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}) + \sum_k N^{-1}(1+\delta_{ik})\mathbb{E}G_{kj}\partial_{ik}(\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}) \\
&=: L_{1,1} + L_{1,2}\,.
\end{aligned}$$

Now,

$$L_{1,1} = -N^{-1}\mathbb{E}(G^2)_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n} - \mathbb{E}\underline{G}G_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}\,,$$

and we get our key cancellation

$$L_{1,1} + (B) = -N^{-1}\mathbb{E}(G^2)_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}\,.$$

Thus

$$\mathbb{E}|\Pi_{ij}|^{2n} = (A) + (B) = \sum_{s=0}^{\ell} L_s + \mathcal{R}_{\ell+1} + (B) = L_{1,2} + \sum_{s=2}^{\ell} L_s + \mathcal{R}_{\ell+1} - N^{-1}\mathbb{E}(G^2)_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}\,. \tag{6.17}$$

Let us look at the last term on RHS of (6.17). By Lemma 6.13, we have

$$|(G^2)_{ij}| \leqslant \sum_k |G_{ik}G_{kj}| \leqslant \Big(\sum_k |G_{ik}|^2\Big)^{1/2} \cdot \Big(\sum_k |G_{kj}|^2\Big)^{1/2} = \frac{(\operatorname{Im} G_{ii} \operatorname{Im} G_{jj})^{1/2}}{\eta}\,,$$

and thus

$$\Big|N^{-1}\mathbb{E}(G^2)_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}\Big| \leqslant \frac{1}{N\eta}\mathbb{E}\big((\operatorname{Im} G_{ii}\operatorname{Im} G_{jj})^{1/2}|\Pi_{ij}|^{2n-1}\big)\,.$$

We can deduce stochastic domination from moment estimates, and we can also use stochastic domination to assist moment estimates. More precisely, by $\max_{ij}|G_{ij} - m_{sc}\delta_{ij}| \prec \phi$, we have

$$\operatorname{Im} G_{ii} \leqslant \operatorname{Im} m_{sc} + |\operatorname{Im} G_{ii} - \operatorname{Im} m_{sc}| \leqslant \operatorname{Im} m_{sc} + |G_{ii} - \operatorname{Im} m_{sc}| \prec \operatorname{Im} m + \phi\,.$$

Let $D = 100n$, and for any fixed $\varepsilon > 0$, we can find $C \equiv C_{D,\varepsilon}$ such that

$$\mathbb{P}(E_\varepsilon) := \mathbb{P}((\operatorname{Im} G_{ii}\operatorname{Im} G_{jj})^{1/2} > (\operatorname{Im} m_{sc} + \phi)N^{\varepsilon}) \leqslant CN^{-D}$$

for all N. Hence

$$\begin{aligned}
&\frac{1}{N\eta}\mathbb{E}\big((\operatorname{Im} G_{ii}\operatorname{Im} G_{jj})^{1/2}|\Pi_{ij}|^{2n-1}\big) = \frac{1}{N\eta}\mathbb{E}\big((\operatorname{Im} G_{ii}\operatorname{Im} G_{jj})^{1/2}|\Pi_{ij}|^{2n-1}(\mathbf{1}_{E_\varepsilon} + \mathbf{1}_{E_\varepsilon^c})\big) \\
&\leqslant \frac{(\operatorname{Im} m_{sc} + \phi)N^{\varepsilon}}{N\eta}\mathbb{E}|\Pi_{ij}|^{2n-1} + \frac{1}{N\eta}\mathbb{E}\big((\operatorname{Im} G_{ii}\operatorname{Im} G_{jj})^{1/2}|\Pi_{ij}|^{2n-1}\mathbf{1}_{E_\varepsilon}\big) \\
&\leqslant \frac{(\operatorname{Im} m_{sc} + \phi)N^{\varepsilon}}{N\eta}\mathbb{E}|\Pi_{ij}|^{2n-1} + \mathbb{P}(E_\varepsilon)\cdot N^{10n}\,,
\end{aligned}$$

where in the last step we used the deterministic bound $\max_{ij}|G_{ij}| \leqslant \eta^{-1} \leqslant N$. Since ε is arbitrary, we arrive at

$$\begin{aligned}\left|N^{-1}\mathbb{E}(G^2)_{ij}\Pi_{ij}^{n-1}\overline{\Pi}_{ij}^{n}\right| &\leqslant \frac{1}{N\eta}\mathbb{E}\big((\operatorname{Im} G_{ii} \operatorname{Im} G_{jj})^{1/2}|\Pi_{ij}|^{2n-1}\big) \\ &\prec \frac{\operatorname{Im} m_{sc}+\phi}{N\eta}\mathbb{E}|\Pi_{ij}|^{2n-1} + O(N^{-90n}) \prec \frac{\operatorname{Im} m_{sc}+\phi}{N\eta}\big(\mathbb{E}|\Pi_{ij}|^{2n}\big)^{\frac{2n-1}{2n}}.\end{aligned} \tag{6.18}$$

We aim for the following estimate.

Lemma 6.14.

$$\mathbb{E}|\Pi_{ij}|^{2n} \prec \sum_{a=1}^{2n}\mathcal{E}^a\big(\mathbb{E}|\Pi_{ij}|^{2n}\big)^{\frac{2n-a}{2n}}, \quad \textit{where} \quad \mathcal{E} := (1+\phi)^3\sqrt{\frac{\operatorname{Im} m_{sc}+\phi}{N\eta}}\,.$$

Apparently, Lemma 6.14 implies $\mathbb{E}|\Pi_{ij}|^{2n} \prec \mathcal{E}^{2n}$. As n is arbitrary, we get the desired estimate (6.4) from Markov's inequality.

By (6.18), the last term on RHS of (6.17) is good. Let us look at other terms in (6.17). There are several terms in $L_{1,2}$. For instance, we have

$$\begin{aligned}&\Big|\sum_k N^{-1}(1+\delta_{ik})\mathbb{E}G_{kj}\partial_{ik}(zG_{ij})\Pi_{ij}^{n-2}\overline{\Pi}_{ij}^{n}\Big| \\ &\leqslant C\sum_k N^{-1}(1+\delta_{ik})\mathbb{E}\big|G_{kj}(G_{ii}G_{kj}+G_{ik}G_{ij})(1+\delta_{ik})^{-1}\Pi_{ij}^{2n-2}\big| \\ &\leqslant CN^{-1}\sum_k \mathbb{E}(|G_{kj}G_{ii}G_{kj}|+|G_{kj}G_{ik}G_{ij}|)|\Pi_{ij}^{2n-2}| \prec (1+\phi)\frac{\operatorname{Im} m_{sc}+\phi}{N\eta}\mathbb{E}|\Pi_{ij}|^{2n-2} \leqslant \mathcal{E}^2\big(\mathbb{E}|\Pi_{ij}|^{2n}\big)^{\frac{2n-2}{2n}}.\end{aligned}$$

Here in the second last term we used (6.15) and Lemma 6.13. Other terms in $L_{1,2}$ works in a similar way. We have

$$L_{1,2} \prec \mathcal{E}^2\big(\mathbb{E}|\Pi_{ij}|^{2n}\big)^{\frac{2n-2}{2n}}. \tag{6.19}$$

We have finished the estimates of L_1, which corresponds to the second cumulant of H_{ik}.

The higher cumulant terms are easier to estimate: recall our assumption

$$\mathbb{E}|\sqrt{N}H_{ij}|^s \leqslant C_s$$

which implies

$$\mathcal{C}_s(\sqrt{N}H_{ij}) \leqslant \widehat{C}_s \quad \text{and} \quad \mathcal{C}_s(H_{ij}) \leqslant \widehat{C}_s N^{-s/2}\,.$$

Here in the last step we used $\mathcal{C}_s(aH_{ij}) = a^s\mathcal{C}_s(H_{ij})$. One readily sees that

$$L_s = \sum_k \frac{1}{s!}\mathcal{C}_{s+1}(H_{ik})\mathbb{E}\partial_{ik}^s(G_{kj}\Pi_{ij}^{s-1}\overline{\Pi}_{ij}^{s}) \prec N^{-(s+1)/2}\sum_k \mathbb{E}|\partial_{ik}^s(G_{kj}\Pi_{ij}^{s-1}\overline{\Pi}_{ij}^{s})| \prec \sum_{a=1}^{(s+1)\wedge 2n}\mathcal{E}^a\mathbb{E}(|\Pi_{ij}|^{2n})^{\frac{2n-a}{2n}}$$

for all $s \geqslant 2$. Together with (6.17) – (6.19), and assming that $\mathcal{R}_{\ell+1}$ is small enough for large, fixed ℓ[7], we finish the proof of Lemma 6.14. This concludes the proof of (6.6).

[7]It is generally true that the remainder term is small enough for large ℓ, and one can refer to [9, Lemma 4.6 (i)] for a proof in a slightly different setting.

6.3.3. Idea of the Proof of Proposition 6.10 (ii). For (6.7), note that the summation in $\underline{\Pi(G)} = \frac{1}{N}\sum_i \Pi_{ii}$ gives rise to one more Ward Identity. We omit the details here.

6.4. Proof of Theorem 6.5: deterministic part

Let $z = E + \mathrm{i}\eta$. In this section, we always consider the same $E \in [-10, 10]$.

Step 1. We first consider $\eta \in [2, 10]$. In this case, $|G_{ij}| \leqslant \eta^{-1} \leqslant 1/2$. Since $|m_{sc}| \leqslant 1$, we have

$$\phi = \max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \leqslant 2\,.$$

By (6.7),

$$1 + z\underline{G} + \underline{G}^2 \prec (1+\phi)^6 \frac{\operatorname{Im} m + \phi}{N\eta} \prec \frac{1}{N\eta}\,,$$

and thus

$$\underline{G} = -\frac{1}{z + \underline{G}}\Big(1 + O_\prec\Big(\frac{1}{N\eta}\Big)\Big)\,.$$

Note that $\operatorname{Im}(z + \operatorname{Im}\underline{G}) \geqslant \operatorname{Im} z \geqslant 2$, and $1 + zm_{sc} + m_{sc}^2 = 0$ implies $m_{sc} = -\frac{1}{z+m_{sc}}$. Hence

$$\underline{G} - m_{sc} = -\frac{1}{z + \underline{G}}\Big(1 + O_\prec\Big(\frac{1}{N\eta}\Big)\Big) + \frac{1}{z + m_{sc}} = \frac{\underline{G} - m_{sc}}{(z + m_{sc})(z + \underline{G})} + O_\prec\Big(\frac{1}{N\eta}\Big)\,,$$

which implies $G - m_{sc} \prec (N\eta)^{-1}$. This proves (6.3). To prove (6.4), we need the following input.

Lemma 6.15. *Recall the definition of* **S** *from* (6.2)*, and let* $z \in \mathbf{S}$*. Suppose at* z *we have* $\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \phi$*,* $\phi \in [N^{-1}, N^{\tau/10}]$*, and* $\underline{G} - m_{sc} \prec \theta$ *for some* $\theta \leqslant N^{-\tau/10}$*. Then*

$$\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \theta + \sqrt{\frac{\operatorname{Im} m_{sc}}{N\eta}} + \frac{1}{N\eta}\,.$$

Proof. By (6.6), we know that

$$\delta_{ij} + zG_{ij} + \underline{G}G_{ij} \prec (1+\phi)^3 \sqrt{\frac{\operatorname{Im} m_{sc} + \phi}{N\eta}}\,.$$

Since

$$\frac{1}{z + \underline{G}} = \frac{1}{z + m_{sc} + \underline{G} - m_{sc}} = \frac{1}{z + m_{sc}} + O_\prec(\theta) = -m_{sc} + O_\prec(\theta)\,,$$

we get

$$G_{ij} - m_{sc}\delta_{ij} \prec \theta + (1+\phi)^3 \sqrt{\frac{\operatorname{Im} m_{sc} + \phi}{N\eta}}\,.$$

In other words, we have

$$\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \theta + (1+\phi)^3 \sqrt{\frac{\operatorname{Im} m_{sc} + \phi}{N\eta}} \tag{6.20}$$

provided that $\max_{ij} |G_{ij} - m_{sc}| \prec \phi$. Iterating (6.20) finitely many times yields the desired result. □

Now for $\eta \geqslant 2$, the condition of Lemma 6.15 holds, with $\underline{G} - m_{sc} \prec \frac{1}{N\eta} =: \theta$. Thus

$$\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \sqrt{\frac{\operatorname{Im} m_{sc}}{N\eta}} + \frac{1}{N\eta}\,,$$

which proves Theorem 6.5 for $\eta \geqslant 2$.

Step 2. Let us consider the general case $\eta \geqslant N^{-1+\tau}$. Recall that Lemma 6.9 states $\Gamma(z) := \max_{ij} |G_{ij}(z)|$. For any $M > 1$, we have $\Gamma(E + \mathrm{i}\eta/M) \leqslant M\Gamma(E + \mathrm{i}\eta)$.

Proof of Lemma 6.9. Fix E and write $\Gamma(\eta) = \Gamma(E + \mathrm{i}\eta)$. For h small enough,

$$|\Gamma(\eta + h) - \Gamma(\eta)| \leqslant \max_{ij} |G_{ij}(E + \mathrm{i}(\eta + h)) - G_{ij}(E + \mathrm{i}\eta)| = |h| \max_{ij} |(G(E + \mathrm{i}(\eta + h))G(E + \mathrm{i}\eta))_{ij}|$$
$$\leqslant |h| \max_{ij} \sum_k |G(E + \mathrm{i}(\eta + h))_{ik}||G_{kj}(E + \mathrm{i}\eta)| \leqslant |h| \sqrt{\frac{\Gamma(\eta + h)\Gamma(\eta)}{(h + \eta)\eta}}\,,$$

where in the second step we used (6.14) and in the last step we used Lemma 6.13. The above implies Γ is locally Lipschitz continuous, and its almost everywhere exists derivative satisfies

$$\left|\frac{\mathrm{d}\Gamma}{\mathrm{d}\eta}\right| \leqslant \frac{\Gamma}{\eta}\,.$$

Thus $\frac{\mathrm{d}}{\mathrm{d}\eta}(\eta\Gamma(\eta)) \geqslant 0$, which implies the desired result. □

To proceed the proof, we need one more result.

Lemma 6.16. *Let $\Delta : \mathbf{S} \to \mathbb{R}$ be a function such that*

$$1 + z\underline{G} + \underline{G}^2 \prec \Delta\,.$$

Suppose $\Delta \in [N^{-2}, N^{-\varepsilon}]$, and $\Delta \equiv \Delta(z)$ is Lipschitz continuous with Lipschitz constant N. Moreover, assume that for each fixed E, the function $\eta \mapsto \Delta(E + \eta)$ is non-increasing. Then

$$\underline{G} - m_{sc} \prec \frac{\Delta}{\operatorname{Im} m_{sc} + \sqrt{\Delta}}\,.$$

The proof of Lemma 6.16 is in fact quite technical (and non-trivial), as the square root $\sqrt{z}$ is unstable when $\operatorname{Im} z$ is close to 0. It can be found in e.g. [4, Lemma 4.5].

Now let us decompose $\{E\} \times [N^{-\tau+1}, 10] = \cup_{k=0}^{n} W_k$, where $W_0 := \{E\} \times [2, 10]$, and $W_k := \{E\} \times [2N^{-\delta k}, 2N^{-\delta(k-1)}]$ for $k \geqslant 1$. Here $\delta := \tau/100$, and $n \leqslant \delta^{-1}$. In particular, $n = O(1)$. The following result will finish the proof of Theorem 6.5.

Lemma 6.17. *For $k = 1, 2, ..., n$. Suppose Theorem 6.5 holds for all $z \in W_{k-1}$, then it also holds for all $z \in W_k$.*

Proof. Let $b_k := E + \mathrm{i}2N^{-\delta(k-1)}$ be the upper edge of W_k (and the lower edge of W_{k-1}). Since

$$\max_{i,j} |G_{ij} - \delta_{ij} m_{sc}| \prec \frac{1}{N\eta} + \sqrt{\frac{\operatorname{Im} m_{sc}}{N\eta}}$$

at b_k, we know that $\max_{ij} |G_{ij}| \prec 1$ at b_k. By Lemma 6.9, $\max_{ij} |G_{ij}| \prec N^{\delta}$ for all $z \in W_k$. Thus

$$\max_{ij} |G_{ij} - \delta_{ij} m_{sc}| \prec N^{\delta}$$

for all $z \in W_k$. Now pick $z \in W_k$. By Proposition 6.10,

$$1 + z\underline{G} + \underline{G}^2 \prec (1 + N^{\delta})^6 \frac{\operatorname{Im} m_{sc} + N^{\delta}}{N\eta}\,,$$

and Lemma 6.16 shows

$$\underline{G} - m_{sc} \prec \frac{1}{(N\eta)^{1/4}} .$$

Now suppose $\underline{G} - m_{sc} \prec \theta$, $\theta \leqslant N^{-\tau/10}$. Lemma 6.15 shows

$$\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \theta + \sqrt{\frac{\operatorname{Im} m_{sc}}{N\eta}} + \frac{1}{N\eta} =: \phi .$$

Applying Proposition 6.10 yields

$$1 + z\underline{G} + \underline{G}^2 \prec \frac{\operatorname{Im} m_{sc} + \phi}{N\eta} = \frac{\operatorname{Im} m_{sc}}{N\eta} + \frac{1}{(N\eta)^2} + \frac{\theta}{N\eta} =: \Delta .$$

By Lemma 6.16, we get

$$\underline{G} - m_{sc} \prec \frac{\Delta}{\operatorname{Im} m_{sc} + \sqrt{\Delta}} \prec \frac{1}{N\eta} + \sqrt{\frac{\theta}{N\eta}} ,$$

provided that $G - m_{sc} \prec \theta$. Iterating the above yields $G - m_{sc} \prec \frac{1}{N\eta}$. Finally, applying Lemma 6.15 with $\theta = \frac{1}{N\eta}$, we get

$$\max_{ij} |G_{ij} - m_{sc}\delta_{ij}| \prec \frac{\operatorname{Im} m_{sc}}{N\eta} + \frac{1}{N\eta} .$$

This finishes the proof. □

Remark 6.18. Theorem 6.5 proves the local law for individual z. In fact, the result can be easily strengthened to

$$\sup_{z \in \mathbf{S}} |\underline{G} - m_{sc}| \prec \frac{1}{N\eta} \quad \text{and} \quad \sup_{z \in \mathbf{S}} \max_{i,j} |G_{ij} - \delta_{ij} m_{sc}| \prec \frac{1}{N\eta} + \sqrt{\frac{\operatorname{Im} m_{sc}}{N\eta}} . \tag{6.21}$$

Obviously, by the definition of Stochastic domination, the local law is true if we take the supremum over a N^{-3}-net of $\mathbf{S}$. One can then prove (6.21) by observing that

$$|G_{ij}(z) - G_{ij}(z+\delta)| \leqslant |\delta| |(G(z)G(z+\delta))_{ij}| \leqslant N^2 |\delta|$$

whenever $z, z+\delta \in \mathbf{S}$.

6.5. Applications of the local law

6.5.1. Proof of Corollary 6.6. We shall also use the following result in complex analysis.

Lemma 6.19 (Helffer–Sjöstrand formula)**.** *Let $f \in C_c^2(\mathbb{R})$. Let $\widetilde{f}$ be the almost analytic extension of f defined by*

$$\widetilde{f}(x + \mathrm{i}y) = f(x) + \mathrm{i}y f'(x). \tag{6.22}$$

Let $\chi \in C_c^\infty(\mathbb{R})$ be a cutoff function satisfying $\chi(0) = 1$, and by a slightly abuse of notation we write $\chi(z) \equiv \chi(\operatorname{Im} z)$. For any $\lambda \in \mathbb{R}$, we have

$$f(\lambda) = \frac{1}{\pi} \int_{\mathbb{C}} \frac{\partial_{\bar{z}}(\widetilde{f}(z)\chi(z))}{\lambda - z} \, \mathrm{d}^2 z .$$

Here $\partial_{\bar{z}} = \frac{1}{2}(\partial_x + \mathrm{i}\partial_y)$ is the antiholomorphic derivative.

Proof. Exercise (it is essentially Cauchy integral formula and Green's theorem). ☐

Fix $\varepsilon > 0$ and let $\eta = N^{-1+\varepsilon}$. Let $I \subset [-10, 10]$ be given. Choose $f \in C_c^\infty(\mathbb{R})$ such that $f(x) = 1$ for $x \in I$, $f(x) = 0$ for $\operatorname{dist}(x, I) \geqslant \eta$, $\|f'\|_\infty = O(\eta^{-1})$, and $\|f''\|_\infty = O(\eta^{-2})$. Let $\widetilde{f}$ be as in Lemma 6.19. Then

$$\int_{\mathbb{R}} f(x)\mu(x)\mathrm{d}x = \frac{1}{N}\sum_i f(\lambda_i) = \frac{1}{N}\sum_i \frac{1}{\pi}\int_{\mathbb{C}} \frac{\partial_{\bar z}(\widetilde{f}(z)\chi(z))}{\lambda_i - z}\,\mathrm{d}^2 z = \frac{1}{\pi}\int_{\mathbb{C}} \partial_{\bar z}(\widetilde{f}(z)\chi(z))\underline{G}(z)\,\mathrm{d}^2 z\,,$$

and similarly, $\int_{\mathbb{R}} f(x)\varrho_{sc}(x)\mathrm{d}x = \frac{1}{\pi}\int_{\mathbb{C}} \partial_{\bar z}(\widetilde{f}(z)\chi(z))m_{sc}(z)\,\mathrm{d}^2 z$. Thus

$$\begin{aligned}&\int_{\mathbb{R}} f(x)(\mu(x) - \varrho_{sc}(x))\mathrm{d}x = \frac{1}{\pi}\int_{\mathbb{C}} \partial_{\bar z}(\widetilde{f}(z)\chi(z))(\underline{G}(z) - m_{sc}(z))\,\mathrm{d}^2 z \\ =&\frac{1}{2\pi}\int_{\mathbb{C}} \big[(f(x) + \mathrm{i}yf'(x))\mathrm{i}\chi'(z) + \mathrm{i}yf''(x)\chi(z)\big](\underline{G}(z) - m_{sc}(z))\mathrm{d}z^2\,.\end{aligned} \tag{6.23}$$

Note that in (6.23), we have the freedom of choosing χ to our advantage (it can depend on N if needed). Let $\chi(z) = 1$ for $|\operatorname{Im} z| \leqslant 1$, and $\chi(z) = 0$ for $|\operatorname{Im} z| \geqslant 2$. By Theorem 6.5, we have

$$\int_{\mathbb{C}_+} f(x)\chi'(z)(\underline{G}(z) - m_{sc}(z))\mathrm{d}z^2 \prec \int_{\mathbb{R}}\int_1^2 |f(x)|\frac{1}{Ny}\mathrm{d}y\mathrm{d}x \prec N^{-1}\,, \tag{6.24}$$

and as $|f'|_1 \leqslant \|f'\|_\infty \cdot |\operatorname{supp}(f')| = O(1)$,

$$\int_{\mathbb{C}_+} yf'(x)\chi'(z)(\underline{G}(z) - m_{sc}(z))\mathrm{d}z^2 \prec \int_{\mathbb{R}}\int_1^2 y|f'(x)|\frac{1}{Ny}\mathrm{d}y\mathrm{d}x \prec N^{-1}\,. \tag{6.25}$$

To estimate the last term on RHS of (6.23), we need the following result.

Lemma 6.20. *We have*

$$\underline{G}(z) \prec \frac{N^\varepsilon}{Ny}$$

for all $0 < y \leqslant \eta = N^{-1+\varepsilon}$.

Proof. By Theorem 6.5, we know that

$$\max_{ij} |G_{ij}(x + \mathrm{i}\eta)| \prec 1\,.$$

Using Lemma 6.9,

$$\max_{ij} |G_{ij}(x + \mathrm{i}y)| = \Gamma(x + \mathrm{i}y) \leqslant \frac{\eta}{y}\Gamma(x + \mathrm{i}\eta) = \frac{N^\varepsilon}{Ny}\,,$$

which implies the desired result. ☐

By Lemma 6.20,

$$\int_{0<y<\eta}\int_{\mathbb{R}} yf''(x)\chi(z)(\underline{G}(z) - m_{sc}(z))\mathrm{d}x\mathrm{d}y \prec \int_{0<y<\eta}\int_{\mathbb{R}} |yf''(x)|\frac{N^\varepsilon}{Ny}\mathrm{d}x\mathrm{d}y \prec \frac{N^\varepsilon}{N}\,. \tag{6.26}$$

In addition,

$$\int_{y\geqslant\eta}\int_{\mathbb{R}} yf''(x)\chi(z)(\underline{G}(z)-m_{sc}(z))\mathrm{d}x\mathrm{d}y = -\int_{y\geqslant\eta}\int_{\mathbb{R}} yf'(x)\chi(z)\partial_x(\underline{G}(z)-m_{sc}(z))\mathrm{d}x\mathrm{d}y$$
$$= \mathrm{i}\int_{y\geqslant\eta}\int_{\mathbb{R}} yf'(x)\chi(z)\partial_y(\underline{G}(z)-m_{sc}(z))\mathrm{d}x\mathrm{d}y$$
$$= \mathrm{i}\int_{\mathbb{R}} \eta f'(x)\chi(\eta)(\underline{G}(x+\mathrm{i}\eta)-m_{sc}(x+\mathrm{i}\eta))\mathrm{d}x - \mathrm{i}\int_{y\geqslant\eta}\int_{\mathbb{R}} \partial_y(yf'(x)\chi(z))(\underline{G}(z)-m_{sc}(z))\mathrm{d}x\mathrm{d}y \prec N^{-1}\,.$$

Inserting the above and (6.24) – (6.26) into (6.23) yields

$$\int_{\mathbb{R}} f(x)(\mu(x)-\varrho_{sc}(x))\mathrm{d}x \prec N^{-1+\varepsilon}\,. \tag{6.27}$$

Thus for any $I\subset[-10,10]$,

$$\mu(I)\leqslant \int f_{I,\eta}(x)\mu(x)\mathrm{d}x \leqslant \int f_{I,\eta}(x)\varrho_{sc}(x)\mathrm{d}x + O_{\prec}(N^{-1+\varepsilon}) \leqslant \varrho_{sc}(I)+O_{\prec}(N^{-1+\varepsilon})\,.$$

Let $I' := \{x : \operatorname{dist}(x,I^c)\geqslant\eta\}$. Then

$$\mu(I)\geqslant \int f_{I',\eta}(x)\mu(x)\mathrm{d}x \geqslant \int f_{I',\eta}(x)\varrho_{sc}(x)\mathrm{d}x + O_{\prec}(N^{-1+\varepsilon}) \geqslant \varrho_{sc}(I)+O_{\prec}(N^{-1+\varepsilon})\,.$$

As ε is arbitrary, we conclude that

$$\mu(I)-\varrho_{sc}(I)\prec N^{-1} \tag{6.28}$$

for all $I\subset[-10,10]$. Note that (6.28) also implies

$$\mu([-2,2]^c) = 1-\mu([-2,2]) = 1-\varrho_{sc}([-2,2])+O_{\prec}(N^{-1})\prec N^{-1}\,.$$

This finishes the proof of Corollary 6.6.

6.5.2. Proof of Corollary 6.7. We shall prove

$$\lambda_i-\gamma_i\prec N^{-1} \tag{6.29}$$

for all $i\in[cN,(1-c)N]$. The edge case is more complicated and we omit here.

Define the eigenvalue counting function $\Sigma:\mathbb{R}\to\{1,2,...,N\}$ by

$$\Sigma(E) := |\{i:\lambda_i\geqslant E\}|\,.$$

Fix $\varepsilon>0$. We have the duality

$$\{\lambda_i\leqslant\gamma_i+N^{-1+\varepsilon}\} = \{\Sigma(\gamma_i+N^{-1+\varepsilon})\leqslant i\}\,.$$

Thus

$$\frac{1}{N}\Sigma(\gamma_i+N^{-1+\varepsilon}) = \mu([\gamma_i+N^{-1+\varepsilon},+\infty))$$
$$= \mu([\gamma_i+N^{-1+\varepsilon},+\infty)) - \varrho_{sc}([\gamma_i+N^{-1+\varepsilon},+\infty)) + \varrho_{sc}([\gamma_i,+\infty)) - \varrho_{sc}([\gamma_i,\gamma_i+N^{-1+\varepsilon}])$$
$$= O_{\prec}(N^{-1}) + \frac{\mathrm{i}}{N} - C_iN^{-1+\varepsilon}\,.$$

Thus $\Sigma(\gamma_i+N^{-1+\varepsilon})\leqslant i$ with very high probability, which implies

$$(\lambda_i-\gamma_i)_+\prec N^{-1+\varepsilon}\,.$$

The same can be said for $(\lambda_i-\gamma_i)_-$. As ε is arbitrary, we get the desired result.

Exercise 6.21. Repeat the above argument near the edge, and show that

$$(\lambda_1 - 2)_- \prec N^{-2/3}\,.$$

In addition, figure out why we cannot use the same steps to deduce

$$(\lambda_1 - 2)_+ \prec N^{-2/3}\,.$$

6.6. The complex case

Recall the definition of Wigner matrices from Definition 5.1. The *symmetry class* of the model is characterized by the second moment of the off-diagonal entries. More precisely, when H is real symmetric, we know that

$$\mathbb{E}H_{ij}^2 = \mathbb{E}|H_{ij}^2| = N^{-1}$$

for all $i \neq j$. When H is complex Hermitian, we have

$$\mathbb{E}H_{ij}^2 = 0 \tag{6.30}$$

for all $i \neq j$. This is because a complex Gaussian random variable $Z \overset{d}{=} \mathcal{N}_{\mathbb{C}}(0, N^{-1})$ satisfy $\mathbb{E}Z^2 = 0$, and we have to require the same for H_{ij}, to ensure the matching of the first two moments.

In this section we explain some steps of the proof of Theorem 6.5, under the assumption (6.30). Our main tool is the following complex version of Lemma 6.11.

Lemma 6.22. *(Complex cumulant expansion) Let h be a complex random variable with all its moments exist. The (p,q)-cumulant of h is defined as*

$$\mathcal{C}^{(p,q)}(h) := (-i)^{p+q} \cdot \left(\frac{\partial^{p+q}}{\partial s^p \partial t^q} \log \mathbb{E} e^{\mathrm{i}sh + \mathrm{i}t\bar{h}}\right)\bigg|_{s=t=0}\,.$$

Let $f : \mathbb{C}^2 \to \mathbb{C}$ be a smooth function, and we denote its holomorphic derivatives by

$$f^{(p,q)}(z_1, z_2) := \frac{\partial^{p+q}}{\partial {z_1}^p \partial {z_2}^q} f(z_1, z_2)\,.$$

Then for any fixed $\ell \in \mathbb{N}$, we have

$$\mathbb{E}f(h, \bar{h})\bar{h} = \sum_{p+q=0}^{\ell} \frac{1}{p!\,q!} \mathcal{C}^{(p,q+1)}(h)\mathbb{E}f^{(p,q)}(h, \bar{h}) + R_{l+1}\,, \tag{6.31}$$

given all integrals in (6.31) exists. Here $R_{\ell+1}$ is the remainder term depending on f and h, and for any $t > 0$, we have the estimate

$$\begin{aligned} R_{\ell+1} = {} & O(1) \cdot \mathbb{E}\big|h^{\ell+2} \cdot \mathbf{1}_{\{|h|>t\}}\big| \cdot \max_{p+q=\ell+1} \big\|f^{(p,q)}(z, \bar{z})\big\|_\infty \\ & + O(1) \cdot \mathbb{E}|h|^{\ell+2} \cdot \max_{p+q=\ell+1} \big\|f^{(p,q)}(z, \bar{z}) \cdot \mathbf{1}_{\{|z|\leqslant t\}}\big\|_\infty\,. \end{aligned}$$

Let $G \equiv G(z) = (H - z)^{-1}$ with $z \in \mathbf{S}_\tau$. Let $m, n \geqslant 1$. Since H is complex hermitian, we have the differential rule

$$\frac{\partial G_{ij}}{\partial H_{kl}} = -G_{ik}G_{lj}\,. \tag{6.32}$$

We can then prove Proposition 6.10 with the aid of Lemma 6.22 and (6.32). The (deterministic) steps from Proposition 6.10 to Theorem 6.5 are identical to the real symmetric case.

6.7. Outro

In Theorem 6.5, we prove the local semicircle law for the centered model H using Lemma 6.11. In history, the first proof of Theorem 6.5 was conducted through a more complicated method named Schur complement, which we will not discuss here. What is only possible with Lemma 6.11 is the local law for Wigner matrices with arbitrary expectations.

Theorem 6.23. *Let H be as in Definition 5.1, and let $A \in \mathbb{C}^{N\times N}$ be a deterministic, complex Hermitian matrix. Consider the random matrix $W := H + A$. Let $z \in \mathbf{S}_\tau$, with $\mathbf{S}_\tau$ as in Definition 6.2. Consider the Green function $\widetilde{G}(z) := (W - z)^{-1}$, and let $M \in \mathbb{C}^{N\times N}$ be the solution of*

$$I + zM - AM + \underline{M}M = 0$$

with positive imaginary part. Let $\mathbf{v}, \mathbf{w} \in \mathbb{S}^{N-1}$, and $B \in \mathbb{C}^{N\times N}$ satisfy $\|B\| = 1$. Then

$$\underline{B\widetilde{G}} - \underline{BM} \prec \frac{1}{N\eta}$$

and

$$\langle \mathbf{v}, \widetilde{G}\mathbf{w}\rangle - \langle \mathbf{v}, M\mathbf{w}\rangle \prec \frac{1}{N\eta} + \sqrt{\frac{\operatorname{Im}\underline{M}}{N\eta}} \tag{6.33}$$

uniformly for all $z \in \mathbf{S}_\tau$.

Theorem 6.23 is stronger than Theorem 6.5 not only in the sense of allowing arbitrary expectations, but also generalizing the local law by allowing test matrices and vectors in the result. In particular, the estimate of the form (6.33) is called the *isotropic law*, and it controls the matrix $\widetilde{G} - M$ in the weak operator sense. By repeating the proof of Corollary 6.8, it is not hard to see that (6.33) implies

$$\max_i |\langle \mathbf{u}_i, \mathbf{v}\rangle| \prec N^{-1/2}\,,$$

where $\mathbf{u}_1, ..., \mathbf{u}_N$ are the eigenvectors of W. This is known as the isotropic delocalization.

Beyond the isotropic law, convergence in the strong operator sense must fail:

$$\big|(\widetilde{G} - M)\mathbf{v}\big| = \sup_{|\mathbf{w}|\leqslant 1} \big|\langle \mathbf{w}, (\widetilde{G} - M)\mathbf{v}\rangle\big| \geqslant \big|\langle \mathbf{u}_i, (\widetilde{G} - M)\mathbf{v}\rangle\big| \geqslant \left|\frac{\langle \mathbf{u}_i, \mathbf{v}\rangle}{\lambda_i - z} - \langle \mathbf{u}_i, M\mathbf{v}\rangle\right|.$$

We have $|\langle \mathbf{u}_i, M\mathbf{v}\rangle| \leqslant \|M\| = O(1)$. On the other hand, for E in the bulk of the spectrum, we can choose i according to E such that

$$\left|\frac{\langle \mathbf{u}_i, \mathbf{v}\rangle}{\lambda_i - z}\right| \geqslant \frac{|\langle \mathbf{u}_i, \mathbf{v}\rangle|}{C\eta}\,.$$

In addition, $|\langle \mathbf{u}_i, \mathbf{v}\rangle| \geqslant N^{-1/2}$ with probability $1 - o(1)$ (this is easy to see at least for GOE, as $\mathbf{u}_i$ is uniformly distributed on the unit sphere). To sum up, we have

$$|(\widetilde{G} - M)\mathbf{v}| \geqslant \frac{1}{C\sqrt{N}\eta}$$

with probability $1 - o(1)$.

To prove Theorem 6.23, the key step is showing

$$I + zG - AG + \underline{G}G \approx 0\,.$$

The proof is in fact surprisingly similar to that of Proposition 6.10, as

$$I + zG - AG + \underline{G}G = HG + \underline{G}G\,.$$

7 Fluctuations

Let H be a Wigner matrix as in Definition 5.1. Unlike the local law, when we study the eigenvalue fluctuations, the symmetry class will play a role in the result. For this reason, we set $\beta \equiv \beta(H) = 1$ if $\mathbb{E}H_{ij}^2 = N^{-1}$ for all $i \neq j$ (real symmetric case), and $\beta = 2$ if $\mathbb{E}H_{ij}^2 = 0$ for all $i \neq j$ (complex Hermitian case).

7.1. Linear eigenvalue statistics

In Sections 7.1 – 7.2, we further assume that $H_{11} \overset{d}{=} \cdots \overset{d}{=} H_{NN}$ and $H_{ij} \overset{d}{=} H_{i'j'}$ for all $i < j$ and $i' < j'$. Set

$$a_2 = N\mathbb{E}H_{11}^2\,, \quad a_3 = N^{3/2}\mathbb{E}H_{11}^3\,, \quad \text{and} \quad s_4 = N^2\mathcal{C}_4(H_{12})\,.$$

We have the following result.

Theorem 7.1. *Let $f \in C^2(\mathbb{R})$. Then*

$$\frac{\operatorname{Tr} f(H) - \mu_f}{\sigma_f} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Here

$$\begin{aligned}\mu_f := &N\int_{-2}^{2} f(x)\rho_{sc}(x)\mathrm{d}x - \frac{1}{2\pi}\Big(\frac{2}{\beta}-1\Big)\int_{-2}^{2}\frac{f(x)}{\sqrt{4-x^2}}\mathrm{d}x + \frac{1}{4}\Big(\frac{2}{\beta}-1\Big)(f(2)+f(-2))\\ &-\frac{a_2-2\beta^{-1}}{2\pi}\int_{-2}^{2} f(x)\frac{2-x^2}{\sqrt{4-x^2}}\mathrm{d}x + \frac{s_4}{2\pi}\int_{-2}^{2} f(x)\frac{x^4-4x^2+2}{\sqrt{4-x^2}}\mathrm{d}x\end{aligned} \tag{7.1}$$

and

$$\begin{aligned}\sigma_f^2 := &\frac{1}{2\beta\pi^2}\int_{-2}^{2}\int_{2}^{2}\frac{(f(y)-f(x))^2}{(x-y)^2}\frac{4-xy}{\sqrt{4-x^2}\sqrt{4-y^2}}\mathrm{d}x\mathrm{d}y\\ &+\frac{a_2-2\beta^{-1}}{4\pi^2}\Big(\int_{-2}^{2} f(x)\frac{x}{\sqrt{4-x^2}}\mathrm{d}x\Big)^2 + \frac{s_4}{2\pi^2}\Big(\int_{-2}^{2} f(x)\frac{2-x^2}{\sqrt{4-x^2}}\mathrm{d}x\Big)^2.\end{aligned} \tag{7.2}$$

In Theorem 7.1, the leading term of $\operatorname{Tr} f(H)$ is $N\int_{-2}^{2} f(x)\rho_{sc}(x)\mathrm{d}x$, which comes from Wigner's semicircle law, and it is of order N. In contrast, the fluctuation of $\operatorname{Tr} f(H)$ sits on much smaller scale 1. In order to reveal it, we also need to identify all $O(1)$ terms on the expectation of $\operatorname{Tr} f(H)$. This leads to the extra terms in μ_f.

From Lemma 6.19 and (6.23), we see that

$$\operatorname{Tr} f(H) - N\int_{\mathbb{R}} f(x)\varrho_{sc}(x))\mathrm{d}x = \frac{N}{\pi}\int_{\mathbb{C}} \partial_{\bar z}(\widetilde{f}(z)\chi(z))(\underline{G}(z)-m_{sc}(z))\,\mathrm{d}^2 z\,. \tag{7.3}$$

It is easy to see that by setting $\chi(z) = 1$ for $|\operatorname{Im} z| \leqslant 1$, using Theorem 6.5 and Lemma 6.20, the above integration is negligible when $|\operatorname{Im} z| \geqslant N^{-\varepsilon}$ for any fixed $\varepsilon > 0$. Thus the essential step is obtaining the fluctuation of $\underline{G}(z) - m_{sc}(z)$, where $\operatorname{Im} z = \eta \geqslant N^{-\varepsilon}$.

Here we illustrate on computing the variance for $\beta = 1$ (the real symmetric case). As $1 + zm_{sc} + m_{sc}^2 = 0$, we have

$$
\begin{aligned}
\mathbb{E}|\underline{G} - m_{sc}|^2 &= \mathbb{E}[-m_{sc}((z + m_{sc})\underline{G} + 1)(\underline{G^*} - m_{sc}^*)] = -m_{sc}\mathbb{E}(\underline{HG} + m_{sc}\underline{G})(\underline{G^*} - m_{sc}^*) \\
&= -m_{sc}N^{-1}\sum_{ij}\mathbb{E}H_{ij}G_{ji}(\underline{G^*} - m_{sc}^*) - m_{sc}^2\mathbb{E}\underline{G}(\underline{G^*} - m_{sc}^*) \\
&= -m_{sc}N^{-1}\sum_{s=1}^{\ell}\frac{1}{s!}\sum_{ij}\mathcal{C}_{s+1}(H_{ij})\mathbb{E}\partial_{ij}^s(G_{ji}(\underline{G^*} - m_{sc}^*)) - m_{sc}^2\mathbb{E}\underline{G}(\underline{G^*} - m_{sc}^*) \\
&=: \sum_{s=1}^{l} L_s^{(1)} - m_{sc}^2\mathbb{E}\underline{G}(\underline{G^*} - m_{sc}^*)\,,
\end{aligned}
\tag{7.4}
$$

where in the third step we used Lemma 6.11 and we ignored the remainder term. Now $L_1^{(1)} =$

$$
\begin{aligned}
&-m_{sc}N^{-1}\sum_{ij}\frac{1 + \delta_{ij} + \delta_{ij}(a_2 - 2)}{N}\mathbb{E}\big[-(1 + \delta_{ij})^{-1}\big((G_{ii}G_{jj} + G_{ji}G_{ji})(\underline{G^*} - m_{sc}^*) + 2G_{ji}N^{-1}(G^{*2})_{ji}\big)\big] \\
&= m_{sc}N^{-2}\sum_{ij}\mathbb{E}\big[\big((G_{ii}G_{jj} + G_{ji}G_{ji})(\underline{G^*} - m_{sc}^*) + 2G_{ji}N^{-1}(G^{*2})_{ji}\big)\big] \\
&\quad + (a_2 - 2)m_{sc}N^{-2}\sum_{i}\mathbb{E}(2G_{ii}^2(\underline{G^*} - m_{sc}^*) + 2N^{-1}G_{ii}(G^{*2})_{ii}) =: L_{1,1}^{(1)} + L_{1,2}^{(1)}\,.
\end{aligned}
$$

We have

$$
\begin{aligned}
&L_{1,1}^{(1)} - m_{sc}^2\mathbb{E}\underline{G}(\underline{G^*} - m_{sc}^*) \\
&= m_{sc}\mathbb{E}(\underline{G}^2 + N^{-1}\underline{G^2})(\underline{G^*} - m_{sc}^*) + m_{sc}N^{-2}\mathbb{E}\underline{GG^{*2}} - m_{sc}^2\mathbb{E}\underline{G}(\underline{G^*} - m_{sc}^*) \\
&= m_{sc}\mathbb{E}\underline{G}|\underline{G} - m_{sc}|^2 + m_{sc}N^{-1}\mathbb{E}\underline{G^2}(\underline{G^*} - m_{sc}^*) + m_{sc}N^{-2}\mathbb{E}\underline{GG^{*2}} \\
&= m_{sc}^2\mathbb{E}|\underline{G} - m_{sc}|^2 + m_{sc}m_{sc}'N^{-1}\mathbb{E}(\underline{G^*} - m_{sc}^*) + m_{sc}N^{-2}\mathbb{E}\left[\frac{\underline{G} - \underline{G^*}}{(z - z^*)^2} - \frac{\underline{G^{*2}}}{z - z^*}\right] + O(N^{-3-10\varepsilon}) \\
&= m_{sc}^2\mathbb{E}|\underline{G} - m_{sc}|^2 + m_{sc}m_{sc}'N^{-1}\mathbb{E}(\underline{G^*} - m_{sc}^*) + m_{sc}N^{-2}\mathbb{E}\left[\frac{m_{sc} - m_{sc}^*}{(z - z^*)^2} - \frac{\overline{m_{sc}'}}{z - z^*}\right] + O(N^{-3-10\varepsilon}) \\
&=: m_{sc}^2\mathbb{E}|\underline{G} - m_{sc}|^2 + L_{1,1,1}^{(1)} + L_{1,1,2}^{(1)} + O(N^{-3-10\varepsilon})
\end{aligned}
$$

Here in the second step we used a cancellation between the first and last term, and in the third step we used (6.17) and

$$
\underline{G^2} - m_{sc}' \prec \frac{1}{N\eta^2}\,,
$$

which is an easy consequence of Theorem 6.5 and Cauchy integral formula. Similarly, by Theorem 6.5 we have

$$
\begin{aligned}
L_{1,2}^{(1)} &= (a_2 - 2)\Big(m_{sc}^3N^{-1}\mathbb{E}(\underline{G^*} - m_{sc}^*) + N^{-2}m_{sc}^2m_{sc}'\Big) + O(N^{-5/2+10\varepsilon}) \\
&=: L_{1,2,1}^{(1)} + L_{1,2,2}^{(1)} + O(N^{-5/2-10\varepsilon})
\end{aligned}
$$

Plugging in the results for $L_{1,1}^{(1)}$ and $L_{1,2}^{(1)}$ into (7.4) yields the *self-consistent equation*

$$\begin{aligned}\mathbb{E}|\underline{G}-m_{sc}|^2 &= (1-m_{sc}^2)^{-1}\Big(L_{1,1,1}^{(1)}+L_{1,1,2}^{(1)}+L_{1,2,1}^{(1)}+O(N^{-5/2+10\varepsilon})+\sum_{s=2}^{l}L_s^{(1)}\Big)\\ &= (1-m_{sc}^2)^{-1}\Big(L_{1,1,1}^{(1)}+L_{1,1,2}^{(1)}+L_{1,2,1}^{(1)}+\sum_{s=2}^{l}L_s^{(1)}\Big)+O(N^{-5/2+11\varepsilon})\,,\end{aligned} \tag{7.5}$$

where in the second step we used

$$|1-m_{sc}^2|^{-1}\leqslant \frac{1}{\sqrt{|E^2-4|+\eta}}\,.$$

In (7.5), $L_{1,1,1}^{(1)}$ corresponds to the second and third term on RHS of (7.1), $L_{1,1,2}^{(1)}$ corresponds to the first term on RHS of (7.2), $L_{1,2,1}^{(1)}$ corresponds to the fourth term on RHS of (7.1), and $L_{1,2,2}^{(2)}$ corresponds to the second term on RHS of (7.2).

The term $L_2^{(1)}$ is negligible for linear eigenvalue statistics, as by (6.16), the indices i and j will appear odd many times in all the resulting terms, which leads to smaller bounds. Vaguely speaking, we have

$$\sum_{ij}G_{ij}^2=\operatorname{Tr}G^2\sim N \quad \text{while} \quad \sum_{ij}G_{ij}^3\sim\sum_{i,j}\delta_{ij}=N\,.$$

Of course, to prove that $L_2^{(1)}$ is indeed small needs some work (which means more expansions). We omit the details here.

In $L_3^{(1)}$, there will be terms that only contain the diagonal entries of G, i.e.G_{ii} and G_{jj}. These are again leading terms which contribute to the last terms on (7.1) and (7.2).

Remark 7.2. (i) By far we have talked about how to compute $\mathbb{E}|\underline{G}-m_{sc}|^2$, but actually the argument works for $\mathbb{E}(\underline{G}-m_{sc})^n(\underline{G^*}-m_{sc}^*)^m$ for all $m,n\geqslant 1$. If we can computed, then we get $N\eta(\underline{G}-m_{sc})$ converges to a complex Gaussian random variable (complex Gaussian distributions are uniquely determined by moments).

(ii) One can then further use (7.3) and compute the arbitrary moment of the linear statistics, and conclude the proof of Theorem 7.1. This would bring an extra difficulty of involving Green functions with different z. Nevertheless, the computation follows the same spirit.

(iii) We will give no more details for how to prove Theorem 7.1. One reason is that to not make the whole thing too computational-oriented. A more important reason is, if you are interested, with the information we provide in Section 6, and Section 7 so far, you can actually proceed the computation yourself: although it is about some current research topic, the tool we have been using are quite elementary (calculus, linear algebra, complex variables...). I trust that motivated readers have the ability to go on their own.

You are also encouraged to try computing $\mathbb{E}|\underline{G}-m_{sc}|^2$ for the complex case $\beta=2$ using Lemma 6.22, and see how the variance depends on β.

7.2. Convergence rate

Let us discuss the convergence rate of the linear statistics. Before diving into the details, let us look at a few examples. For a random variable X, we abbreviate $\langle X\rangle := X-\mathbb{E}X$. Let

$$h_{ij}:=\sqrt{N}H_{ij}\,,$$

so that $(h_{ij})_{1\leqslant i,j\leqslant N}$ are order 1 random variables.

Example 7.3. Recall the Kolmogorov-Smirnov distance Δ from (3.9).

(i) Let $f(x) = x$. Then

$$\operatorname{Tr} f(H) = \operatorname{Tr} H = \frac{h_{11} + \cdots + h_{NN}}{\sqrt{N}} \xrightarrow{d} \mathcal{N}(0, \sigma_1^2) =: Z_1 \,.$$

By Theorem 3.11,

$$\Delta(\operatorname{Tr} f(H), Z_1) = O(N^{-1/2})\,.$$

From Remark 3.19, we know that the above rate is sharp. However, one may argue that it is a special case, as only the diagonal entries of H are involved.

(ii) Let $f(x) = x^2$. Then

$$\operatorname{Tr} f(H) = \operatorname{Tr} H^2 = \sum_{ij=1}^{N} \frac{h_{ij}^2}{N}\,.$$

It is a sum of $O(N^2)$ number of random variables, each of size N^{-1}. Thus we should apply the CLT and Berry-Esseen bound with $n = N^2$. This leads to

$$\langle \operatorname{Tr} f(H)\rangle := \operatorname{Tr} f(H) - \mathbb{E} \operatorname{Tr} f(H) \xrightarrow{d} \mathcal{N}(0, \sigma_2^2) =: Z_2\,, \quad \text{and} \quad \Delta(\langle \operatorname{Tr} f(H)\rangle, Z_2) = O(N^{-1})\,.$$

So one may think that only the trivial case gives a weak rate $O(N^{-1/2})$, and all other cases have the stronger rate $O(N^{-1})$. However, this is not the case.

(iii) Let $f(x) = x^3$. Then

$$\operatorname{Tr} f(H) = \operatorname{Tr} H^3 = N^{-3/2} \sum_{ijk} h_{ij} h_{jk} h_{ki} = N^{-3/2} \sum_{ij} h_{ij} h_{ji} h_{ii} + \cdots$$

Note that the SLLN says $N^{-1}\sum_j h_{ij}h_{ji} \xrightarrow{a.s.} 1$, thus $\operatorname{Tr} H^3$ again contains the contribution $N^{-1/2}\sum_i h_{ii} = \operatorname{Tr} H$, which converges to Z_1 with rate $O(N^{-1/2})$. In fact, this slow rate appear in all $\operatorname{Tr} H^{2k+1}$ for $k \in \mathbb{N}$.

The above example shows that there is a "trace part" of the linear statistics which contributes to the slow convergence rate $O(N^{-1/2})$. We are then motivated to identify this slow converging part explicitly and remove it, in order to get a universal rate $O(N^{-1})$.

For a test function f, define

$$\tau_f := \frac{1}{\pi}\int_{-\pi}^{\pi} f(2\cos\theta)\cos\theta \mathrm{d}\theta = \frac{1}{2\pi}\int_{-2}^{2} \frac{f(x)x}{\sqrt{4-x^2}}\mathrm{d}x\,,$$

and we set the parameter

$$\mathcal{X} \equiv \mathcal{X}(f,\gamma,H) := \begin{cases} 0 & \text{if} \quad (1-\gamma)\tau_f \mathbb{E}h_{11}^3 = 0 \\ 1 & \text{otherwise}\,. \end{cases} \tag{7.6}$$

Recall μ_f, σ_f from (7.1) and (7.2). For $\gamma \in \mathbb{R}$, define

$$\mu_{f,\gamma} = -\frac{a_2(1-\gamma)^2 - a_2}{2\pi}\int_{-2}^{2} f(x)\frac{2-x^2}{\sqrt{4-x^2}}\mathrm{d}x\,, \quad \text{and} \quad \sigma_{f,\gamma}^2 = \frac{a_2(1-\gamma)^2 - a_2}{4\pi^2}\Big(\int_{-2}^{2} f(x)\frac{x}{\sqrt{4-x^2}}\mathrm{d}x\Big)^2\,.$$

Recall the Kolmogorov-Smirnov distance from (3.9).

Since the result for linear function $f(x) = ax + b$ follows from Theorem 3.11, we exclude this trivial case from our discussion. We have the following result.

Theorem 7.4. *Fxe $\gamma \in \mathbb{R}$. Let $f \in C^5(\mathbb{R})$ be independent of N and suppose that f is not linear. Consider the shifted linear statistics*

$$\mathcal{Z}_{f,\gamma} := \frac{\operatorname{Tr} f(H) - \mu_{f,\gamma} - \frac{1}{2}\gamma\tau_f \operatorname{Tr} H}{\sigma_{f,\gamma}} . \tag{7.7}$$

Let $Z \overset{d}{=} \mathcal{N}(0,1)$. Then for any $\varepsilon > 0$, there exists constant $C_{f,\varepsilon} > 0$ such that

$$\Delta(\mathcal{Z}_{f,\varepsilon}, Z) \leqslant C_{f,\varepsilon}(\mathcal{X} N^{-1/2+\varepsilon} + N^{-1+\varepsilon}) .$$

Remark 7.5. From the definition of $\mathcal{X}$ in (7.6), the conditions to have the rate $O(N^{-1+\varepsilon})$ is three-fold. First, the slow rate $O(N^{-1/2+\varepsilon})$ comes from the diagonal part $\frac{1}{2}c_1^f \operatorname{Tr} H$ of the linear statistics. Once we fully subtract this term from $\operatorname{Tr} f(H)$, i.e. when $\gamma = 1$, the remaining part of the LES will have a unified $O(N^{-1+\kappa})$ convergence rate.

Second, in case $\gamma \neq 1$ but the test function f satisfies $c_1^f = 0$, the rate is again $O(N^{-1+\varepsilon})$. Especially, it recovers the rate for $\operatorname{Tr} f(H)$ in the toy case $f(x) = x^2$.

Third, if $\gamma \neq 1$ and $c_1^f \neq 0$, the diagonal part $\frac{1}{2}c_1^f \operatorname{Tr} H$ will play a role in the linear statistics. The object $\operatorname{Tr} H$, is simply a sum of i.i.d. random variable, and in general it has a slow convergence rate $O(N^{-1/2})$ towards Gaussian distribution. This is true even if $\mathbb{E}h_{11}^3 = 0$, as it was observed for $\mathbb{P}(h_{11} = 1) = \mathbb{P}(h_{11} = -1) = \frac{1}{2}$ in Remark 3.19. However, Theorem 7.4 shows that if $\mathbb{E}h_{11}^3 = 0$, the convergence rate of our shifted linear statistics will still degenerate to $O(N^{-1+\kappa})$. This is due to the fact that the Gaussianity of the off-diagonal distribution $\operatorname{Tr} f(H) - \frac{1}{2}c_1^f \operatorname{Tr} H$ can further smooth out the difference between the distribution of $\frac{1}{2}c_1^f \operatorname{Tr} H$ and Gaussian, as long as $\mathbb{E}H_{11}^3 = 0$.

We also have a companion result on the lower bound. Let us denote

$$\mathring{\mathcal{Z}}_{f,\gamma} := \frac{\operatorname{Tr} f(H) - \frac{1}{2}\gamma c_1^f \operatorname{Tr} H - \mathbb{E} \operatorname{Tr} f(H)}{\sqrt{\operatorname{Var}(\operatorname{Tr} f(H) - \frac{1}{2}\gamma c_1^f \operatorname{Tr} H)}} .$$

For the lower bound of the convergence rate, we study the above quantity with mean 0 and variance 1, instead of $\mathcal{Z}_{f,\gamma}$ in (7.7). Otherwise, one needs to exclude the possibility that the bias of the centralization or the scaling may be responsible for the lower bound of the convergence rate.

Theorem 7.6. *Let us adopt the assumptions of Theorem 7.4. We have*

$$\mathbb{E}\mathring{\mathcal{Z}}_{f,\gamma}^3 = (\operatorname{Var}(\operatorname{Tr} f_\gamma(H)))^{-\frac{3}{2}} \Big(-\frac{1}{8}(1-\gamma)^3 \tau_f^3 \mathbb{E}h_{11}^3 N^{-\frac{1}{2}} + \tau'_{f,\gamma,\beta} N^{-1} \Big) + O(N^{-\frac{3}{2}}) . \tag{7.8}$$

Here $\tau'_{f,\gamma,\beta}$ is a constant which is non-zero for most f [8]*. As a consequence, for any fixed $\varepsilon > 0$, we have*

$$\Delta(\mathring{\mathcal{Z}}_{f,\gamma}, Z) \geqslant C'_{f,\kappa}\Big(\mathcal{X} N^{-1/2-\varepsilon} + |\tau'_{f,\gamma,\beta}| N^{-1-\varepsilon}\Big) \tag{7.9}$$

for some constant $C'_{f,\varepsilon} > 0$.

We will only say a few words about the proof of Theorem 7.4: it makes use of Lemma 3.15, the Esseen's inequality. We get the rate of $O(N^{-1+\varepsilon})$ by estimating the characteristic function of $\mathcal{Z}_{f,\gamma}$ for $T = N^{1-\varepsilon}$. Of course, the detailed computation involves the usual Green function approach, cumulant expansion, as well as Lemma 6.19. In practice, Theorem 7.4 should hold even with the rate $O(N^{-1})$, but we were unable to estimate the characteristic function for $t \asymp N$.

[8] The detailed formula of $\tau'_{f,\gamma,\beta}$ can be found in [2, Theorem 1.5].

7.3. Mesoscopic statistics

In Sections 7.1 – 7.2, we consider the object $\operatorname{Tr} f(H)$, where f is a function independent of N. One can imagine that for most of those f, all eigenvalues of H are relevant in the linear eigenvalue statistics. In other words, $\operatorname{Tr} f(H)$ concerns the *global* spectrum, and it is thus called *macroscopic* statistics.

Similar to the local law, we also have the "local" version of the linear statistics. Let f be a function independent of N, and we assume that f vanishes at infinity. Let $E \in R$ and $\eta > 0$. The *mesoscopic linear statistics* consider the fluctuation of the object

$$\operatorname{Tr} f_\eta(H) := \operatorname{Tr} f\Big(\frac{H-E}{\eta}\Big)$$

where $\eta \geqslant N^{-1+\tau}$ for some fixed $\tau > 0$. We have the following result.

Theorem 7.7. *Fix $\tau > 0$. Let $f \in C^2(\mathbb{R})$ and $f(x) = O(|x|^{-2})$ uniformly for all $x \in \mathbb{R}$. Let $E \in [-2+\tau, 2-\tau]$. Then*

$$\operatorname{Tr} f_\eta(H) - \int_{\mathbb{R}} N f_\eta(x) \varrho_{sc}(x) \mathrm{d}x \xrightarrow{d} \mathcal{N}\big(0, V(f)\big) \tag{7.10}$$

for any $\eta \in [N^{-1+\tau}, N^{-\tau}]$. Here

$$V(f) \equiv V(f,\beta) := \frac{1}{2\beta\pi^2} \int_{\mathbb{R}^2} \Big(\frac{f(x)-f(y)}{x-y}\Big)^2 \mathrm{d}x\mathrm{d}y\,.$$

When E is ± 2, the same result holds for any $\eta \in [N^{-2/3+\tau}, N^{-\tau}]$.

Notice that the RHS of (7.10) is independent of η and E: it gives the same result as long as we are in the mesoscopic scale. To see why this is the case, simply observe that

$$V(f_\eta) = \frac{1}{2\beta\pi^2} \int_{\mathbb{R}^2} \Big(\frac{f((x-E)/\eta) - f((y-E)/\eta)}{x-y}\Big)^2 \mathrm{d}x\mathrm{d}y = \frac{1}{2\beta\pi^2} \int_{\mathbb{R}^2} \Big(\frac{f(x)-f(y)}{x-y}\Big)^2 \mathrm{d}x\mathrm{d}y = V(f).$$

Comparing $V(f)$ and σ_f^2 in Theorem 7.1, we see that in the mesoscopic scale, we no longer have the contribution from the diagonal entries of the fourth cumulant. In addition, as x and y are very close in the expression of $V(f_\eta)$, the factor

$$\frac{4-xy}{\sqrt{4-x^2}\sqrt{4-y^2}}$$

in the first term of σ_f^2 is also gone. You can imitate the argument in Section 7.1 yourself to see why this is the case.

As a special class of test functions, we also have the CLT for the trace of the Green functions.

Theorem 7.8. *Fix $\tau > 0$. Let $G(z) = (H-z)^{-1}$. Assume $z = E + \mathrm{i}\eta$, $E \in [-2+\tau, 2-\tau]$, and $\eta \in [N^{-1+\tau}, N^{-\tau}]$. Then*

$$N\eta(\underline{G} - m_{sc}) \xrightarrow{d} \mathcal{N}_{\mathbb{C}}\Big(0, \frac{1}{2\beta}\Big)\,.$$

When E is ± 2, the same result holds for any $\eta \in [N^{-2/3+\tau}, N^{-\tau}]$.

There is one more thing that we can talk about in the mesoscopic case. In Theorems 7.7 and 7.8, we cover the cases $\mathbb{E}H_{ij}^2 = N^{-1}$ (real symmetric, $\beta = 1$), and $\mathbb{E}H_{ij}^2 = 0$ (complex Hermitian case, $\beta = 2$). It is then natural consider the case $\mathbb{E}H_{ij}^2 = aN^{-1}$ for $a \in [0,1]$: let H_1 be a real symmetric Wigner matrix and let H_2 be an independent complex Hermitian Wigner matrix. What is the mesoscopic behavior of

$$H = \sqrt{a}H_1 + \sqrt{1-a}H_2?$$

At what value of a will the result transits from GUE statistics to GOE statistics? What happens at the transition period?

Surprisingly, the next result shows the transition a depends on the scale η you look at! It happens sharply at $1 - a \sim \eta$ (implicitly, this means a depends on N).

Theorem 7.9. *Fix $\tau > 0$. Let $f \in C^2(\mathbb{R})$ and $f(x) = O(|x|^{-2})$ uniformly for all $x \in \mathbb{R}$. Let $E \in [-2+\tau, 2-\tau]$ and $\eta \in [N^{-1+\tau}, N^{-\tau}]$. Assume $\mathbb{E}H_{ij}^2 = aN^{-1}$ for all $i \neq j$, where $a \in [0,1]$. Suppose*

$$\frac{\sqrt{4-E^2}(1-a)}{\eta} \longrightarrow b \in [0,\infty]$$

as $N \to \infty$, then

$$\operatorname{Tr} f\Big(\frac{H-E}{\eta}\Big) - \mathbb{E}\operatorname{Tr} f\Big(\frac{H-E}{\eta}\Big) \xrightarrow{d} \mathcal{N}\big(0, V_b(f)\big)$$

as $N \to \infty$. Here

$$V_b(f) := \frac{1}{4\pi^2}\int (f(x)-f(y))^2\bigg(\frac{1}{(x-y)^2} + \frac{(x-y)^2 - b^2}{\big((x-y)^2+b^2\big)^2}\bigg)\,\mathrm{d}x\,\mathrm{d}y\,.$$

Theorem 7.9 shows that we have the GOE statistics only when $1 - a \ll \eta$, i.e. the matrix H is very close to a real symmetric one. As long as $1 - a \gg \eta$, we jump to the GUE statistics. This is quite surprising: the component $\sqrt{1-a}H_2$ starts to dominate when $\sqrt{1-a} \gg \sqrt{\eta}$.

The underling reason for this phenomenon can be understood from the perspective of *Dyson Brownian Motion*. Let $(\lambda_i(0))_{i=1}^N$ be the eigenvalues of a Hermitian matrix H_0. Consider the system of stochastic ODE

$$\mathrm{d}\lambda_i = \sqrt{\frac{2}{N\beta}}\mathrm{d}B_i - \lambda_i \mathrm{d}t + \frac{1}{N}\sum_{j\neq i}\frac{\mathrm{d}t}{\lambda_i - \lambda_j}$$

for $i = 1, 2, ..., N$, where $\beta = 1, 2$, and B_i are standard Brownian motions. In this case, $(\lambda_i(t))_{i=1}^N$ are the eigenvalues of

$$H_t := \sqrt{\mathrm{e}^{-t}}H_0 + \sqrt{1-\mathrm{e}^{-t}}V \approx \sqrt{1-t}H_0 + \sqrt{t}V \quad \text{if} \;\; t \ll 1\,.$$

Here V is independent of H_0, and it is GOE for $\beta = 1$, and V is GUE for $\beta = 2$. In the case that H_0 is a real symmetric Wigner matrix and V is GUE, at the spectral scale $N^{-1} \leqslant \eta \ll 1$, the Dyson Brownian Motion reaches equilibrium whenever $t \gg \eta$.

7.4. Microscopic statistics

I think anyone who reads to this point can imagine that for Wigner matrices, it is very hard to compute anything about the local eigenvalue statistics, i.e. things happen on scale N^{-1}. However, more can be done for Gaussian matrices. Let $\mu_1 \geqslant \cdots \geqslant \mu_N$ be the eigenvalues of GOE/GUE, and we can show that the joint eigenvalue density is explicitly given by

$$p_N(\mu_1, ..., \mu_N) = \frac{1}{Z_{\beta,N}} \prod_{i<j} |\mu_i - \mu_j|^\beta \mathrm{e}^{-\frac{\beta N}{4} \sum_k \mu_k^2} \tag{7.11}$$

(as always, $\beta = 1$ corresponds to GOE. and $\beta = 2$ corresponds to GUE), direct computation is possible for this special case. Here are a few things that we know about Gaussian Ensembles. The two-point correlation function is defined as

$$p_N^{(2)}(x_1, x_2) = \frac{N!}{(N-2)!} \int_{\mathbb{R}^{N-2}} p_N(x_1, x_2, \mu_3 ..., \mu_N) \mathrm{d}\mu_3 \cdots \mathrm{d}\mu_N ,$$

which can thought as the eigenvalue density of the matrix at points x_1, x_2. As we are interested the local behavior, we define the 2-point correlation function as

$$p_E(u, v) := \frac{1}{(N\varrho_{sc}(E))^2} p_N^{(2)}\Big(E + \frac{u}{N\varrho_{sc}(E)}, E + \frac{v}{N\varrho_{sc}(E)}\Big) .$$

Theorem 7.10. *Fix $\tau > 0$. For GOE and GUE, we have $\lim_{N\to\infty} p_E(u, v) = Y_\beta(u - v)$ for all $E \in [-2 + \tau, 2 - \tau]$, where*

$$Y_2(u) = -s(u)^2 := -\Big(\frac{\sin \pi u}{\pi u}\Big)^2 , \quad \textit{and} \quad Y_1(u) := -s'(u) \int_u^\infty s(v)\mathrm{d}v - s(u)^2 . \tag{7.12}$$

We can also talk about the fluctuation of individual eigenvalues. Let us start with bulk ones.

Theorem 7.11. *Fix $\tau > 0$. Let γ_i be the typical location of μ_i on the semicircle. Then for all $i \in [\tau N, (1 - \tau)N]$, we have*

$$\frac{\mu_i - \gamma_i}{\sqrt{\frac{4 \log N}{(4-\gamma_i^2)\beta N^2}}} \xrightarrow{d} \mathcal{N}(0, 1) . \tag{7.13}$$

The extra $\sqrt{\log N}$ in the fluctuation of μ_i can roughly be understood through the "pushing" of its massive number of right and left eigenvalues. For edge eigenvalues, we no longer have this extra factor.

Theorem 7.12. *We have*

$$N^{2/3}(\mu_1 - 2) \xrightarrow{d} \mathrm{TW}_\beta ,$$

where TW_β is the Tracy-Widom distribution [12]. Similar results holds for all μ_k if $\min\{k, N - k\}$ is bounded.

We also have results on the eigenvalue fluctuations near the edge.

Theorem 7.13. *Let $i \to \infty$ as $N \to \infty$. Then*

$$c\frac{\mu_i - \gamma_i}{(\log i)^{1/2} N^{-2/3} i^{-1/3}} \xrightarrow{d} \mathcal{N}(0, 1) ,$$

where $c = (3/2)^{1/3}\pi\beta^{1/2}$.

All the above results are essentially explicit computations based on (7.11). Luckily, we have the following generalization.

Theorem 7.14. *Theorems 7.10 – 7.13 remain valid for Wigner matrices having the same symmetry class. In addition, if we denote the eigenvalues of a Wigner matrix by $\lambda_1 \geqslant \ldots \geqslant \lambda_N$. Fix $\tau > 0$. For all $i \in [\tau N, (1-\tau)N]$ and $f \in \mathbb{C}_c^\infty$, we have*

$$\lim_{N\to\infty} \mathbb{E}[f(N\varrho_{sc}(\gamma_i))(\mu_i - \mu_{i+1}) - f(N\varrho_{sc}(\gamma_i)(\lambda_i - \lambda_{i+1}))] = 0\,.$$

The actual formula of the gap distribution is quite nasty and we omit here. You can think that they are approximately $\frac{\pi s}{2}\mathrm{e}^{-\pi s^2/4}$ for $\beta = 1$ and $\frac{32 s^2}{\pi^2}\mathrm{e}^{-4s^2/\pi}$ for $\beta = 2$.[9]

We conclude this Section on some rough ideas about how Theorem 7.14 was proved. Let H be a real symmetric Wigner matrix with eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_N$. The proof goes in three steps.

(i) Local density law: for all Wigner matrices we have

$$\max_i |\lambda_i - \gamma_i| = O(N^{-c}) \tag{7.14}$$

for some fixed $c > 0$.

(ii) Let V be a GOE independent of H. How that for any fixed $\varepsilon > 0$, $\sqrt{1 - N^{-1+\varepsilon}}H + N^{-1/2+\varepsilon}V$ has the same microscopic statistics as GOE.

(iii) Remove the small GOE component.

Here Step (i) is what we discussed in Section 6. For Step (ii), we consider the Dyson Brownian motion

$$\mathrm{d}\lambda_i(t) = \sqrt{\frac{2}{N}}\mathrm{d}B_i - \lambda_i(t)\mathrm{d}t + \frac{1}{N}\sum_{j\neq i}\frac{\mathrm{d}t}{\lambda_i(t) - \lambda_j(t)}$$

for $i = 1, 2, ..., N$. Here $\lambda_i(0) = \lambda_i$. Then $(\lambda_1(t), ..., \lambda_N(t))$ are the eigenvalues of

$$H_t := \sqrt{\mathrm{e}^{-t}}H + \sqrt{1 - \mathrm{e}^{-t}}V \approx \sqrt{1-t}H + \sqrt{t}V \quad \text{if} \quad t \ll 1\,.$$

Provided that initially the particles are close to their typical location (7.14), it can be shown that for spectral statistics at scale η, H_t will reach equilibrium $H_\infty = V$ as long as $t \gg \eta$. This shows that H and $\widetilde{H} := \sqrt{1 - N^{-1+\varepsilon}}H + N^{-1/2+\varepsilon}V$ has the same microscopic statistics.

Step (iii) is a Lindberg replacement method, similar to the second proof of Theorem 3.1. Recall in Theorem 3.1, we replaced N random variables $X_1, ..., X_n$ one by one by standard Gaussian, using a matching of their first two moments. Here we need to replace $O(N^2)$ many random variables, namely $(H_{ij})_{1\leqslant i\leqslant j\leqslant N}$. The replacement works since in addition to the first two moment matching, we also have

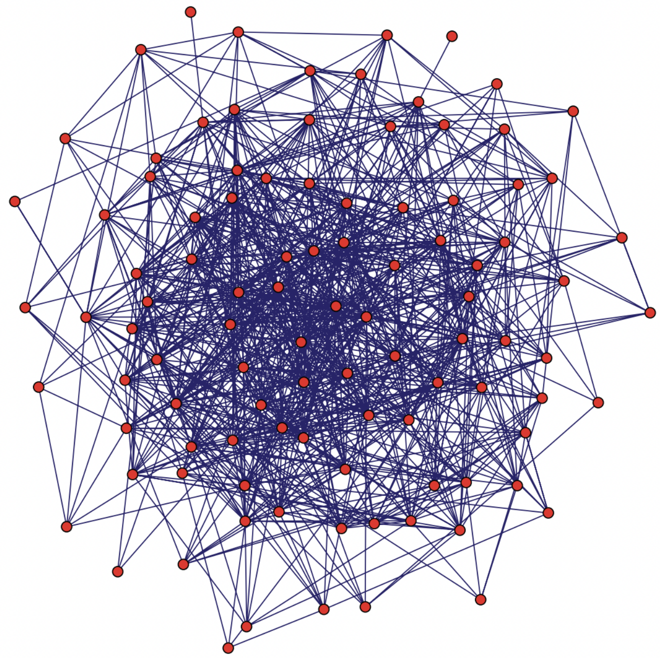
$$\mathbb{E}\widetilde{H}_{ij}^3 = \mathbb{E}H_{ij}^3 + O(N^{-5/2+\varepsilon}) \quad \text{and} \quad \mathbb{E}\widetilde{H}_{ij}^4 = \mathbb{E}H_{ij}^4 + O(N^{-3+\varepsilon})\,.$$

[9] This was calculated by Wigner for 2×2 matrices. So we are basically claiming something like π is roughly 3, which is more than 95% accurate.

8 Collection of other results

8.1. Random graphs

8.1.1. The Erdős-Rényi graph. Arguably the simplest graph model is the Erdős-Rényi graph $\mathcal{G}(N,p)$. It is a simple, undirected graph with N vertexes, and every edge is connected with probability p, completely independent of other edges. Below is a simulation of the model when $N = 100$ and $p = 0.36$.



As in the context of random matrix theory, here N is our fundamental large parameter, and we are interested in the case when N is large. The connecting probability p also could depend on N, and as we shall see, the most interesting case is when $p \to 0$ as $N \to \infty$.

Let $\mathcal{A} = \mathcal{A}^T = \{0,1\}^{N\times N}$ be the adjacency matrix of the graph. It has independent upper triangular entries satisfying

$$\mathcal{A}_{ij} = \begin{cases} 1 & \text{with probability } p\,; \\ 0 & \text{with probability } 1-p\,. \end{cases}$$

For numerical convenience, we consider the rescaled matrix

$$A := \frac{1}{\sqrt{Np(1-p)}}\mathcal{A}$$

so that $\mathrm{Var}(A_{ij}) = N^{-1}$. We further write

$$A = \mathcal{H} + \sqrt{\frac{Np}{1-p}}\,|\mathbf{e}\rangle\langle\mathbf{e}|$$

so that $\mathcal{H}$ is a centered matrix. As $\mathrm{Var}\,\mathcal{H}_{ij} = N^{-1}$, the typical size of $\mathcal{H}_{ij}$ is of order $O(N^{-1/2})$, while

$$\mathbb{E}A_{ij} = \sqrt{\frac{p}{N(1-p)}} \gg N^{-1/2}$$

when $p \ll 1$. One naturally expect that $\mathbb{E}A$ gives a huge change on the spectrum of $\mathcal{H}$. The next result shows that it does not.

Lemma 8.1 (Cauchy interlacing)**.** *Let $\widehat{\lambda}_1 \geqslant \cdots \geqslant \widehat{\lambda}_N$ and $\lambda_1 \geqslant \cdots \geqslant \lambda_N$ be the eigenvalues of A and $\mathcal{H}$ respectively. Then*

$$\widehat{\lambda}_{i+1} \leqslant \lambda_i \leqslant \widehat{\lambda}_i$$

for all $i = 1, 2, ..., N-1$.

Proof. Write $\widehat{G}(z) = (A - z)^{-1}$, $G(z) = (H - z)^{-1}$, and $f = \sqrt{\frac{Np}{1-p}}$. The resolvent identity says

$$G - \widehat{G} = f\widehat{G}|\mathbf{e}\rangle\langle\mathbf{e}|G\,,$$

which implies $\langle \mathbf{e}, G\mathbf{e}\rangle - \langle \mathbf{e}, \widehat{G}\mathbf{e}\rangle = f\langle \mathbf{e}, \widehat{G}\mathbf{e}\rangle \cdot \langle \mathbf{e}, G\mathbf{e}\rangle$. As a result,

$$\langle \mathbf{e}, \widehat{G}\mathbf{e}\rangle^{-1} - \langle \mathbf{e}, G\mathbf{e}\rangle^{-1} = f\,.$$

By spectral decomposition,

$$\bigg(\sum_i \frac{|\langle \widehat{\mathbf{u}}_i, \mathbf{e}\rangle|^2}{\widehat{\lambda}_i - z}\bigg)^{-1} = f + \bigg(\sum_i \frac{|\langle \mathbf{u}_i, \mathbf{e}\rangle|^2}{\lambda_i - z}\bigg)^{-1}$$

Thus $z \in \mathbb{R}$ is a eigenvalue of $\mathcal{H}$ if and only if the RHS of the above is f. In addition, the LHS of the above has N zeros and $N-1$ singularities, and it is decreasing away from the singularities. □

Lemma 8.1 tells that the expectation only shifts the eigenvalues $\lambda_2, ..., \lambda_N$ by a tiny portion, and thus in many situations, studying $\mathcal{H}$ already gives us a lot of information about A. Let examine what properties the entries of $\mathcal{H}$ have. By construction, $\mathbb{E}\mathcal{H}_{ij} = 0$ and $\mathbb{E}\mathcal{H}_{ij}^2 = 1$ for all i, j, which match the case of Wigner matrices. However, for $p \leqslant 1/2$, we have

$$\mathbb{E}\mathcal{H}_{ij}^k = O_k\Big(\frac{1}{N \cdot (Np)^{(k-2)/2}}\Big) = O\Big(\frac{1}{N^{k/2}p^{(k-2)/2}}\Big)$$

for all $k \geqslant 3$. So for $p \ll 1$, the higher moments of $\mathcal{H}$ decays much slower compared to Wigner matrices. This is the key new property that we see for sparse random graphs.

Before we proceed further, we need to answer one question: how small can p get? Certainly p cannot be e^{-N}, where all entries of A will be 0 with very high probability. A slightly more careful examination shows that the threshold is $\log /N$: if $p \leqslant (1-\varepsilon)\log N/N$, then there will be a probability $1 - o(1)$ that the graph has an isolated vertex. In this case, a row and column of A is zero, and A has an eigenvector localized in one atom (see e.g. [10, Proposition 5.9]). This is fundamentally different from the eigenvector delocalization for Wigner matrices Corollary 6.8, and the behavior of the eigenvectors also change. In this notes, we consider the supercritical regime $p \geqslant (1+\varepsilon)\log N/N$, or slightly more relaxing, $p \geqslant N^{-1+\varepsilon}$. We have the following result.

Theorem 8.2. *Fix $\varepsilon > 0$. When $N^{-1+\varepsilon} \leqslant p \leqslant 1/2$, inside the bulk of the spectrum, the gap distribution and 2-point correlations of $\mathcal{H}$ (and A) coincide with those of GOE.*

On the other hand, the fluctuations of individual eigenvalues are not universal.

Theorem 8.3. *Fix $\varepsilon > 0$ and let $N^{-1+\varepsilon} \leqslant p \leqslant N^{-\varepsilon}$. If $i \in [\varepsilon N, (1/2-\varepsilon)N] \cup [(1/2+\varepsilon)N, (1-\varepsilon N)]$, we have*

$$\frac{\widehat{\lambda}_i - \mathbb{E}\widehat{\lambda}}{\gamma_i\sqrt{\frac{1}{2}\mathbb{E}\mathcal{H}_{12}^4}} \xrightarrow{d} \mathcal{N}(0,1)\,. \tag{8.1}$$

If $i/N \to 1/2$, we have

$$\frac{\widehat{\lambda}_i - \mathbb{E}\widehat{\lambda}_i}{\sqrt{\mathbb{E}\mathcal{H}_{12}^4}} \xrightarrow{d} 0\,.$$

Analogue results also hold for λ_i.

Theorem 8.3 shows that the bulk eigenvalues of sparse matrices fluctuate on the scale $\sqrt{\mathbb{E}\mathcal{H}_{12}^4} \asymp \frac{1}{N\sqrt{p}}$. The reason that we have this non-universal contribution is quite simple: $\frac{1}{N\sqrt{p}} \gg \frac{\log N}{N}$ as long as $p \leqslant N^{-\varepsilon}$. In other words, the slow decay of higher moments creates a large oscillation that outscale the $\log N/N$ Gaussian that we see in the GOE case.

The edge statistics of sparse matrices are more dedicated and much more important. We have the following rather trivial estimate on the largest eigenvalue of A.

Lemma 8.4. *Fix $\varepsilon > 0$ and let $N^{-1+\varepsilon} \leqslant p \leqslant 1/2$. We have*

$$\widehat{\lambda}_1 = \sqrt{\frac{Np}{1-p}}(1+o(1))$$

with very high probability.

In particular, the above result shows that the largest eigenvalue of A sits very far away from the rest of the spectrum. One of the prominent problem in graph theory is the distribution of $\lambda_1 - \lambda_2$, the *spectral gap*, which contains the information of the connectivity of the graph. If $\lambda_1 = \lambda_2$, the the graph is disconnected, and if the spectral gap is large, the graph is very-well connected. The spectral gap if Erdős-Rényi graphs is now well understood after a series of works in 2011-2022.

Theorem 8.5. *Fix $\varepsilon > 0$ and let $N^{-1+\varepsilon} \leqslant p \leqslant 1/2$.*

(i) When $N^{-2/3+\varepsilon} \leqslant p \leqslant 1/2$, we have

$$N^{2/3}\big(\widehat{\lambda}_2 - \mathbb{E}\widehat{\lambda}_2\big) \xrightarrow{d} \mathrm{TW}_1\,.$$

In other words, we have edge universality.

(ii) When $N^{-1+\varepsilon} \leqslant p \leqslant N^{-2/3-\varepsilon}$, we have

$$\frac{\widehat{\lambda}_2 - \mathbb{E}\widehat{\lambda}_2}{\sqrt{2\mathbb{E}\mathcal{H}_{12}^4}} \xrightarrow{d} \mathcal{N}(0,1)\,. \tag{8.2}$$

One might now notice the similarities between (8.1) and (8.2): in fact, they have the same source. Let

$$\mathcal{Z} := \frac{1}{N}\operatorname{Tr}\mathcal{H}^2 - 1\,.$$

It is not hard to see that

$$\frac{\mathcal{Z}}{\sqrt{\mathbb{E}\mathcal{H}_{12}^4}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad \text{and} \quad \sqrt{\mathbb{E}\mathcal{H}_{12}^4} \asymp \frac{1}{N\sqrt{p}}\,.$$

What was actually proved is, when $N^{-1+\varepsilon} \leqslant p \leqslant N^{-2/3-\varepsilon}$,

$$\left|\widehat{\lambda}_i - \mathbb{E}\widehat{\lambda}_i - \frac{\gamma_i}{2}\mathcal{Z}\right| \prec \frac{1}{N^{1+\delta}\sqrt{p}}$$

for all $i = 2, 3, ..., N$. Let us interpret the meaning of $\mathcal{Z}$. Recall that H was made from the adjacency matrix A through normalization and centering. So the fluctuation of $\mathcal{Z}$ is proportional to that of

$$\operatorname{Tr}\mathcal{A}^2 = \sum_{ij} \mathcal{A}_{ij}\mathcal{A}_{ji} = \sum_{ij} \mathcal{A}_{ij}\,, \tag{8.3}$$

which is nothing but the total degree of the graph.

Beyond the leading fluctuation, there is still a good interest to identify all the noise random variables and recover the Tracy-Widom distribution. We have the following result.

Theorem 8.6. *For $N^{-1+\varepsilon} \leqslant p \leqslant 1/2$, we can find an explicit random variable $\mathcal{X}$ such that*

$$\lim_{N\to\infty} \mathbb{P}\Big(N^{2/3}\big(\widehat{\lambda}_2 - \mathbb{E}\widehat{\lambda}_2 - \mathcal{X}\big) \leqslant s\Big) = F_1(s)\,.$$

In the above $\mathcal{X}$, the first two order terms are

$$\mathcal{Z} = \frac{1}{N}\Big(\sum_{ij} H_{ij}^2 - \frac{1}{N}\Big)$$

and

$$\mathcal{Z}_2 := \frac{1}{N}\Big(\sum_{ij} H_{ij}^4 - \frac{1}{N^3p^2}\Big) + \frac{2}{N}\sum_{ijk} H_{ij}^2\Big(H_{ik}^2 - \frac{1}{N}\Big)\,.$$

Recall that $\mathcal{Z}$ fluctuates on the scale $\frac{1}{N\sqrt{p}}$. It is not hard to see that $\mathcal{Z}_2$ is on the scale

$$\frac{1}{N^{3/2}p} = \frac{1}{N\sqrt{p}} \cdot \frac{1}{\sqrt{Np}}\,.$$

When $p \ll N^{-5/6}$, $\mathcal{Z}_2$ bigger than the Tracy-Widom law, and thus also relevant. In general, the kth order term in $\mathcal{X}$ is of size

$$\frac{1}{N\sqrt{p}} \cdot \Big(\frac{1}{\sqrt{Np}}\Big)^k\,.$$

If $p = N^{-1+\varepsilon}$ for some small $\varepsilon > 0$, we need to identify the first $O(\varepsilon^{-1})$ order terms in $\mathcal{X}$. This is theoretically possible, but the computation required grows exponentially with k, and by far we have no closed formula for $\mathcal{X}$.

We can also interpret the graphical meaning of $\mathcal{Z}_2$. Let $d_1, ..., d_N$ be the degrees of my graph $G(N, p)$. Similar to (8.3), the first term of $\mathcal{Z}_2$ is proportional to the fluctuation of $\sum_i d_i$, the total degree of the graph. The second term of $\mathcal{Z}_2$ is proportional to the fluctuation of

$$\sum_{ijk} \mathcal{A}_{ij}^2\mathcal{A}_{ik}^2 = \sum_i d_i^2\,.$$

So one might imagine that the noise random variables might be a function of $d_1, ..., d_N$. Unfortunately, this is not the case: the third order term of $\mathcal{X}$ contains the contribution

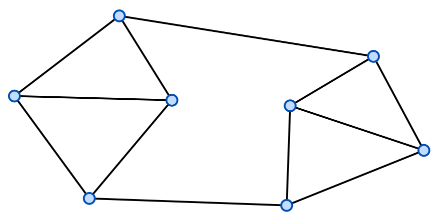
$$\sum_{ijkl} H_{ij}H_{jk}H_{kl} \propto \sum_{jk} d_j\mathcal{A}_{jk}d_k\,,$$

and it cannot be fully described by the graph degrees.

Currently, it is still a mystery what $\mathcal{X}$ (and also $\mathbb{E}\lambda_2$) look like when p is close to N^{-1}. However, there is another graph model that always have the Tracy-Widom law near the edge.

8.1.2. Random regular graph. A d regular graph on N vertices is a simple graph that all vertices have degree d. Given integers N, d, there are finitely many such graphs. A random regular graph $G_d(N)$ is an ensemble that takes a uniform probability measure on all d regular graph on N vertices.

As always, d is a quantity that can depend on N. Naively, we have the range $d \in \{1, 2, ..., N-1\}$. However, the cases $d = 1, 2$ are trivial: the graph is always disconnected. Together with symmetry, we are interested in the regime $3 \leqslant d \leqslant N/2$.



$$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

We have the following result.

Theorem 8.7. *Assume $3 \leqslant d \leqslant N/2$, and let $\widetilde{\mathcal{A}}$ be the adjacency matrix of $G_d(N)$. Let $d = \widetilde{\lambda}_1 \geqslant \widetilde{\lambda}_2 \geqslant \cdots \geqslant \widetilde{\lambda}_N$ be the eigenvalues of $\widetilde{\mathcal{A}}$. We have*

$$\left(\frac{d(d-1)}{(d-2)^2}N\right)^{2/3}\left(\frac{\lambda_2 + d/N}{(d-1)(N-d)/N} - 2\right) \xrightarrow{d} \mathrm{TW}_1$$

as $N \to \infty$.

Theorem 8.7 was proved very recently (December 2024).[10] Although the result has a rather complicated dependence on d, it is at the same time completely explicit and uniform: the fluctuation is universal as long as $d \geqslant 3$. In addition, one can check that the terms in $\mathcal{X}$ that we found out in Theorem 8.6 are all deterministic for random regular graphs.

We close the discussion here with some initial ideas of how to handle this model. It is clear that the cumulant expansion, Lemma 6.11, which can be used for Wigner matrices and Erdős-Rényi graphs, no longer works for random regular graphs: the entries of $\widetilde{\mathcal{A}}$ are highly dependent.

The method that overcomes this obstacle is *switching*, or sometimes called *local resampling*. Here we illustrate the method in the regime $N^{\tau} \leqslant d \leqslant N^{2/3\tau}$ for some fixed $\tau > 0$. In this regime, one can show that the empirical eigenvalue density of $\widetilde{\mathcal{A}}/\sqrt{d-1}$ converges weakly to the semicircle density on $[-2, 2]$. Let us define the Green function via $\widetilde{G}(z) := (\widetilde{A}/\sqrt{d-1} - z)^{-1}$. Then

$$\mathbb{E}\big[\delta_{ij} + z\widetilde{G}_{ij} + \widetilde{G}_{ij}\underline{\widetilde{G}}\big] = \frac{1}{\sqrt{d-1}}\sum_k \mathbb{E}\widetilde{A}_{ik}\widetilde{G}_{kj} + \mathbb{E}\widetilde{G}_{ij}\underline{\widetilde{G}}\,.$$

Similar to the Wigner case, where we needed to compute $\mathbb{E}H_{ij}F(H)$, where F consists of the Green

[10] Technically speaking, the regimes $N^{1/3-\varepsilon} \leqslant d \leqslant N^{2/3+\varepsilon}$ and $C \ll d \ll N^{\varepsilon}$ is still undone, but we believe that they can be proved using the techniques developed in the extreme regimes.

function of H, here we need to compute $\mathbb{E}\widetilde{A}_{ij}F(\widetilde{A})$. It is not hard to show that

$$
\begin{aligned}
\mathbb{E}\widetilde{A}_{ij}F(\widetilde{A}) &= \frac{1}{N-d}\sum_{k}\mathbb{E}\widetilde{A}_{ij}(1-\widetilde{A}_{ik})F(\widetilde{A}) = \frac{1}{(N-d)d}\sum_{kl}\mathbb{E}\widetilde{A}_{ij}(1-\widetilde{A}_{ik})\widetilde{A}_{kl}F(\widetilde{A})\\
&= \frac{1}{(N-d)d}\sum_{kl}\mathbb{E}\widetilde{A}_{ij}(1-\widetilde{A}_{ik})\widetilde{A}_{kl}(1-\widetilde{A}_{jl})F(\widetilde{A}) + \text{error}\\
&= \frac{1}{(N-d)d}\sum_{kl:i,j,k,l\,distinct}\mathbb{E}\widetilde{A}_{ij}(1-\widetilde{A}_{ik})\widetilde{A}_{kl}(1-\widetilde{A}_{jl})F(\widetilde{A}) + \text{error}\,.
\end{aligned}
$$

Here in the third step we made use of the bound $d \leqslant N^{2/3-\tau}$. Now for indices i,j,k,l, we define the signed adjacency matrices

$$
(\Delta_{ij})_{xy} := \delta_{ix}\delta_{jy} + \delta_{iy}\delta_{jx}\,, \quad \xi_{ij}^{kl} = \Delta_{ij} + \Delta_{kl} - \Delta_{ik} - \Delta_{jl}\,.
$$

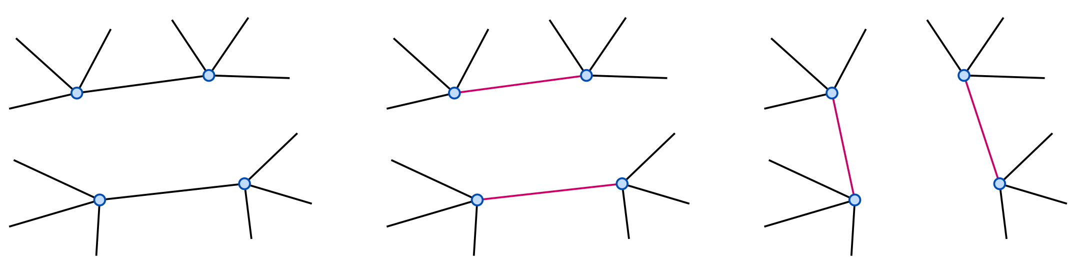
Since we are taking the uniform measure on all d-regular graphs on vertices, we can perform the following switching.

Lemma 8.8. *Let i,j,k,l be distinct indices. Let F be a function which depends on the random graph $\widetilde{\mathcal{A}}$, and possibly on the indices i,j,k,l. We have*

$$
\mathbb{E}\widetilde{A}_{ij}(1-\widetilde{A}_{ik})\widetilde{A}_{kl}(1-\widetilde{A}_{jl})F(\widetilde{\mathcal{A}}) = \mathbb{E}\widetilde{A}_{ik}(1-\widetilde{A}_{ij})\widetilde{A}_{jl}(1-\widetilde{A}_{kl})F(\widetilde{\mathcal{A}}+\xi_{ij}^{kl})\,,
$$

where ξ and χ are defined in (2.5) and (2.6) respectively.

Lemma 8.8 is the fundamental formula that we use to generate non-trivial transformations in our equation, so that the computations concerning the random regular graphs can proceed. To see why it is true, you can refer to the following picture.



Of course, there is still a long way to go before we can prove Theorem 8.7. We do not pursue this here.

8.2. Eigenvector distributions

Recall that a GOE can be written as

$$
H^o = \frac{O+O^T}{\sqrt{2N}}\,,
$$

where $O \in \mathbb{R}^{N\times N}$ is a matrix with i.i.d. standard Gaussian entries. It is easy to see that GOE is orthogonal invariant, meaning

$$
B^T H^o B \overset{d}{=} H^o
$$

for any orthogonal matrix B. Now let $\mathbf{u} \in \mathbb{S}^{N-1}$ be a eigenvector of H^o. Then $B^T\mathbf{u}$ is an eigenvector of $B^T H^o B$. As a result,

$$B\mathbf{u} \stackrel{d}{=} \mathbf{u}$$

for all orthogonal matrix B. In other words,

$$\sqrt{N}\mathbf{u} \stackrel{d}{\approx} \mathbf{g} \tag{8.4}$$

for large N, where $\mathbf{g}$ denotes the standard Gaussian vector in $\mathbb{R}^N$. Now the natural question is, what happens to a general Wigner matrix?

Theorem 8.9. *Let H be a real symmetric Wigner matrix with eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_N$ and associated eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_N \in \mathbb{S}^{N-1}$. Then for any $i, j \in \{1, ..., N\}$, we have*

$$\sqrt{N}\mathbf{u}_i(j) \xrightarrow{d} \mathcal{N}(0,1)$$

as $N \to \infty$.

Beyond entrywise convergence, (8.4) can in fact be verified for Wigner matrices in all deterministic directions.

Theorem 8.10. *Let us adopt the assumptions of Theorem 8.9. Let $\mathbf{v} \in \mathbb{S}^{N-1}$. Then for all $i \in \{1, 2, ..., N\}$, we have*

$$\sqrt{N}\langle \mathbf{v}, \mathbf{u}_i \rangle \xrightarrow{d} \mathcal{N}(0,1)$$

as $N \to \infty$.

We can even go further than Theorem 8.10. Observe

$$\langle \sqrt{N}\mathbf{u}_i, \mathbf{v} \rangle^2 = \langle \sqrt{N}\mathbf{u}_i, M\sqrt{N}\mathbf{u}_i \rangle \stackrel{d}{\approx} \mathcal{N}(0,1)^2\,,$$

where $M = \mathbf{v}\mathbf{v}^T \in \mathbb{R}^{N\times N}$ is of rank 1. What if $M = \sum_k \mu_k \mathbf{v}_k \mathbf{v}_k^T$ is of *high rank*? In this case

$$\langle \sqrt{N}\mathbf{u}_i, B\sqrt{N}\mathbf{u}_i \rangle = \sum_k \mu_k \langle \sqrt{N}\mathbf{u}_i, \mathbf{v}_k \rangle^2\,,$$

and the following result, known as fluctuations of quantum ergodicity, gives a CLT of the CLT.

Theorem 8.11. *Let us adopt the assumptions of Theorem 8.9. Let $M \in \mathbb{R}^{N\times N}$ be deterministic and real-symmetric. Suppose $\operatorname{Tr} M^2 \geqslant N^{\varepsilon}\|M\|^2 > 0$ and M is traceless.Then for all $i \in \{1, 2, ..., N\}$, we have*

$$\frac{1}{\sqrt{2\operatorname{Tr} M^2}}\langle \sqrt{N}\mathbf{u}_i, M\sqrt{N}\mathbf{u}_i \rangle \xrightarrow{d} \mathcal{N}(0,1) \stackrel{d}{\approx} \frac{1}{\sqrt{2\operatorname{Tr} M^2}}\langle \mathbf{g}, M\mathbf{g} \rangle\,. \tag{8.5}$$

The assumption $\operatorname{Tr} M^2 \geqslant N^{\varepsilon}\|M\|^2 > 0$ essentially says M has large, even full rank. The traceless condition on M does nothing but removing the mean $\langle \mathbf{u}_i, I\mathbf{u}_i \rangle = 1$.

One can also talk about the eigenvectors of sparse matrices.

Theorem 8.12. *Let $A, \mathcal{H}$ be as in Section 8.1, with eigenvectors $\widehat{\mathbf{u}}_1, ..., \widehat{\mathbf{u}}_N$ and $\mathbf{u}_1^{\mathcal{H}}, ..., \mathbf{u}_N^{\mathcal{H}}$ respectively. Suppose that $N^{-1+\tau} \leqslant p \leqslant 1/2$ for some fixed $\tau > 0$.*

(i) Theorems 8.9 and 8.10 remain valid for $\mathbf{u}_1^{\mathcal{H}}, ..., \mathbf{u}_N^{\mathcal{H}}$.

(ii) *Theorem 8.9 remain valid for* $\widehat{\mathbf{u}}_1, ..., \widehat{\mathbf{u}}_N$*.*

(iii) *Let* $\mathbf{e} = N^{-1/2}(1, 1, ..., 1) \in \mathbb{S}^{N-1}$ *and* $\mathbf{v} \in \mathbb{S}^{N-1}$ *such that* $\mathbf{e} \perp \mathbf{v}$*. Then*

$$\sqrt{N}\langle \mathbf{v}, \widehat{\mathbf{u}}_i \rangle \xrightarrow{d} \mathcal{N}(0, 1)$$

for all $i \in 2, 3, ..., N$*. In addition,*

$$\max_{i=2,\dots,N} |\langle \mathbf{e}, \widehat{\mathbf{u}}_i \rangle| \prec \frac{1}{N\sqrt{p}} \quad \text{and} \quad 1 - \langle \mathbf{e}, \widehat{\mathbf{u}}_1 \rangle \prec \frac{1}{N^{3/2}p} \,.$$

Theorem 8.12 (iii) shows that, the eigenvectors $\widehat{\mathbf{u}}_2, ..., \widehat{\mathbf{u}}_N$ are almost orthogonal to $\mathbf{e}$. The reason is quite simple:

$$\mathbb{E}A = \sqrt{\frac{Np}{1-p}}\,|\mathbf{e}\rangle\langle\mathbf{e}|$$

is very large, which makes $\widehat{\mathbf{u}}_1$ very close to $\mathbf{e}$.

8.3. The non-Hermitian case

By far we have only talked about Hermitian random matrices, and it is generally perfectly nature to also consider non-Hermitian ones.

Let $Q \in \mathbb{C}^{N\times N}$ be a random matrix with i.i.d. entries, and

$$\mathbb{E}Q_{ij} = 0, \quad \mathbb{E}|Q_{ij}^2| = N^{-1} \quad \text{and} \quad \mathbb{E}|Q_{ij}^k| \leqslant C_k N^{-k/2}$$

for all $k \geqslant 3$. Let θ_1,...,θ_N be the eigenvalue of Q. As the eigenvalues are now in general complex, we need to compute the joint complex moments in order to extract useful information. However,

$$\sum_i \theta_i^m \overline{\theta}_i^n \neq \operatorname{Tr} Q^m \overline{Q}^n \,, \operatorname{Tr} Q^m Q^{*n} \,.$$

The reason is quite simple: $Q, \overline{Q}, Q^*$ do not have the same set of eigenvectors.

It took mathematicians many years to figure out how to study the eigenvalues of non-Hermitian random matrices. In the 1980s, Girko came up with the following simple idea.

Theorem 8.13. *For a non-Hermitian matrix* Q*, let us define its shifted Hermitization by*

$$S^z := \begin{pmatrix} 0 & Q - z \\ Q^* - \overline{z} & 0\,. \end{pmatrix} \in \mathbb{C}^{2N\times 2N}$$

Then for any $f \in C_c^\infty(\mathbb{C})$*, we have*

$$\sum_i f(\theta_i) = \frac{\mathrm{i}}{4\pi}\int_{\mathbb{C}}\int_0^\infty \nabla^2 f(z) \operatorname{Tr}(S^z - \mathrm{i}\eta)^{-1} \mathrm{d}\eta\, \mathrm{d}^2 z \,.$$

Proof. Our starting point is the potential identity $\nabla^2 \log|z| = 2\pi\delta_0$, and as a result,

$$\int_{\mathbb{C}} \log|\theta_i - z| \nabla^2 f(z) \mathrm{d}^2 z = 2\pi f(\theta_i)\,.$$

Let us denote the singular values of $Q - z$ by σ_i^z. We have

$$\sum_i f(\theta_i) = \frac{1}{2\pi} \int_{\mathbb{C}} \sum_i \log|\theta_i - z| \nabla^2 f(z) \mathrm{d}^2 z = \frac{1}{2\pi} \int_{\mathbb{C}} \log \prod_i |\theta_i - z| \nabla^2 f(z) \mathrm{d}^2 z$$

$$= \frac{1}{2\pi} \int_{\mathbb{C}} \log \prod_i \sigma_i^z \nabla^2 f(z) \mathrm{d}^2 z = \frac{1}{2\pi} \int_{\mathbb{C}} \sum_i \log \sigma_i^z \nabla^2 f(z) \mathrm{d}^2 z \,.$$

Now notice that the eigenvalues of S^z are $\pm\sigma_i^z$. Together with

$$\int_0^\infty \frac{1}{\sigma_i^z - \mathrm{i}\eta} + \frac{1}{-\sigma_i^z - \mathrm{i}\eta} \mathrm{d}\eta = 2\mathrm{i} \log \sigma_i^z \,,$$

we have

$$\sum_i \log \sigma_i^z = \frac{1}{2\mathrm{i}} \int_0^\infty \operatorname{Tr}(S^z - \mathrm{i}\eta)^{-1} \mathrm{d}\eta \,.$$

The desired result then follows. □

Theorem 8.13 transfers the study of θ_i to something we are familiar of: the Green function of the Hermitian matrix S^z. This Green function

$$s := \frac{1}{2N}(S^z - \mathrm{i}\eta)^{-1} \,,$$

due to the shift variable z, in general satisfies the cubic equation

$$s^3 + 2\mathrm{i}\eta s^2 + (1 - \eta^2 - |z^2|)s + \mathrm{i}\eta \approx 0 \,.$$

We know many things concerning non-Hermitian models, for instance the following local circular law, which can be thought as an analogue of Theorem 6.5.

Theorem 8.14. *Let $\mu(z) = N^{-1} \sum_i \delta_{\theta_i}(z)$ be the empirical eigenvalue density of Q. For any ball $B \subset \mathbb{C}$, we have*

$$\int_B \mu(z) \mathrm{d}^2 z = \int_B \frac{1}{\pi} \mathbf{1}_{|z| \leqslant 1} \mathrm{d}^2 z + O_\prec(N^{-1}) \,.$$

We conclude by reminding you that most of the results we stated in Sections 6 – 8 can be proved using the simple calculus tools we introduced: you are welcome to try if interested. The topics of random matrices and random graphs are very rich and there are many things one can talk about, but for the matter of this course we will stop here.

References

[1] Andrew Barbour. *Asymptotic Expansions Based on Smooth Functions in the Central Limit Theorem.* Probability Theory and Related Fields, **72** (1986), pages 289–303.

[2] Zhigang Bao, Yukun He. *Quantitative CLT for linear eigenvalue statistics of Wigner matrices.* Annals of Applied Probability, **33** (2023), 5171-5207.

[3] Jordan Bell. *The Berry-Esseen theorem.*

[4] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. *Isotropic local laws for sample covariance and generalized Wigner matrices.* Elect. J. Prob. **19** (2014), no. 33, 54 pages.

[5] Raphaël Cerf and Pierre Petit. *A Short Proof of Cramér's Theorem in* $\mathbb{R}$. The American Mathematical Monthly, **118** (2014), 925–931.

[6] Rick Durrett. *Probability: Theory and Examples.*

[7] P. J. Fitzsimmons. *On a Proof of the Strong Law of Large Numbers.* arXiv preprint: 2111.05766.

[8] Rita Giuliano. *The Local Limit Theorem and the Almost Sure Local Limit Theorem.*

[9] Yukun He and Antti Knowles. *Mesoscopic eigenvalue statistics for Wigner matrices.* Annals of Applied Probability, **27** (2017), 1510-1550.

[10] R. van der Hofstad. *Random graphs and complex networks.* Volume **1**, Cambridge Series in Statistical and Probabilistic Mathematics, 2017

[11] 林正炎， 陆传容， 苏中根。《概率极限理论基础》。

[12] C. A. Tracy and H. Widom. *Level-spacing distributions and the Airy kernel.* Communications in Mathematical Physics, **159** (1994), 151–174.

[13] Roman Vershynin. *A friendly proof of the Berry-Esseen theorem.* arXiv preprint: 2602.06234.

[14] Irina Shevtsova. *On the absolute constants in the Berry–Esseen type inequalities for identically distributed summands.* Dokl. Akad. Nauk **456** (2014), no. 6, 650–654.