Advanced Theory of Probability, Fall 2019
高等概率论笔记

何憾[1]

版本 1.0.0 Beta

[1]访问我的个人主页, 查看其他笔记

Contents

Chapter 1

Indenpendence

Througut this note, we denote by $\mathbb{N}$ all the non-negative integers and by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space we work on.

The measure theory is a "linear" theory that could not describe the dependence structure of events or random variables. We enter the realm of probability theory exactly at this point, where we define independence of events and random variables. *Independence* is a pivotal notion of probability theory, and the computation of dependencies is one of the theory's major tasks.

1.1 Independence of Events

1.1.1 Independence of Events

We consider two events A and B as (stochastically) independent if the occurrence of A does not change the probability that B also occurs. Formally, two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\,. \tag{1.1}$$

Example 1.1 (Rolling a dice twice)**.** Consider the random experiment of rolling a dice twice. Hence $\Omega = \{1, \cdots, 6\}^2$ endowed with the σ -algebra $\mathcal{F} = 2^{\Omega}$ and the uniform distribution $\mathbb{P}$.

(i) Two events A and B should be independent, e.g., if A depends only on the outcome of the first roll and B depends only on the outcome of the second roll. Formally, we assume that there are sets $\tilde{A}, \tilde{B} \subset \{1, \cdots, 6\}$ such that

$$A = \tilde{A} \times \{1, \cdots, 6\} \text{ and } B = \{1, \cdots, 6\} \times \tilde{B}$$

Now we check that A and B are independent. To this end, we compute $\mathbb{P}(A) = \#A/36 = \#\tilde{A}/6$ and $\mathbb{P}(B) = \#B/36 = \#B/6$. Furthermore,

$$\mathbb{P}(A \cap B) = \frac{\#(\tilde{A} \times \tilde{B})}{36} = \frac{\#\tilde{A}}{6} \cdot \frac{\#\tilde{B}}{6} = \mathbb{P}(A) \cdot \mathbb{P}(B) .$$

(ii) Independence can occur also in less obvious situations. For instance, let A be the event where the sum of the two rolls is odd,

$$A = \{(\omega_1, \omega_2) \in \Omega : \omega_1 + \omega_2 \in \{3, 5, 7, 9, 11\}\}$$

and let B be the event where the first roll gives at most a three

$$B = \{(\omega_1, \omega_2) \in \Omega : \omega_1 \in \{1, 2, 3\}\}$$

Although it might seem that these two events are entangled in some way, they are independent. Indeed, it is easy to check that $\mathbb{P}(A) = \mathbb{P}(B) = 1/2$ and $\mathbb{P}(A \cap B) = 1/4$.

What is the condition for three events A_1, A_2, A_3 to be independent? Of course, any of the pairs $(A_1, A_2), (A_1, A_3)$ and (A_2, A_3) has to be independent. However, we have to make sure also that the simultaneous occurrence of A_1 and A_2 does not change the probability that A_3 occurs. Hence, it is not enough to consider pairs only. Formally, we call three events A_1, A_2 and A_3 independent if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) \quad \text{for all } i, j \in \{1, 2, 3\}, i \neq j \tag{1.2}$$

and

$$\mathbb{P}(A_1 \cap A_1 \cap A_3) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)\,\mathbb{P}(A_3) . \tag{1.3}$$

We should emphasize that (1.2) does not imply (1.3), and (1.3) does not imply (1.2).

Example 1.2 (Rolling a dice three times)**.** Consider the random experiment of rolling a three times. Hence $\Omega = \{1, \cdots, 6\}^3$ endowed with the σ -algebra $\mathcal{F} = 2^{\Omega}$ and the uniform distribution $\mathbb{P}$.

(i) (1.2) does not imply (1.3): Consider now the events

$$\begin{aligned} A_1 &:= \{\omega \in \Omega : \omega_1 = \omega_2\} \\ A_2 &:= \{\omega \in \Omega : \omega_2 = \omega_3\} \\ A_3 &:= \{\omega \in \Omega : \omega_1 = \omega_3\} \end{aligned}$$

Then $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{6}$. Furthermore, $\#(A_i \cap A_j) = 6$ if $i \neq j$, hence $\mathbb{P}(A_i \cap A_j) = \frac{1}{36}$. Hence (1.2) holds. On the other hand, we have $\#(A_1 \cap A_2 \cap A_3) = 6$, thus $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{1}{36} \neq \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6}$. Thus (1.3) does not hold and so the events A_1, A_2, A_3 are not independent.

(ii) (1.3) does not imply (1.2). Consider now the events

$$\begin{aligned} A_1 &:= \{1,2,3,4\} \times \{1,2,3\} \times \{1,2,3\} \\ A_2 &:= \{1,3,4,5\} \times \{1,3,4\} \times \{1,3,4\} \\ A_3 &:= \{1,4,5,6\} \times \{1,4,5\} \times \{1,4,5\} \end{aligned}$$

Clearly, $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{6}$. Since $A_1 \cap A_2 \cap A_3 = \{(1,1,1)\}$, (1.3) holds. To check (1.2), note that

$$A_1 \cap A_2 = \{1,3,4\} \times \{1,3\} \times \{1,3\}$$

hence $\mathbb{P}(A_1 \cap A_2) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{18} \neq \frac{1}{6} \times \frac{1}{6}$, (1.2) does not hold.

(iii) If we assume that for any $i = 1,2,3$ the event A_i depends only on the outcome of the i th roll, then the events A_1, A_2 and A_3 are independent. Indeed, as in the preceding example, there are sets $\tilde{A}_1, \tilde{A}_2, \tilde{A}_3 \subset \{1, \cdots, 6\}$ such that

$$\begin{aligned} A_1 &= \tilde{A}_1 \times \{1, \cdots, 6\}^2 \\ A_2 &= \{1, \cdots, 6\} \times \tilde{A}_2 \times \{1, \cdots, 6\} \\ A_3 &= \{1, \cdots, 6\}^2 \times \tilde{A}_3 \end{aligned}$$

The validity of (1.2) follows as in Example 1.1 (i). In order to show (1.3), we compute

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{\#\left(\tilde{A}_1 \times \tilde{A}_2 \times \tilde{A}_3\right)}{216} = \prod_{i=1}^{3} \frac{\#\tilde{A}_i}{6} = \prod_{i=1}^{3} \mathbb{P}(A_i)$$

Definition 1.1. Let I be an arbitrary index set and let $(A_i)_{i\in I}$ be an arbitrary family of events. The family $\{A_i\}_{i\in I}$ is called **independent** if for any finite subset $J \subset I$ the product formula holds:

$$\mathbb{P}(\cap_{j\in J} A_j) = \prod_{j\in J} \mathbb{P}(A_j) \ . \tag{1.4}$$

$\{A_i\}_{i\in I}$ is called **pairwise independent** if for any distinct $i, j \in I$, A_i and A_j are independent.

Remark. We can see that a family of events $\{A_i\}_{i\in I}$ is independent iff for any finite subset $J \subset I$, $\{A_i\}_{i\in J}$ is independent.

Proposition 1.1. *$\{A_i\}_{i\in I}$ is independent iff the family of r.v.'s $\{1_{A_i}\}_{i\in I}$ is independent.*

Example 1.3 (Euler's prime number formula)**.** The Riemann zeta function is defined by the Dirichlet series

$$\zeta(s) := \sum_{n=1}^{\infty} n^{-s} \quad \text{for } s \in (1, \infty) \tag{1.5}$$

Euler's prime number formula is a representation of the Riemann zeta function as an infinite product

$$\zeta(s) = \prod_{p\in\mathcal{P}} \left(1 - p^{-s}\right)^{-1} \tag{1.6}$$

where $\mathcal{P} := \{p \in \mathbb{N} : p \text{ is prime } \}$.

We give a probabilistic proof for this formula. Let $\Omega = \mathbb{N}$, and for fixed $s > 1$ define $\mathbb{P}$ on 2^Ω by

$$\mathbb{P}\{n\} = \zeta(s)^{-1} n^{-s} \text{ for } n \in \mathbb{N}\,.$$

Let $p\mathbb{N} = \{pn : n \in \mathbb{N}\}$ and $\mathcal{P}_n = \{p \in \mathcal{P} : p \leqslant n\}$. We consider $p\mathbb{N} \subset \Omega$ as an event.

Assertion: $\mathbb{P}(p\mathbb{N}) = p^{-s}$ and $(p\mathbb{N})_{p\in\mathcal{P}}$ is independent.

Indeed, for $k \in \mathbb{N}$ and mutually distinct $p_1, \cdots, p_k \in \mathcal{P}$, we have $\bigcap_{i=1}^k (p_i\mathbb{N}) = (p_1 \cdots p_k)\,\mathbb{N}$. Thus

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^k (p_i\mathbb{N})\right) &= \sum_{n=1}^\infty \mathbb{P}\left(\{p_1 \cdots p_k n\}\right) \\ &= \zeta(s)^{-1} (p_1 \cdots p_k)^{-s} \sum_{n=1}^\infty n^{-s} \\ &= (p_1 \cdots p_k)^{-s} = \prod_{i=1}^k \mathbb{P}\left(p_i\mathbb{N}\right)\,.\end{aligned}$$

By proposition the family $((p\mathbb{N})^c, p \in \mathcal{P})$ is also independent, whence

$$\begin{aligned}\zeta(s)^{-1} = \mathbb{P}(\{1\}) &= \mathbb{P}\left(\cap (p\mathbb{N})^c\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{p\in\mathcal{P}_n} (p\mathbb{N})^c\right) \\ &= \lim_{n\to\infty} \prod_{p\in\mathcal{P}_n} (1 - \mathbb{P}(p\mathbb{N})) = \prod_{p\in\mathcal{P}} \left(1 - p^{-s}\right)\,.\end{aligned}$$

1.1.2 Independence of classes of events

Now we extend the notion of independence from families of events to families of classes of events.

Definition 1.2. Let I be an arbitrary index set and let $\mathcal{A}_i \subset \mathcal{F}$ for all $i \in I$. The family $\{\mathcal{A}_i\}_{i\in I}$ is called **independent** if, for any finite subset $J \subset I$ and any choice of $A_j \in \mathcal{A}_j$, $j \in J$, we have

$$\mathbb{P}\left(\cap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}\left(A_j\right)\,. \tag{1.7}$$

The most important case is the indenpendence of class of σ-fields $\{\mathcal{F}_i\}_{i\in I}$, where $\mathcal{F}_i \subset \mathcal{F}$ is σ-field for all $i \in I$.

Theorem 1.2. *Assume that $\mathcal{A}_i$ is π-system for all i in I. Then $\{\mathcal{A}_i\}_{i\in I}$ is independent implies that $\{\sigma(\mathcal{A}_i)\}_{i\in I}$ is independent.*

Proof. For any finite subset J of I. we will show that for any choice $A_j \in \sigma(\mathcal{A}_j)$, $j \in J$,

$$\mathbb{P}\left(\cap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}\left(A_j\right) .$$

Asuume $J = \{1, 2\cdots, n\}$. By induction, we only need to show for any given $A_j \in \mathcal{A}_j$, $2 \leqslant j \leqslant n$,

$$\mathbb{P}\left(\cap_{j=1}^{n} A_j\right) = \prod_{j=1}^{n} \mathbb{P}\left(A_j\right) , \forall\, A_1 \in \sigma(\mathcal{A}_1)\,. \tag{1.8}$$

Let

$$\mathcal{G} \coloneqq \{A_1 \in \sigma(\mathcal{A}_1) : (1.8) \text{ holds } .\}$$

Then $\mathcal{A}_1 \subset \mathcal{G}$, and $\mathcal{G}$ is a λ-systerm. Using π-λ systerm theorem, we get $\mathcal{G} = \mathcal{A}_1$. □

Corollary 1.3. *The family $\{\mathcal{F}_i\}_{i\in I}$ of σ-fields is independent. For any partition $\{I_k\}_{k\in K}$ of I, let $\mathcal{G}_k = \sigma(\cup_{i\in I_k}\mathcal{F}_i)$, then $\{\mathcal{G}_k\}_{k\in K}$ is also independent.*

Proof. Note that $\cup_{i\in I_k}\mathcal{F}_i$ are π-systerm, and $\{\cup_{i\in I_k}\mathcal{F}_i\}_{k\in K}$ is independent, thus by Theorem 1.2 we get the desired result. □

1.2 Independent Random Variables

Now that we have studied independence of events, we want to study independence of random variables. Here also the definition ends up with a product formula. Formally, however, we can also define independence of random variables via independence of the σ-fields they generate. This is the reason why we studied independence of classes of events in the last section.

Let I be an arbitrary index set. For each $i \in I$, let $X_i : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be a random variable with generated σ-algebra $\sigma(X_i) = X_i^{-1}(\mathcal{B})$.

Definition 1.3. The family $(X_i)_{i\in I}$ of random variables is called **independent** if the family $\{\sigma(X_i)\}_{i\in I}$ of σ-fields is independent.

Obviously, $(X_i)_{i\in I}$ is independent if and only if for any finite subset J of I, $(X_i)_{i\in J}$ is independent.

So it's sufficient to discuss the independence of finite many r.v.'s. It's easy to see that $X_1, \cdots, X_n$ is independent if and only if, for any choice of $B_i \in \mathcal{B}, 1 \leqslant i \leqslant n$, where $\mathcal{B}$ is the Borel algebra on $\mathbb{R}$, we have

$$\mathbb{P}(X_1 \in B_1, \cdots, X_n \in B_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i) . \tag{1.9}$$

In other words, X_i has distribution $\mathbb{P}_{X_i}$, then the distribution of $(X_1, \cdots, X_n)$ is the product probability measure $\mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}$.

By Theorem 1.2, to check the independence of $X_1, \cdots, X_n$, it's sufficient to check for any choice of $x_i \in \mathbb{R}, 1 \leqslant i \leqslant n$, if the product formular holds:

$$\mathbb{P}(X_1 \leqslant x_1, \cdots, X_n \leqslant x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leqslant x_i) , \tag{1.10}$$

i.e.,

$$F_{1,\cdots,n}(x_1, \cdots, x_n) = \prod_{i\in J} F_i(x_j) . \tag{1.11}$$

where $F_{1,\cdots,n}$ is the *joint distribution function* of $(X_1, \cdots, X_n)$. In addition, if any $F_{1,\cdots,n}$ has *density* $f_{1,\cdots,n}$ and F_i has density f_i. In this case, $X_1, \cdots, X_n$ are independent if and only if

$$f_{1,\cdots,n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_i(x_i) , \text{ for any } (x_1, \cdots, x_n) \in \mathbb{R}^n . \tag{1.12}$$

Our next goal is to prove that functions of disjoint collections of independent random variables are independent.

Theorem 1.4. *If $\{X_{i,j} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m(i)\}$ is independent. Let $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ be measurable function for all i, then $f_i\left(X_{i,1}, \cdots, X_{i,m(i)}\right)$ are independent.*

A concrete special case of Theorem 1.4 that we will use in a minute is: if $X_1, \cdots, X_n$ are independent then $X = X_1$ and $Y = X_2 \cdots X_n$ are independent. Later, when we study sums $S_m = X_1 + \cdots + X_m$ of independent random variables $X_1, \cdots, X_n$ we will use Tit to conclude that if $m < n$ then $S_n - S_m$ is independent of the indicator function of the event $\{\max_{1 \leqslant k \leqslant m} S_k > x\}$.

1.2.1 Expectation, Variance

Independent random variables allow for a rich calculus. For example, we can compute the expectation of a product of two indenpendent random variables and distribution of a sum of two independent random variables.

Theorem 1.5. *If $X_1, \cdots, X_n$ are independent and have (a) $X_i \geq 0$ for all i, or (b) $\mathbb{E}\,|X_i| < \infty$ for all i then*

$$\mathbb{E} \prod_{i=1}^{n} X_i = \prod_{i=1}^{n} \mathbb{E} X_i$$

i.e., the expectation on the left exists and has the value given on the right.

Proof. Using *variable substitution* and *Fubini theorem.* □

Theorem 1.6. *If $X_1, \cdots, X_n$ are independent and $\mathrm{Var}(X_i) < \infty$, then*

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i)\,. \tag{1.13}$$

In fact, A family of random variables $\{X_i\}_{i \in I}$ with $\mathbb{E} X_i^2 < \infty$ is said to be **uncorrelated** if we have

$$\mathbb{E}\,(X_i X_j) = \mathbb{E} X_i \mathbb{E} X_j \quad \text{ whenever } i \neq j\,,$$

if then, It's easy to check that

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Var}\left(X_1\right) + \cdots + \mathrm{Var}\left(X_n\right) .$$

It can happen that X, Y are uncorrelated, i.e., $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$, but not are independent. For any example, let X normally distributed with mean 0 and variance 1, then $\mathbb{E}(X|X|) = \mathbb{E}X \cdot \mathbb{E}|X| = 0$, but X and $|X|$ are not independent.

1.2.2 Sums of Independent Random Variables

Theorem 1.7. *If X and Y are independent, $F(x) = \mathbb{P}(X \leqslant x)$, and $G(y) = \mathbb{P}(Y \leqslant y)$, then*

$$\mathbb{P}(X + Y \leqslant z) = \int F(z - y) dG(y) \tag{1.14}$$

The integral on the right-hand side is called the **convolution** of F and G and is denoted $F * G(z)$.

Proof. Let μ and ν be the probability measures with distribution functions F and G. Since for fixed y

$$\int 1_{\{x+y\leqslant z\}} \mu(dx) = \int 1_{(-\infty, z-y]}(x) \mu(dx) = F(z - y)$$

then

$$\begin{aligned} P(X + Y \leqslant z) &= \iint 1_{(x+y\leqslant z)} \mu(dx)\nu(dy) \\ &= \int F(z - y)\nu(dy) = \int F(z - y) dG(y) . \end{aligned}$$

The last equality is just a change of notation, we regard $dG(y)$ as a shorthand for "integrate with respect to the measure ν induced by G." □

To treat concrete examples, we need a special case of Theorem 1.7.

Corollary 1.8. *Theorem 2.1.16. Suppose that X with density f and Y with distribution function G are independent. Then $X+Y$ has density*

$$h(x) = \int f(x-y)dG(y) \tag{1.15}$$

When Y has density g, the last formula can be written as

$$h(x) = \int f(x-y)g(y)dy \tag{1.16}$$

Proof. From Theorem 1.7, the definition of density function, and Fubini's theorem, we get

$$\begin{aligned}\mathbb{P}(X+Y \leqslant z) &= \int F(z-y)dG(y) = \iint_{-\infty}^{z} f(x-y)dxdG(y) \\ &= \int_{-\infty}^{z}\int f(x-y)dG(y)dx\,.\end{aligned}$$

The last equation says that $X+Y$ has density $h(x) = \int f(x-y)dG(y)$. □

Corollary 1.8 plus some ugly calculus allows us to treat two standard examples. These facts should be familiar from undergraduate probability.

Example 1.4. The gamma density with parameters $\alpha > 0$ and $\lambda > 0$ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}\frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases}$$

In particular, gamma$(1,\lambda)$ is exactly exponential(λ).

If $X = \text{gamma}(\alpha,\lambda)$ and $Y = \text{gamma}(\beta,\lambda)$ are independent then $X+Y$ is gamma$(\alpha+\beta,\lambda)$. Consequently if $X_1, \cdots X_n$ are independent exponential(λ) $r.v.'s$, then $X_1 + \cdots + X_n$, has a gamma(n, λ) distribution.

To see this, writing $f_{X+Y}(z)$ for the density function of $X+Y$ and using Corollary 1.8

$$\begin{aligned}f_{X+Y}(x) &= \int_0^x \frac{\lambda^\alpha(x-y)^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda(x-y)}\frac{\lambda^\beta y^{\beta-1}}{\Gamma(\beta)}e^{-\lambda y}dy \\ &= \frac{\lambda^{\alpha+\beta}e^{-\lambda x}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x (x-y)^{\alpha-1}y^{\beta-1}dy\,.\end{aligned}$$

so it suffices to show the integral is $x^{\alpha+\beta-1}\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. To do this, we begin by changing variables $y = xu, dy = xdu$ to get

$$\int_0^x (x-y)^{\alpha-1}y^{\beta-1}dy = x^{\alpha+\beta-1}\int_0^1 (1-u)^{\alpha-1}u^{\beta-1}du = x^{\alpha+\beta-1}B(\alpha,\beta)\,,$$

Where B is beta function, and we konw $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

Example 1.5. $X =$ normal(μ, a) and $Y =$ normal(ν, b) are independent then $X + Y =$ normal$(\mu+\nu, a+b)$.

It is enough to prove the result for $\mu = \nu = 0$. Suppose $Y_1 =$ normal $(0, a)$ and $Y_2 =$ normal $(0, b)$. Then Corollary 1.8 implies

$$f_{Y_1+Y_2}(z) = \frac{1}{2\pi\sqrt{ab}}\int e^{-x^2/2a}e^{-(z-x)^2/2b}dx\,.$$

Dropping the constant in front, the integral can be rewritten as

$$\begin{aligned}
&\int \exp\left(-\frac{bx^2+ax^2-2axz+az^2}{2ab}\right)dx\\
&= \int \exp\left(-\frac{a+b}{2ab}\left\{x^2-\frac{2a}{a+b}xz+\frac{a}{a+b}z^2\right\}\right)dx\\
&= \int \exp\left(-\frac{a+b}{2ab}\left\{\left(x-\frac{a}{a+b}z\right)^2+\frac{ab}{(a+b)^2}z^2\right\}\right)dx\,.
\end{aligned}$$

Since $-\{a/(a+b)\}^2+\{a/(a+b)\} = ab/(a+b)^2$. Factoring out the term that does not depend on x, the last integral

$$\begin{aligned}
&= \exp\left(-\frac{z^2}{2(a+b)}\right)\int \exp\left(-\frac{a+b}{2ab}\left(x-\frac{a}{a+b}z\right)^2\right)dx\\
&= \exp\left(-\frac{z^2}{2(a+b)}\right)\sqrt{2\pi ab/(a+b)}\,,
\end{aligned}$$

since the last integral is the normal density with parameters $\mu = az/(a+b)$ and $\sigma^2 = ab/(a+b)$ without its proper normalizing constant. Reintroducing the constant we dropped at the beginning,

$$f_{Y_1+Y_2}(z) = \frac{1}{2\pi\sqrt{ab}}\sqrt{2\pi ab/(a+b)}\exp\left(-\frac{z^2}{2(a+b)}\right)\,.$$

Example 1.6. $X =$ Poisson (λ) and $Y =$ Poisson (μ) are independent then $X + Y =$ Poisson $(\lambda + \nu)$.

To see this, by Theorem 1.7, for any $n \geqslant 0$,

$$\begin{aligned}\mathbb{P}(X+Y=n) &= e^{-\mu}e^{-\lambda}\sum_{m=0}^{n}\frac{\mu^m}{m!}\frac{\lambda^{n-m}}{(n-m)!} \\ &= e^{-(\mu+\lambda)}\frac{1}{n!}\sum_{m=0}^{n}\binom{n}{m}\mu^m\lambda^{n-m} = e^{-(\mu+\lambda)}\frac{(\mu+\lambda)^n}{n!}\,.\end{aligned}$$

Chapter 2

Law of Large Numbers

In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer to the expected value as more trials are performed.

The LLN is important because it guarantees stable long-term results for the averages of some random events. For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins. Any winning streak by a player will eventually be overcome by the parameters of the game. It is important to remember that the law only applies (as the name indicates) when a large number of observations is considered. There is no principle that a small number of observations will coincide with the expected value or that a streak of one value will immediately be "balanced" by the others.

2.1 Weak Law of Large Numbers

2.1.1 L^2 Weak Laws

Our first set of weak laws come from computing variances and using Chebyshev's inequality.

Theorem 2.1 (L^2 weak law). *Let $X_1, X_2, \cdots$ be uncorrelated random variables with $\mathbb{E}X_n = \mu$ and $\mathrm{Var}(X_n) \leqslant C < \infty$ for all n. Let $S_n = X_1 + \cdots + X_n$, then as $n \to \infty$,*

$$\frac{S_n}{n} \to \mu \text{ in } L^2 .$$

Proof. To prove L^2 convergence, observe that $\mathbb{E}\left(\frac{S_n}{n}\right) = \mu$, hence

$$\mathbb{E}\left|\frac{S_n}{n} - \mu\right|^2 = \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}(X_j) \leqslant \frac{Cn}{n^2} .$$

So as $n \to \infty$, $\frac{S_n}{n} \to \mu$ in L^2. □

The most important special case of the L^2 weak law occurs when $\{X_n\}$ are independent and identically distributed or i.i.d. for short. L^2 weak law tells us in this case that if $EX_n^2 < \infty$, then $\frac{S_n}{n}$ converges to $\mu = \mathbb{E}X_1$ in probability. We are going to give some applications of the L^2 weak law.

Example 2.1 (Polynomial approximation)**.** Let f be a continuous function on $[0, 1]$, We will use L^2 weak law to find some polynomials $\{f_n\}$, such that as $n \to \infty$

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0 . \tag{2.1}$$

Let $\{X_n\}$ be i.i.d. r.v.'s obeying binomial$(1, x)$, and $S_n = X_1 + \cdots + X_n$. Since $\frac{S_n}{n}$ converges to x in probability and f is continous, $f(\frac{S_n}{n})$ converges to $f(x)$ in probability. By Lebesgue domainted convergence theorem we have $\mathbb{E}f(\frac{S_n}{n}) \to f(x)$, and $\mathbb{E}f(\frac{S_n}{n})$ are exactly ploynomials. Thus let

$$f_n(x) \coloneqq \mathbb{E}f\left(\frac{S_n}{n}\right) = \sum_{m=0}^{n}\binom{n}{m}x^m(1-x)^{n-m}f\left(\frac{m}{n}\right) .$$

Thus, it's natural to ask if $\mathbb{E}f(\frac{S_n}{n})$ can approximate f uniformly. From the proof of L^2 weak law we have

$$\mathbb{E}\left|\frac{S_n}{n} - x\right|^2 \leqslant \frac{\mathrm{Var}(X_1)}{n} \leqslant \frac{1}{n}$$

Claim: if f is uniformly continuous on $[0,1]$, then for any $\epsilon > 0$, there is a $M = M_\epsilon > 0$ such that

$$|f(x) - f(y)| \leqslant M|x-y| + \epsilon\,, \quad \text{for any } x, y \in [0,1]\,.$$

This claim is a classical exercise in the course of mathematic analysis. Thus we have

$$\begin{aligned} |f_n(x) - f(x)| &\leqslant \mathbb{E}\left| f\left(\frac{S_n}{n}\right) - f(x)\right| \\ &\leqslant M\,\mathbb{E}\left|\frac{S_n}{n} - x\right| + \epsilon \\ &\leqslant M\left[\mathbb{E}\left|\frac{S_n}{n} - x\right|^2\right]^{1/2} + \epsilon \\ &\leqslant \frac{M}{\sqrt{n}} + \epsilon\,. \end{aligned}$$

Thus (2.1) holds. $f_n(x)$ is called the **Bernstein polynomial of degree** n **associated with** f.

Example 2.2 (A high-dimensional cube is almost the boundary of a ball.)**.** We choose a point in $(-1,1)^n$ arbitary, denote as $(X_1, \cdots, X_n)$. So $X_1, X_2, \cdots$ be independent and uniformly distributed on $(-1,1)$. Let $Y_i = X_i^2$, which are independent since they are functions of independent random variables. $\mathbb{E}Y_i = 1/3$ and $\mathrm{Var}\,(Y_i) \leqslant \mathbb{E}Y_i^2 \leqslant 1$, so Theorem 2.1 implies

$$\frac{X_1^2 + \cdots + X_n^2}{n} \to \frac{1}{3} \quad \text{in probability}\ .$$

Let

$$A_{n,\epsilon} = \left\{ x \in \mathbb{R}^n : (1-\epsilon)\sqrt{\frac{n}{3}} < \|x\|_2 < (1+\epsilon)\sqrt{\frac{n}{3}} \right\}\,.$$

If we let $|S|$ denote the Lebesgue measure of S, then the last conclusion implies that, for any $\epsilon > 0$,

$$\frac{1}{2^n}\,|A_{n,\epsilon} \cap (-1,1)^n| \to 1\,,$$

or, in words, most of the volume of the cube $(-1,1)^n$ comes from $A_{n,\epsilon}$, which is almost the boundary of the ball of radius $\sqrt{\frac{n}{3}}$.

2.1.2 Triangular arrays

Many classical limit theorems in probability concern arrays random variables

$$\begin{array}{ccccccc} X_{11} & & & & & & \\ X_{21} & X_2 & & & & & \\ X_{31} & X_{32} & X_{33} & & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ X_{n1} & X_{n2} & X_{n3} & \cdots & X_{nn} & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{array}$$

and investigate the limiting behavior of their row sums

$$S_n = \sum_{k=1}^{n} X_{n,k} \,.\ \text{for } n \geqslant 1 \,.$$

In most cases, we assume that the random variables on each row are independent.

From the proof of L^2 weak law, it's easy to give a L^2 weak law for triangular arrays as following:

Theorem 2.2. *For each n let $X_{n,k}, 1 \leqslant k \leqslant n$, be independent. Let $S_n = X_{n,1} + \cdots + X_{n,n}$. If $\{b_n\}$ satisfies $\frac{\mathrm{Var}(S_n)}{b_n^2} \to 0$, then*

$$\frac{S_n - \mathbb{E}S_n}{b_n} \to 0 \quad \text{in } L^2 \,.$$

Proof. To prove L^2 convergence, observe that

$$\mathbb{E}\left|\frac{S_n - \mathbb{E}S_n}{b_n}\right|^2 = \frac{\mathrm{Var}(S_n)}{b_n^2} \to 0$$

which gives the disired result. □

Remark. In fact, we didn't use the condition $X_{n,1}, \cdots, X_{n,n}$ are independent, here S_n can be any sequence of random variables.

We will now give three applications of Theorem 2.2.

Example 2.3 (Coupon collector's problem)**.** Let $X_1, X_2, \cdots$ be i.i.d. uniform on $\{1, 2, \cdots, n\}$. To motivate the name, think of collecting coupons. Suppose that the i th item we collect is chosen at ridandom from the set of possibilities and is independent of the previous choices. Let $Y_0 = 0$ and $Y_m \coloneqq |\{X_1, \cdots, X_m\}|$. We can see that $(Y_m)_{m\geqslant 0}$ is a *Markov Chain* with finite state. Let

$$\tau_k^n = \inf\{m \geqslant 0 : Y_m = k\}$$

be the first time we have k different items. In this problem, we are interested in the asymptotic behavior of $T_n = \tau_n^n$, the time to collect a complete set. It is easy to see that $\tau_0^n = 0$ and $\tau_1^n = 1$.

For $1 \leqslant k \leqslant n$, Let $X_{n,k} \coloneqq \tau_k^n - \tau_{k-1}^n$ represents the time to get a choice different from our first $k-1$, so $X_{n,k}$ has a geometric distribution with parameter $1 - \frac{k-1}{n}$ and is independent of the earlier waiting times $X_{n,j}$, $1 \leqslant j < k$, by strong Markov property. (Another method is to compute the joint distribution $\mathbb{P}(X_{n,1} = 1, \cdots, X_{n,k} = m_k)$, find that it is exactly the product of marginal distributions (边缘分布)). Thus,

$$\begin{aligned}
\mathbb{E}T_n &= \sum_{k=1}^{n} \left(1 - \frac{k-1}{n}\right)^{-1} = n \sum_{m=1}^{n} \frac{1}{m} \sim n \log n\,, \\
\operatorname{Var}(T_n) &\leqslant \sum_{k=1}^{n} \left(1 - \frac{k-1}{n}\right)^{-2} = n^2 \sum_{m=1}^{n} \frac{1}{m^2} \leqslant n^2 \sum_{m=1}^{\infty} \frac{1}{m^2}\,.
\end{aligned}$$

Taking $b_n = n \log n$ and using Thoerem 2.2 , it follows that

$$\frac{T_n - n\sum_{m=1}^{n} m^{-1}}{n \log n} \to 0 \quad \text{in probability}$$

and hence

$$\frac{T_n}{n \log n} \to 1 \quad \text{in probability}\,.$$

For a concrete example, take $n = 365$, i.e., we are interested in the number of people we need to meet until we have seen someone with every birthday. In this case the limit theorem says it will take about $365 \log 365 = 2153.46$ tries to get a complete set. Note that the number of trials is 5.89 times the number of birthdays.

REMARK. This interesting example combines Markov chain and law of large numbers : we find some i.i.d. r.v.'s by using strong Markov property and then analysis the limiting behavior of their sums.

Example 2.4 (Random permutations)**.** Let Ω_n consist of the $n!$ permutations (i.e., one-to-one mappings from $\{1, \cdots, n\}$ onto $\{1, \cdots, n\}$) and make this into a probability space by assuming all the permutations are equally likely. This application of the weak law concerns the cycle structure of a random permutation π, so we begin by describing the decompostion of a permutation into cycles.

Consider the sequence $1, \pi(1), \pi(\pi(1)), \cdots$ Eventually, $\pi^k(1) = 1$. When it does, we say the first cycle is completed and has length k. To start the second cycle, we pick the smallest integer i not in the first cycle and look at $i, \pi(i), \pi(\pi(i)), \cdots$ until we come back to i. We repeat the construction until all the elements are accounted for. For example, if the permutation is

i	1	2	3	4	5	6	7	8	9
$\pi(i)$	3	9	6	8	2	1	5	4	7

then the cycle decomposition is (136)(2975)(48).

Let $X_{n,k} = 1$ if a right parenthesis (括号) occurs after the k th number in the decomposition, $X_{n,k} = 0$ otherwise and let $S_n = X_{n,1} + \cdots + X_{n,n} =$ the number of cycles. (In the example, $X_{9,3} = X_{9,7} = X_{9,9} = 1$, and the other $X_{9,m} = 0$).

Claim: $X_{n,1}, \cdots, X_{n,n}$ are independent and $\mathbb{P}(X_{n,j} = 1) = \frac{1}{n-j+1}$.

To show this, we compute that for any $1 \leqslant j \leqslant n$ and $1 \leqslant a_1 < \cdots < a_j = n$,

$$\begin{aligned}&\mathbb{P}(X_{n,k} = 1, k \in \{a_1, \cdots, a_j\}; X_{n,k} = 0, k \notin \{a_1, \cdots, a_j\}) \\ &= \frac{1}{(n-a_1)\cdots(n-a_j)} = \prod_{k \in \{a_1, \cdots, a_j\}} \frac{1}{n-k+1} \cdot \prod_{k \notin \{a_1, \cdots, a_j\}} \frac{n-k}{n-k+1}.\end{aligned}$$

Hence $\mathbb{P}(X_{n,k} = 1) = \frac{1}{n-k+1}$ and $X_{n,1}, \ldots, X_{n,n}$ are independent.

To check the conditions of Theorem 2.2, now note

$$\mathbb{E}S_n = \frac{1}{n+1} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1\,,$$
$$\mathrm{Var}\,(S_n) = \sum_{k=1}^{n} \mathrm{Var}\,(X_{n,k}) \leqslant \sum_{k=1}^{n} \mathbb{E}\left(X_{n,k}^2\right) = \sum_{k=1}^{n} \mathbb{E}\,(X_{n,k}) = \mathbb{E}S_n\,.$$

Now $\mathbb{E}S_n \sim \log n$, so if $b_n = (\log n)^{0.5+\epsilon}$ with $\epsilon > 0$, then

$$\frac{S_n - \sum_{m=1}^{n} \frac{1}{m}}{(\log n)^{0.5+\epsilon}} \to 0 \text{ in probability } . \tag{2.2}$$

Taking $\epsilon = 0.5$ we have that

$$\frac{S_n}{\log n} \to 1 \text{ in probability } .$$

But (2.2) says more. We will see in Example 3.27 that (2.2) is false if $\epsilon = 0$.

In the preceeding two examples, S_n is the sum of independent r.v.'s, and we give an example concerning the sum of dependent r.v.'s.

Example 2.5 (An occupancy problem)**.** Suppose we put $r = r(n)$ balls at random in n boxes, i.e., all n^r assignments of balls to boxes have equal probability. Let $A_i = \{\text{the } i\text{-th box is empty}\}$, and $N_n = \sum_{i=1}^{n} 1_{A_i}$. It is easy to see that

$$\mathbb{P}\,(A_i) = \left(1 - \frac{1}{n}\right)^r \quad \text{and} \quad \mathbb{E}N_n = n\left(1 - \frac{1}{n}\right)^r .$$

A little calculus shows that if $\frac{r(n)}{n} \to c$, then $\frac{\mathbb{E}N_n}{n} \to e^{-c}$. To compute the variance of N_n ,we observe that

$$\mathbb{E}N_n^2 = \mathbb{E}\left(\sum_{m=1}^{n} 1_{A_m}\right)^2 = \sum_{1 \leqslant k,m \leqslant n} \mathbb{P}\,(A_k \cap A_m)\,.$$
$$\mathrm{Var}\,(N_n) = \mathbb{E}N_n^2 - (\mathbb{E}N_n)^2 = \sum_{1 \leqslant k,m \leqslant n} \mathbb{P}\,(A_k \cap A_m) - \mathbb{P}\,(A_k)\,\mathbb{P}\,(A_m)$$
$$= n(n-1)\left[\left(1 - \frac{2}{n}\right)^r - \left(1 - \frac{1}{n}\right)^{2r}\right] + n\left[\left(1 - \frac{1}{n}\right)^r - v(1 - \frac{1}{n})^{2r}\right] .$$

The first term comes from $k \neq m$ and the second from $k = m$. Since

$$\left(1 - \frac{2}{n}\right)^{r(n)} \to e^{-2c} \quad \text{and} \quad \left(1 - \frac{1}{n}\right)^{r(n)} \to e^{-c},$$

it follows easily from the last formula that $\frac{\mathrm{Var}(N_n)}{n^2} \to 0$. Taking $b_n = n$ in Theorem2.2 now we have

$$\frac{N_n}{n} \to e^{-c} \quad \text{in probability .}$$

2.1.3 Truncation

Note that in Theorem 2.2, we suppose that the variance of $X_{n,k}$ exists. However, we have no reason to ensure this. In this subsection, we introduce a useful method to treat general r.v. which do not have a variance or even expected. This method will be used in all the notes, is really important.

To **truncate** a random variable X at level M, where $M > 0$, means to consider

$$\bar{X} = X 1_{\{|X| \leqslant M\}} = \begin{cases} X, & \text{if } |X| \leqslant M. \\ 0, & \text{if } |X| > M. \end{cases} \tag{2.3}$$

Now, to extend the weak law to random variables without a finite second moment, we will truncate and then use Chebyshev's inequality. We begin with a very general but also very useful result. Its proof is easy because we have assumed what we need for the proof. Later we will have to work a little to verify the assumptions in special cases, but the general result serves to identify the essential ingredients in the proof.

Theorem 2.3 (Weak law for triangular arrays). *For each n let $X_{n,k}$, $1 \leqslant k \leqslant n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and $\bar{X}_{n,k} = X_{n,k} 1_{\{|X_{n,k}| \leqslant b_n\}}$. Suppose that as $n \to \infty$,*

$$\textit{(i)} \ \sum_{k=1}^{n} \mathbb{P}\left(|X_{n,k}| > b_n\right) \to 0, \quad \textit{and (ii)} \ \frac{1}{b_n^2} \sum_{k=1}^{n} \mathbb{E}\bar{X}_{n,k}^2 \to 0\,. \tag{2.4}$$

Put $\bar{S}_n = \sum_{k=1}^n \bar{X}_{n,k}$, *then*

$$\frac{S_n - \mathbb{E}\bar{S}_n}{b_n} \to 0 \text{ in probability } .$$

Proof. Note that

$$\mathbb{P}\left(\left|\frac{S_n - \mathbb{E}\bar{S}_n}{b_n}\right| > \epsilon\right) \leqslant \mathbb{P}\left(S_n \neq \bar{S}_n\right) + \mathbb{P}\left(\left|\frac{\bar{S}_n - \mathbb{E}\bar{S}_n}{b_n}\right| > \epsilon\right) .$$

To estimate the first term, we note that

$$\mathbb{P}\left(S_n \neq \bar{S}_n\right) \leqslant \mathbb{P}\left(\cup_{k=1}^n \left\{\bar{X}_{n,k} \neq X_{n,k}\right\}\right) \leqslant \sum_{k=1}^n \mathbb{P}\left(|X_{n,k}| > b_n\right) \to 0 .$$

For the second term, use Chebyshev' s inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\bar{S}_n - \mathbb{E}\bar{S}_n}{b_n}\right| > \epsilon\right) &\leqslant \frac{1}{\epsilon^2}\mathbb{E}\left|\frac{\bar{S}_n - \mathbb{E}\bar{S}_n}{b_n}\right|^2 = \frac{1}{\epsilon^2 b_n^2}\operatorname{Var}\left(\bar{S}_n\right) \\ &= \frac{1}{\epsilon^2 b_n^2}\sum_{k=1}^n \operatorname{Var}\left(\bar{X}_{n,k}\right) \leqslant \frac{1}{\epsilon^2 b_n^2}\sum_{k=1}^n \mathbb{E}\bar{X}_{n,k}^2 \to 0 . \quad \square\end{aligned}$$

Remark. Why we choose the truncation level $\{b_n\}$ satisfying (i) and (ii)? Note that (i) implies $S_n - \bar{S}_n$ converges to 0 in probability, and (ii) implies that

$$\frac{\bar{S}_n - \mathbb{E}\bar{S}_n}{b_n} \to 0 \text{ in probability} .$$

These two reason reflects the key to choose the truncation level :

> *On the one hand, the truncated r.v.'s satisfy the desired properties. On the other hand, the truncation error is so small that the desired properties holds for the original r.v.'s.*

We should always keep this in mind when using the method of truncation.

Weak law of large numbers From Theorem 2.3 we get the following result for a single sequence. To do this, we need the following lemma, which will be useful several times below.

Lemma 2.4. *If $Y \geqslant 0$ and $p > 0$ then $\mathbb{E}(Y^p) = \int_0^\infty py^{p-1}\mathbb{P}(Y > y)dy$.*

Proof. Using the definition of expected value, Fubini's theorem (for nonnegative random variables), and then calculating the resulting integrals gives

$$\begin{aligned}\int_0^\infty py^{p-1}\mathbb{P}(Y > y)\,\mathrm{d}y &= \int_0^\infty \int_\Omega py^{p-1}1_{\{Y>y\}}\,\mathrm{d}\mathbb{P}\,\mathrm{d}y \\ &= \int_\Omega \int_0^\infty py^{p-1}1_{\{Y>y\}}\,\mathrm{d}y\,\mathrm{d}\,\mathbb{P} \\ &= \int_\Omega \int_0^Y py^{p-1}\,\mathrm{d}y\,\mathrm{d}\mathbb{P} = \int_\Omega Y^p\,\mathrm{d}\mathbb{P} = \mathbb{E}Y^p\end{aligned}$$

which is the desired result. □

Theorem 2.5 (WLLN). *Let $X_1, X_2, \cdots$ be i.i.d. with*

$$x\,\mathbb{P}(|X_1| > x) \to 0\,, \quad \text{as } x \to \infty\,.$$

Put $\mu_n = \mathbb{E}X_1 1_{\{|X_1|\leqslant n\}}$, $S_n = X_1 + \cdots + X_n$ for each n. Then

$$\frac{S_n}{n} - \mu_n \to 0 \text{ in probability}$$

Remark. For any $\epsilon > 0$, Applying Lemma 2.4 with $p = 1 - \epsilon$, we see that

$$x\,\mathbb{P}(|X_1| > x) \to 0 \Rightarrow \mathbb{E}\,|X_1|^{1-\epsilon} < \infty,$$

so the assumption in Theorem 2.5 is not much weaker than finite mean.

Proof. We will apply Theorem 2.3 with $X_{n,k} = X_k$ and $b_n = n$. First we observe that

$$\sum_{k=1}^n \mathbb{P}(|X_{n,k}| > n) = n\mathbb{P}(|X_i| > n) \to 0\,.$$

So we need to show that

$$\frac{1}{b_n^2}\sum_{k=1}^n \mathbb{E}\bar{X}_{n,k}^2 = \frac{1}{n}\mathbb{E}\bar{X}_{n,1}^2 \to 0\,.$$

Now return to the proof. $\bar{X}_{n,1} = X_1 1_{\{|X_1| \leqslant n\}}$ imply

$$\mathbb{E}\left(\bar{X}_{n,1}^2\right) = \int_0^\infty 2y\mathbb{P}\left(\left|\bar{X}_{n,1}\right| > y\right) dy \leqslant \int_0^n 2y\mathbb{P}\left(|X_1| > y\right) dy\,,$$

since $\mathbb{P}\left(\left|\bar{X}_{n,1}\right| > y\right) = 0$ for $y \geqslant n$ and $= \mathbb{P}\left(|X_1| > y\right) - \mathbb{P}\left(|X_1| > n\right)$ for $y \leqslant n$. We claim that $y\mathbb{P}\left(|X_1| > y\right) \to 0$ as $y \to \infty$ implies

$$\frac{\mathbb{E}\left(\bar{X}_{n,1}^2\right)}{n} \leqslant \frac{1}{n}\int_0^n 2y\mathbb{P}\left(|X_1| > y\right) dy \to 0\,.$$

In fact we only need to change the variable y to nt,

$$\frac{1}{n}\int_0^n 2y\mathbb{P}\left(|X_1| > y\right) dy = \int_0^1 2nt\mathbb{P}\left(|X_1| > nt\right) dt \to 0\,.$$

by Lebesgue domainted limit theorem. □

Finally, we have the weak law in its most familiar form.

Corollary 2.6. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}\,|X_i| < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mathbb{E}X_1 \text{ in probability}$$

Proof. Two applications of the dominated convergence theorem imply

$$x\mathbb{P}\left(|X_1| > x\right) \leqslant \mathbb{E}\,|X_1|\,1_{\{|X_1|>x\}} \to 0 \quad \text{as } x \to \infty$$

and

$$\mu_n = \mathbb{E}X_1 1_{\{|X_1| \leqslant n\}} \to \mathbb{E}X_1\,,$$

which gives the desired result. □

Example 2.6. For an example where the weak law does not hold, suppose $X_1, X_2, \cdots$ are independent and have a *Cauchy distribution*:

$$\mathbb{P}\left(X_n \leqslant x\right) = \int_{-\infty}^x \frac{dt}{\pi\left(1+t^2\right)}\,.$$

As $x \to \infty$,

$$\mathbb{P}(|X_1| > x) = 2\int_x^\infty \frac{dt}{\pi(1+t^2)} \sim \frac{2}{\pi}\int_x^\infty t^{-2}dt = \frac{2}{\pi}x^{-1}.$$

We claim that there is no sequence of constants μ_n so that $\frac{S_n}{n} - \mu_n$ converges to 0 in probability. In fact, we will show that $\frac{S_n}{n}$ always has the same distribution as X_1 in Example 3.19.

Example 2.7 (The"St. Petersburg paradox"). Let $X_1, X_2, \cdots$ be i.i.d. random variables, and the distribution is given by

$$\mathbb{P}\left(X_i = 2^j\right) = \frac{1}{2^j} \text{ for } j \geqslant 1.$$

In words, you win 2^j dollars if it takes j tosses to get a heads. The paradox here is that $\mathbb{E}X_1 = \infty$, but you clearly wouldn't pay an infinite amount to play this game. An application of Theorem 2.3 will tell us how much we should pay to play the game n times.

In this example, let $X_{n,k} = X_k$. To apply Theorem 2.3, we have to pick b_n. To do this, we are guided by the principle that we want to take b_n as small as we can and have (2.4) hold. With this in mind, we observe that if m is an integer

$$\mathbb{P}(X_1 \geqslant 2^m) = \sum_{j=m}^\infty \frac{1}{2^j} = \frac{1}{2^{m-1}}.$$

Let $m(n) = \log_2 n + K(n)$ where $K(n) \to \infty$ and is chosen so that $m(n)$ is an integer (and hence the displayed formula is valid). Letting $b_n = 2^{m(n)}$, we have

$$n\mathbb{P}(X_1 \geqslant b_n) = \frac{n}{2^{m(n)-1}} = \frac{1}{2^{K(n)-1}} \to 0$$

and we observe that if $\bar{X}_{n,k} = X_k 1_{(|X_k| \leqslant b_n)}$ then

$$\mathbb{E}\bar{X}_{n,k}^2 = \sum_{j=1}^{m(n)} 2^{2j} \cdot \frac{1}{2^j} \leqslant 2^{m(n)} \sum_{k=0}^\infty \frac{1}{2^k} = 2b_n.$$

Since $K(n) \to \infty$,

$$\frac{1}{b_n^2}\sum_{k=1}^n \mathbb{E}\bar{X}_{n,k}^2 = \frac{2n}{b_n} = \frac{2}{K(n)} \to 0.$$

Cmpute

$$\sum_{k=1}^{n} \mathbb{E}\bar{X}_{n,k} = \sum_{k=1}^{n} \sum_{j=1}^{m(n)} 2^j \frac{1}{2^j} = nm(n)\,.$$

Applying Theorem 2.3, we have

$$\frac{S_n - nm(n)}{n2^{K(n)}} \to 0 \text{ in probavility}\,.$$

We pick $K(n)$ such that $\frac{m(n)}{2^{K(n)}} \to 1$, for example, suppose that $K(n) \leqslant \log_2 \log_2 n$ for large n, then the last conclusion holds with the denominator replaced by $n \log_2 n$, and it follows that

$$\frac{S_n}{n \log_2 n} \to 1 \text{ in probability}\,.$$

Returning to our original question, we see that a fair price for playing n times is $\$ \log_2 n$ per play. When $n = 1024$, this is \$10 per play. Nicolas Bernoulli wrote in 1713, "There ought not to exist any even halfway sensible person who would not sell the right of playing the game for 40 ducates (per play)." If the wager were 1 ducat, one would need $2^{40} \approx 10^{12}$ plays to start to break even.

EXERCISE

Exercise 1. Generalize Lemma 2.4 to conclude that if $H(x) = \int_{(-\infty,x]} h(y)dy$ with $h(y) \geqslant 0$, then

$$\mathbb{E}H(X) = \int_{-\infty}^{\infty} h(y)\mathbb{P}(X \geqslant y)dy$$

An important special case is $H(x) = \exp(\theta x)$ with $\theta > 0$.

Exercise 2 (Weak law for positive variables). Suppose $X_1, X_2, \cdots$ are i.i.d. non-negative r.v.'s, and $\mathbb{P}(X_1 > x) > 0$ for all $x > 0$. Let

$$\mu(s) = \int_0^s x dF(x)\,,\ \nu(s) = \frac{\mu(s)}{s(1 - F(s))}\,.$$

If $\nu(s) \to \infty$ as $s \to \infty$. Show that

(i) We can pick $b_n \geqslant 1$ so that $n\mu(b_n) = b_n$ (this works for large n).

(ii) $\frac{S_n}{b_n} \to 1$ in probability.

EXERCISE 3 (Monte Carlo integration). Let f be a measurable function on $[0,1]$ with $\int_0^1 |f(x)|dx < \infty$. Let $U_1, U_2, \cdots$ be independent and uniformly distributed on $[0,1]$, and let

$$I_n = \frac{f(U_1) + \cdots + f(U_n)}{n}.$$

(i) Show that $I_n \to I := \int_0^1 f(x)dx$ in probability.

(ii) Suppose $\int_0^1 |f(x)|^2 dx < \infty$. Use Chebyshev's inequality to estimate

$$\mathbb{P}\left(|I_n - I| > \frac{a}{\sqrt{n}}\right).$$

2.2 Borel-Canteli Lemma

If A_n is a sequence of subsets of Ω, we let $\limsup A_n$ be the event that infinitely many $\{A_n\}$ happens , i.e. ,

$$\limsup A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$$

and let $\liminf A_n$ be the event that all but finitely many $\{A_n\}$ happens , i.e. ,

$$\liminf A_n = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n$$

The names lim sup and lim inf can be explained by noting that

$$\limsup_{n\to\infty} 1_{A_n} = 1_{\{\limsup A_n\}} \quad \liminf_{n\to\infty} 1_{A_n} = 1_{\{\liminf A_n\}}$$

It is common to write $\limsup A_n = \{\omega : \omega \in A_n \text{ i.o. }\}$, where i.o. stands for infinitely often. An easy but useful theorem which illustrates the use of this notation is :

Theorem. *$X_n \to 0$ a.s. if and only if for all $\epsilon > 0$,*

$$\mathbb{P}\left(|X_n| > \epsilon \text{ i.o. }\right) = 0\,.$$

First Borel-Cantelli lemma The next result should be familiar from measure theory even though its name may not be, so the proof will be left as an exercise.

Theorem 2.7 (B-C lemma). *If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}\left(A_n \text{ i.o. }\right) = 0\ .$$

Remark. The converse of the Borel-Cantelli lemma is trivially false, we will now give a counterexample.

Example 2.8. Let $\Omega = (0,1), \mathcal{F} =$ Borel sets, $\mathbb{P} =$ Lebesgue measure. If $A_n = (0, a_n)$ where $a_n \to 0$ as $n \to \infty$ then $\limsup A_n = \varnothing$, but if $a_n \geqslant 1/n$, we have $\sum a_n = \infty$.

Applications We give some typical applications of the Borel-Cantelli lemma. The first is to prove a necessary and sufficient condition of convegence in probability.

Example 2.9. *$X_n \to X$ in probability if and only if for every subsequence $X_{n(m)}$ there is a further subsequence $X_{n(m_k)}$ that converges to X a.s..*

Let ϵ_k be a sequence of positive numbers that $\downarrow 0$. For each k, there is an $n(m_k) > n(m_{k-1})$ so that $\mathbb{P}\left(\left|X_{n(m_k)} - X\right| > \epsilon_k\right) \leqslant 2^{-k}$. Since

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|X_{n(m_k)} - X\right| > \epsilon_k\right) < \infty\,,$$

the Borel-Cantelli lemma implies $\mathbb{P}\left(\left|X_{n(m_k)} - X\right| > \epsilon_k \text{ i.o. }\right) = 0$, i.e., $X_{n(m_k)}$ converges to X a.s.

To show the converse, we use arguement by contradiction. If X_n doesn't converge to X in probability, then there exises some ϵ, $\delta > 0$ and a sequence $\{X_{n(m)}\}$ so that

$$\mathbb{P}\left(\left|X_{n(m)} - X\right| > \epsilon\right) > \delta$$

Thus any subsequence $X_{n(m_k)}$ doesn't converges almost surely to X, which is a contradiction.

As our second application of the Borel-Cantelli lemma,we get our first strong law of large numbers:

Example 2.10 ("Weak" strong law of large number)**.** Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_1 = \mu$. Let $S_n = X_1 + \cdots + X_n$. We will prove the strong law of large numbers, and by letting $X_i' = X_i - \mu$, we can suppose without loss of generality that $\mu = 0$. Then we want to show

$$\frac{S_n}{n} \to 0 \text{a.s.}\,.$$

It suffices to show

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon \text{ i.o.}\right) = 0\,,$$

by B-C lemma, we only need to show

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon\right) < \infty\,.$$

By Chebyshev inequality,

$$\mathbb{P}\left(|S_n| > n\epsilon\right) \leqslant \frac{\mathbb{E}\left(S_n^2\right)}{\epsilon^2 n^2} \leqslant \frac{\mathrm{Var}(X_1)}{\epsilon^2 n}\,.$$

But $\sum 1/n = \infty$, so this method failed. However, we can increase the order of the moments in Chebyshev inequality, note that it's noe easy to compute $\mathbb{E}|S_n|^3$, but

$$\mathbb{E}S_n^4 = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4 = \mathbb{E}\sum_{1\leqslant i,j,k,l\leqslant n} X_i X_j X_k X_l$$

Terms in the sum of the form $\mathbb{E}\left(X_i^3 X_j\right), \mathbb{E}\left(X_i^2 X_j X_k\right)$, and $\mathbb{E}\left(X_i X_j X_k X_l\right)$ are 0 (if i, j, k, l are distinct) since the expectation of the product is the product of the expectations, and in each case one of the terms has expectation 0. The only terms that do not vanish are those of the form $\mathbb{E}X_i^4$ and $\mathbb{E}X_i^2 X_j^2 = \left(\mathbb{E}X_i^2\right)^2$. There are n and $3n(n-1)$ of these terms, respectively. (In the second case we can pick the two indices in $n(n-1)/2$ ways, and with the indices fixed, the term can arise in a total of 6 ways.) The last observation implies

$$\mathbb{E}S_n^4 = n\mathbb{E}X_1^4 + 3\left(n^2 - n\right)\left(\mathbb{E}X_1^2\right)^2$$

Assume $\mathbb{E}X_1^4 < \infty$, then $\mathbb{E}S_n^4 \leqslant Cn^2$ for some constant C. Chebyshev's inequality gives us

$$\mathbb{P}\left(|S_n| > n\epsilon\right) \leqslant \frac{\mathbb{E}\left(S_n^4\right)}{(n\epsilon)^4} \leqslant \frac{C}{n^2\epsilon^4}\,.$$

Summing on n and using the Borel-Cantelli lemma gives $\mathbb{P}\left(|S_n| > n\epsilon \text{ i.o.}\right) = 0$. for any $\epsilon > 0$. Thus we proved

Theorem. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = \mu$ and $\mathbb{E}X_i^4 < \infty$. Put $S_n = X_1 + \cdots + X_n$ then*

$$\frac{S_n}{n} \to \mu \text{ a.s.}$$

Second Borel-cantelli lemma For independent events, the necessary condition for $\mathbb{P}(\limsup A_n) > 0$ is sufficient for $\mathbb{P}(\limsup A_n) = 1$.

Theorem 2.8 (The second Borel-Cantelli lemma.). *If the events A_n are independent then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ implies $\mathbb{P}(A_n \text{ i.o. }) = 1$.*

Proof. Note that

$$\mathbb{P}(A_n \text{ i.o. }) = \lim_{k\to\infty} \mathbb{P}(\cup_{n=k}^{\infty} A_n) \ .$$

For given k,

$$\mathbb{P}(\cup_{n=k}^{\infty} A_n) = 1 - \mathbb{P}(\cap_{n=k}^{\infty} A_n^c) = 1 - \prod_{n=k}^{\infty} (1 - \mathbb{P}(A_n)) \ ,$$

since $\sum_{n=k}^{\infty} \mathbb{P}(A_n) = \infty$,

$$\prod_{n=k}^{\infty} (1 - \mathbb{P}(A_n)) = 0 \, .$$

Thus for each k,

$$\mathbb{P}(\cup_{n=k}^{\infty} A_n) = 1 \, ,$$

which gives the desired result. □

A typical application of the second Borel-Cantelli lemma is:

Example 2.11. Let $X_1, X_2, \cdots$ be i.i.d. We now find necessary and sufficient conditions for

$$\frac{X_n}{n} \to 0 \ \text{ a.s.} \quad \text{and} \quad \frac{\max_{1\leqslant m\leqslant n} X_m}{n} \to 0 \ \text{ a.s.}$$

Clearly, we should use B-C lemma to deal with convergence almost sure.

(i) Note that (i) holds iff for all $\epsilon > 0$, $\mathbb{P}(|X_n| \geqslant n\epsilon \text{ i.o.}) = 0$. By second B-C lemma, since $\{X_n\}$ is i.i.d., this is equivalent to

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geqslant n\epsilon) < \infty \Leftrightarrow \sum_{n=1}^{\infty} \mathbb{P}(|X_1| \geqslant n\epsilon) < \infty \Leftrightarrow \mathbb{E}|X_1| < \infty \, .$$

Moreover, if $\mathbb{E}|X_1| = \infty$, by the second B-C lemma we have $|X_n| \geqslant Mn$ i.o. for any $M > 0$, thus

$$\limsup_{n\to\infty} \frac{|X_n|}{n} \to \infty \, , \quad \text{a.s.}$$

(ii) First, note that

$$\frac{\max_{m\leqslant n} X_m}{n} \to 0 \Leftrightarrow \limsup_{n\to\infty} \frac{X_n}{n} \leqslant 0\,.$$

Then $\limsup \frac{X_n}{n} \leqslant 0$ iff for any $\epsilon > 0$, $\mathbb{P}(X_n \geqslant n\epsilon \text{ i.o.}) = 0$ By second B-C lemma, since $\{X_n\}$ is i.i.d., this is equivalent to

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \geqslant n\epsilon) < \infty \Leftrightarrow \sum_{n=1}^{\infty} \mathbb{P}(X_1 \geqslant n\epsilon) < \infty \Leftrightarrow \mathbb{E}X_1^+ < \infty\,.$$

The next example is both intresting and technical.

Example 2.12 (Head runs)**.** Let $\{X_n : n \in \mathbb{Z}\}$, be i.i.d. with

$$\mathbb{P}\left(X_n = 1\right) = \mathbb{P}\left(X_n = -1\right) = \frac{1}{2}\,.$$

Let

$$l_n = \begin{cases} \max\{m \geqslant 1 : X_{n-m+1} = \cdots = X_n = 1\}\,, & X_n = 1\,. \\ 0\,, & X_n = -1\,. \end{cases}$$

be the length of the run of $+1$'s at time n. Let

$$L_n = \max_{1\leqslant m\leqslant n} l_m$$

be the longest run at time n.

In fact $(l_n)_{n\geqslant 1}$ is Markov chian on $\mathbb{N}$, more precisely, is *success-run chian.* The reason we let $\{X_n\}$ indexed by $\mathbb{Z}$ instead of $\mathbb{N}$ is that we want $(l_n)_{n\geqslant 1}$ starting at equlibrium, i.e., for all n,

$$\mathbb{P}\left(l_n = k\right) = \frac{1}{2^{i+1}} \text{ for } k \geqslant 0\,.$$

The result we are going to prove to show, by B-C lemma, is

Theorem. *Almost surely,*

$$\textit{(i)}\ \limsup_{n\to\infty} \frac{l_n}{\log_2 n} = 1, \quad \textit{and (ii)}\ \lim_{n\to\infty} \frac{L_n}{\log_2 n} = 1\,.$$

Proof of (i). Clearly, it suffices to show for any given $\epsilon > 0$,

$$\mathbb{P}\left(\frac{l_n}{\log_2 n} > 1+\epsilon, \quad \text{i.o.}\right) = 0\,, \quad \text{and} \quad \mathbb{P}\left(\frac{l_n}{\log_2 n} > 1-\epsilon, \quad \text{i.o.}\right) = 1 \quad (2.5)$$

The first equation in (2.5) follows from the first B-C lemma, since

$$\mathbb{P}\left(l_n > (1+\epsilon)\log_2 n\right) \leqslant \frac{1}{2^{(1+\epsilon)\log_2 n}} = \frac{1}{n^{1+\epsilon}}\,,$$

is summable. But we can not use the second B-C lemma to show the second equation in (2.5), since $(l_n)_{n\geqslant 1}$ is not independent !

To find a sequence of indenpendent r.v.'s to use second B-C lemma, let $r_1 = 1$, $r_2 = 2$ and

$$r_n = r_{n-1} + [\log_2 n]\ .$$

Let

$$A_n = \{X_m = 1 \text{ for } r_{n-1} < m \leqslant r_n\}\ ,$$

Then have

$$\mathbb{P}\left(A_n\right) = \frac{1}{2^{[\log_2 n]}} \sim \frac{1}{n}\,,$$

so it follows from the second Borel Cantelli lemma that $\mathbb{P}(A_n \text{ i.o.}) = 1$, and hence $l_{r_n} \geqslant [\log_2 n]$ i.o., since $r_n \leqslant n\log_2 n$ we have

$$\frac{l_{r_m}}{\log_2\left(r_n\right)} \geqslant \frac{[\log_2 n]}{\log_2 n + \log_2\log_2 n} > 1-\epsilon\,.\ \text{i.}\ \text{o.}$$

and the second equation in (2.5) follows. □

Proof of (ii). It follows from Exercise 4 that

$$\limsup_{n\to\infty} \frac{L_n}{\log_2 n} = 1 \quad \text{a.s.}$$

It suffices to show that,

$$\liminf_{n\to\infty} \frac{L_n}{\log_2 n} \geqslant 1 \quad \text{a.s.}$$

or equivalently, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{L_n}{\log_2 n} < 1-\epsilon, \quad \text{i.o.}\right) = 0\,.$$

We want to use B-C lemma, however

$$\mathbb{P}\left(\frac{L_n}{\log_2 n} < 1-\epsilon\right) = \mathbb{P}\left(\bigcap_{m=1}^{n} \{l_m < (1-\epsilon)\log_2 m\}\right),$$

is hard wo compute since $(l_n)_{n\geqslant 1}$ is not independent. Thus, as the proof of (i), to find a sequence of independent r.v.'s, we break the first n trials into disjoint blocks of length $\lceil (1-\epsilon)\log_2 n\rceil$ [1] on which the variables are all 1 with probability

$$\frac{1}{2^{\lceil (1-\epsilon)\log_2 n\rceil}} \geqslant \frac{1}{n^{1-\epsilon}}.$$

to conclude that if n is large enough so that

$$\frac{n}{\lceil (1-\epsilon)\log_2 n\rceil} \geqslant \frac{n}{\log_2 n},$$

then

$$\mathbb{P}\left(L_n \leqslant (1-\epsilon)\log_2 n\right) \leqslant \left(1-\frac{1}{n^{1-\epsilon}}\right)^{\frac{n}{\log_2 n}} \leqslant \exp\left(\frac{1}{n^{\epsilon}\log_2 n}\right),$$

which is summable, so the Borel-Cantelli lemma implies

$$\liminf_{n\to\infty} \frac{L_n}{\log_2 n} \geqslant 1 \quad \text{a.s.} \qquad \square$$

A generalization of second B-C lemma The next result extends the second Borel-Canteli lemma and sharpens its conclusion. B-C lemma asserts that if $\{A_n\}$ are indenpendet, $\sum_{n=1}^{\infty}\mathbb{P}(A_n)=\infty$, then $\sum_{m=1}^{n} 1_{A_m} \uparrow \infty$ as $n\to\infty$. Then next theorem gives the "speed" of divergence.

Theorem 2.9. *Suppose $A_1, A_2, \cdots$ are pairwise independent and $\sum_{n=1}^{\infty}\mathbb{P}(A_n) = \infty$, then as $n\to\infty$*

$$\frac{\sum_{m=1}^{n} 1_{A_m}}{\sum_{m=1}^{n}\mathbb{P}(A_m)} \to 1 \quad \text{a.s.}$$

[1] $\lceil x\rceil := \inf\{k\in\mathbb{Z} : k\geqslant x\}$.

Proof. Let $X_m = 1_{A_m}$ and let $S_n = X_1 + \cdots + X_n$, the desired result is exactly

$$\frac{S_n}{\mathbb{E}S_n} \to 1 \quad \text{a.s.}$$

It's natural to ask if we can show this by B-C lemma, so we want to show

$$\mathbb{P}\left(\frac{|S_n - \mathbb{E}S_n|}{\mathbb{E}S_n} > \epsilon \text{ i.o.}\right) = 0\,. \tag{2.6}$$

Clearly, we have $\operatorname{Var}(S_n) \leqslant \mathbb{E}(S_n)$, and Chebyshev's inequality implies

$$\mathbb{P}\left(\frac{|S_n - \mathbb{E}S_n|}{\mathbb{E}S_n} > \epsilon\right) \leqslant \frac{\operatorname{Var}(S_n)}{\epsilon^2 (\mathbb{E}S_n)^2} \leqslant \frac{1}{\epsilon^2 \mathbb{E}S_n}\,.$$

The problem is, we don't konw the series $\sum_{n=1}^{\infty} \frac{1}{\mathbb{E}S_n}$ is convergent or not, so we cannot use B-C lemma to deduce (2.6).

However, we can try to prove the result for a subsequence $\{S_{n_k}\}$ first, and then try to use some methods like *squeeze theorem* to show the result for $\{S_n\}$ also holds.

Let $n_k = \inf\{n : \mathbb{E}S_n \geqslant k^2\}$, let $T_k = S_{n_k}$ and note that the definition and $\mathbb{E}X_m \leqslant 1$ imply $k^2 \leqslant \mathbb{E}T_k \leqslant k^2 + 1$, thus

$$\mathbb{P}\left(\frac{|T_k - \mathbb{E}T_k|}{\mathbb{E}T_k} > \epsilon\right) \leqslant \frac{1}{\epsilon^2 \mathbb{E}T_k} \leqslant \frac{1}{\epsilon^2 k^2}\,,$$

which is summable, then we have

$$\mathbb{P}\left(\frac{|T_k - \mathbb{E}T_k|}{\mathbb{E}T_k} > \epsilon \text{ i.o.}\right) = 0\,.$$

Since ϵ is arbitrary, it follows that

$$\frac{T_k}{\mathbb{E}T_k} \to 1 \quad \text{a.s.}$$

To show $\frac{S_n}{\mathbb{E}S_n} \to 1$ a.s., take an ω so that $\frac{T_k(\omega)}{\mathbb{E}T_k} \to 1$ and observe that if $n_k \leqslant n < n_{k+1}$ then

$$\frac{T_k(\omega)}{\mathbb{E}T_{k+1}} \leqslant \frac{S_n(\omega)}{\mathbb{E}S_n} \leqslant \frac{T_{k+1}(\omega)}{\mathbb{E}T_k}$$

To show that the terms at the left and right ends $\to 1$, we rewrite the last inequalities as

$$\frac{\mathbb{E}T_k}{\mathbb{E}T_{k+1}} \cdot \frac{T_k(\omega)}{\mathbb{E}T_k} \leqslant \frac{S_n(\omega)}{\mathbb{E}S_n} \leqslant \frac{T_{k+1}(\omega)}{\mathbb{E}T_{k+1}} \cdot \frac{\mathbb{E}T_{k+1}}{\mathbb{E}T_k}$$

Clearly, it suffices to show $\frac{\mathbb{E}T_{k+1}}{\mathbb{E}T_k} \to 1$, but this follows from

$$k^2 \leqslant \mathbb{E}T_k \leqslant \mathbb{E}T_{k+1} \leqslant (k+1)^2 + 1\,. \qquad \square$$

REMARK. We will encounter the method in this proof again in the proof of SLLN, Theorem 2.10.

Example 2.13 (Record values)**.** Let $X_1, X_2, \cdots$ be a sequence of random variables and think of X_k as the distance for an individual's k th high jump or shot-put toss so that $A_k = \{X_k > \sup_{1\leqslant j<k} X_j\}$ is the event that a record occurs at time k. Ignoring the fact that an athlete's performance may get better with more experience or that injuries may occur, we will suppose that $X_1, X_2 \cdots$ are i.i.d. with a distribution $F(x)$ that is continuous. Even though it may seem that the occurrence of a record at time k will make it less likely that one will occur at time $k+1$, we claim

Lemma. *$\{A_k\}$ is independent, and $\mathbb{P}(A_k) = \frac{1}{k}$.*

Proof. Since F is countinuous function, $\mathbb{P}(X_i = X_j) = 0$ for any i, j. Then note that for any $1 \leqslant i \leqslant k$, by symmetry,

$$\begin{aligned}\mathbb{P}(A_k) &= \int_{\mathbb{R}^k} 1_{\{x_k > \sup\limits_{1\leqslant j<k} x_j\}}\, dF(x_1)\cdots dF(x_k) \\ &= \int_{\mathbb{R}^k} 1_{\{x_i > \sup\limits_{\substack{1\leqslant j\leqslant k \\ j\neq i}} x_j\}}\, dF(x_1)\cdots dF(x_k)\,.\end{aligned}$$

and

$$\sum_{i=1}^{k} \int_{\mathbb{R}^k} 1_{\{x_i > \sup\limits_{\substack{1\leqslant j\leqslant k \\ j\neq i}} x_j\}}\, dF(x_1)\cdots dF(x_k) = 1\,.$$

Thus $\mathbb{P}(A_k) = \frac{1}{k}$. Pick any $i < j$, then for the same reason for any $1 \leqslant m \leqslant i$

$$\begin{aligned}\mathbb{P}(A_i \cap A_j) &= \int_{\mathbb{R}^j} 1_{\{x_i > \sup\limits_{1\leqslant l<i} x_l\}} 1_{\{x_j > \sup\limits_{1\leqslant l<j} x_l\}} dF(x_1)\cdots dF(x_j) \\ &= \int_{\mathbb{R}^j} 1_{\{x_m > \sup\limits_{\substack{1\leqslant l\leqslant i \\ l\neq i}} x_j\}} 1_{\{x_j > \sup\limits_{1\leqslant l<j} x_l\}} dF(x_1)\cdots dF(x_j)\,,\end{aligned}$$

and

$$\sum_{m=1}^{i} \int_{\mathbb{R}^j} 1_{\{x_m > \sup\limits_{\substack{1\leqslant l\leqslant i \\ l\neq i}} x_j\}} 1_{\{x_j > \sup\limits_{1\leqslant l<j} x_l\}} dF(x_1)\cdots dF(x_j) = \mathbb{P}(A_j)\,.$$

Thus $\mathbb{P}(A_i \cap A_j) = \frac{1}{ij}$, so they are independent. Using the same method, we can show that $\{A_n\}$ is independent. □

Using Theorem 2.9 and the by now familiar fact that $\sum_{m=1}^{n} \frac{1}{m} \sim \log n$, we have

Theorem. *If $R_n = \sum_{m=1}^{n} 1_{A_m}$ is the number of records at time n, then*

$$\frac{R_n}{\log n} \to 1, \text{ a.s.}$$

We should emphasize that the last result is independent of the distribution F (as long as it is continuous).

EXERCISE

Exercise 4. Suppose $\{x_n\}$ and $\{a_n\}$ are two sequence of real numbers, $a_n \uparrow \infty$. Let $M_n = \max_{1\leqslant m\leqslant n} x_m$ and let $c \in [0, \infty)$. Then

$$\limsup_{n\to\infty} \frac{x_n}{a_n} = c \Leftrightarrow \limsup_{n\to\infty} \frac{M_n}{a_n} = c$$

Exercise 5. Let A_n be a sequence of independent events with $\mathbb{P}(A_n) < 1$ for all n. Show that $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = 1$ implies $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and hence $\mathbb{P}(A_n \text{ i.o.}) = 1$
Hint: $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = 1$ implies $\mathbb{P}(\cap_{n=1}^{\infty} A_n^c) = 0$.

EXERCISE 6. If $\mathbb{P}(A_n) \to 0$ and $\sum_{n=1}^{\infty} \mathbb{P}(A_n^c \cap A_{n+1}) < \infty$. Prove that

$$\mathbb{P}(A_n \text{ i.o. }) = 0.$$

Hint: $\{A_n \text{ i.o.}\} \cap \{A_n^c \text{ i.o.}\} \subset \{A_{n+1} \cap A_n^c \text{ i.o.}\}$

EXERCISE 7 (Kochen-Stone lemma). Suppose $\sum \mathbb{P}(A_k) = \infty$. Show that if

$$\limsup_{n\to\infty} \frac{\left[\sum_{k=1}^{n} \mathbb{P}(A_k)\right]^2}{\sum_{1\leqslant j,k\leqslant n} \mathbb{P}(A_j \cap A_k)} = \alpha > 0$$

then $\mathbb{P}(A_n \text{ i.o. }) \geqslant \alpha$. The case $\alpha = 1$ contains the second Borel-Cantelli lemma.

Hint: let $X_n = \sum_{k=1}^{n} 1_{A_k}$, then by C-B-S inequality, we have

$$\frac{(\mathbb{E}X_n)^2}{\mathbb{E}X_n^2} \leqslant \mathbb{P}(X_n > 0)\,.$$

EXERCISE 8. If X_n is any sequence of random variables, there are constants $c_n \to \infty$ so that

$$\frac{X_n}{c_n} \to 0 \quad \text{a.s.}$$

Hint: take b_n statisfying $\mathbb{P}(X_1 > b_n) \leqslant \frac{1}{2^n}$, then let $c_n = nb_n$.

EXERCISE 9. Let $X_1, X_2, \cdots$ be independent. Show that $\sup_{n\geqslant 1} X_n < \infty$ a.s. if and only if $\sum_{n=1}^{\infty} P(X_n > A) < \infty$ for some A.

Hint: note that $\{\sup_{n\geqslant 1} X_n < \infty\} = \cup_m \liminf\{X_n \leqslant m\}$.

EXERCISE 10. Let $X_1, X_2, \cdots$ be i.i.d. with exponential(1) distribution, is the holding time in a Poisson preocess with rate 1. Let $M_n = \max_{1\leqslant m\leqslant n} X_m$ be the longest holding time. Show that, almost surely,

$$\text{(i) } \limsup_{n\to\infty} \frac{X_n}{\log n} = 1, \quad \text{and (ii) } \lim_{n\to\infty} \frac{M_n}{\log n} = 1\,.$$

Hint: see Example 2.12.

EXERCISE 11. Let X_n be independent Poisson r.v. 's with $\mathbb{E}X_n = \lambda_n$, and let $S_n = X_1 + \cdots + X_n$. Show that if $\sum \lambda_n = \infty$ then

$$\frac{S_n}{\mathbb{E}S_n} \to 0 \quad \text{a.s.}$$

Hint: note that if independent, Poisson (λ) + Poisson (μ) = Poisson $(\lambda+\mu)$.

2.3 Strong Law of Large Numbers (I)

We are now ready to give Etemadi' s (1981) proof of

Theorem 2.10 (SLLN). *Let $X_1, X_2, \cdots$ be pairwise independent, identically distributed r.v.'s with $\mathbb{E}|X_1| < \infty$. Let $\mathbb{E}X_1 = \mu$ and $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to \mu \quad \text{a.s.}$$

Proof. It natural to ask if we can use B-C lemma and Chebyshev inequality to show this, however, we don't konw $\mathbb{E}|X_1|^2 < \infty$ or not. As in the proof of the WLLN, we begin by truncation. A first question is, how to choose the truncation level ?

Let $Y_n = X_n 1_{\{|X_n| \leqslant b_n\}}$ for all n. Clearly, a good idea is that we ensure

$$\mathbb{P}\left(X_n \neq Y_n \text{ i.o.}\right) = 0\,, \tag{2.7}$$

then it suffices to show

$$\frac{T_n}{n} \to \mu \quad \text{a.s.} \tag{2.8}$$

where $T_n = Y_1 + \cdots Y_n$ for each n. Note that, by B-C lemma, (2.7) holds iff

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| > b_n\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > b_n\right) < \infty\,.$$

Thus we let $b_n = n$ for each n, and then (2.7) holds.

To show (2.8), by Chebyshev's inequality,

$$\mathbb{P}\left(\frac{|T_n - \mathbb{E}T_n|}{n} > \epsilon\right) \leqslant \frac{\mathrm{Var}\left(T_n\right)}{\epsilon^2\, n^2}\,.$$

The problem is, we don't konw the series $\sum_{n=1}^{\infty} \frac{\mathrm{Var}(T_n)}{n^2}$ is convergent or not, so we cannot use B-C lemma.

As in the the proof of Theorem 2.9, we can prove the result first for a subsequence and then use monotonicity to control the values in between. But note that we need that $\{T_n\}$ is increasing, Etemadi's inspiration was

that since $\{X_n^+\}$ and $\{X_n^-\}$ satisfy the assumptions of the theorem and $X_n = X_n^+ - X_n^-$, we can without loss of generality suppose $X_n \geqslant 0$.

We now let pick $k(n)$. Chebyshev's inequality implies that if $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|T_{k(n)} - \mathbb{E}T_{k(n)}\right| > \epsilon k(n)\right) \leqslant \epsilon^{-2} \sum_{n=1}^{\infty} k(n)^{-2} \operatorname{Var}\left(T_{k(n)}\right)$$
$$= \epsilon^{-2} \sum_{n=1}^{\infty} k(n)^{-2} \sum_{m=1}^{k(n)} \operatorname{Var}\left(Y_m\right) = \epsilon^{-2} \sum_{m=1}^{\infty} \operatorname{Var}\left(Y_m\right) \sum_{n:k(n)\geqslant m} k(n)^{-2},$$

where we have used Fubini's theorem to interchange the two summations of nonnegative terms. To guarantee the sum is finite, we need the following lemma.

Lemma 2.11. $\sum_{n=1}^{\infty} \frac{\operatorname{Var}(Y_n)}{n^2} \leqslant C\,\mathbb{E}\,|X_1| < \infty$, *where* C *is a constant.*

Proof. To bound the sum, we observe for any n,

$$\operatorname{Var}\left(Y_n\right) \leqslant \mathbb{E}\left(Y_n^2\right) = \int 1_{\{|y|\leqslant n\}} y^2 dF(y),$$

where F is the c.d.f. of X_1. Thus we have

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var}\left(Y_n\right)}{n^2} \leqslant \sum_{n=1}^{\infty} \frac{1}{n^2} \int 1_{\{|y|\leqslant n\}} y^2 dF(y) = \int y^2 \sum_{n=1}^{\infty} \frac{1}{n^2} 1_{\{|y|\leqslant n\}} dF(y)$$
$$\leqslant \int C\,|y| dF(y) = C\,\mathbb{E}|X_1| < \infty.$$

as required. □

We are going to make the series $\sum_{n:k(n)\geqslant m} k(n)^{-2} \ll 1/m^2$, by Lemma 2.11. We pick $k(n)$ carefully. First, we can not let $k(n)$ be ploynomial n^α since

$$\sum_{n\geqslant m^{1/\alpha}} \frac{1}{n^{2\alpha}} \sim \int_{m^{1/\alpha}}^{\infty} \frac{1}{x^{2\alpha}} dx \sim \frac{1}{m^{2-1/\alpha}}.$$

Next, let's try exponential function, let $k(n) = \alpha^n$. Then we find that

$$\sum_{n:\alpha\geqslant m^{1/n}} \alpha^{-2n} \sim \frac{1}{m^2}.$$

Thus we let $k(n) = [\alpha^n]$ and $[\alpha^n] \geqslant \alpha^n/2$ for $n \geqslant 1$, so summing the geometric series and noting that the first term domaineted by $\leqslant m^{-2}$:

$$\sum_{n:\alpha^n \geqslant m} [\alpha^n]^{-2} \leqslant 4 \sum_{n:\alpha^n \geqslant m} \alpha^{-2n} \leqslant C_\alpha \frac{1}{m^2} .$$

Where $C_\alpha > 0$ is a constant relative to α. Combining our computations shows

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|T_{k(n)} - \mathbb{E}T_{|k(n)}\right| > \epsilon k(n)\right) \leqslant C_\alpha \sum_{m=1}^{\infty} \frac{\operatorname{Var}(Y_m)}{\epsilon^2 m^2} < \infty$$

by Lemma 2.11. Since ϵ is arbitrary

$$\frac{T_{k(n)} - \mathbb{E}T_{k(n)}}{k(n)} \to 0 \quad \text{a.s.}$$

The dominated convergence theorem implies $\mathbb{E}Y_n \to \mu$, so $\mathbb{E}T_{k(n)}/k(n) \to \mu$ and we have shown

$$\frac{T_{k(n)}}{k(n)} \to \mu \quad \text{a.s.}$$

To handle the intermediate values, we observe that if $k(n) \leqslant m < k(n+1)$, then

$$\frac{T_{k(n)}}{k(n+1)} \leqslant \frac{T_m}{m} \leqslant \frac{T_{k(n+1)}}{k(n)} ,$$

so recalling $k(n) = [\alpha^n]$, we have $\frac{k(n+1)}{k(n)} \to \alpha$ and

$$\frac{1}{\alpha}\mu \leqslant \liminf_{n\to\infty} \frac{T_n}{n} \leqslant \limsup_{n\to\infty} \frac{T_n}{n} \leqslant \alpha\mu .$$

Since $\alpha > 1$ is arbitrary, the proof is complete. □

Theorem 2.12 (SLLN for i.i.d. sequence). *Let $\{X_n\}$ be i.i.d. r.v.'s and $S_n = X_1 + \cdots + X_n$. Then there exists some $\mu \in \mathbb{R}$ such that*

$$\frac{S_n}{n} \to \mu \quad \text{a.s.}$$

if and only if $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$.

Proof. If $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$, by Theorem 2.10 the theorem holds.

If exists some $\mu \in \mathbb{R}$ such that $\frac{S_n}{n} \to \mu$, then we have

$$\frac{X_n}{n} \to 0 \text{ a.s.}$$

By Example 2.11, we konw $\mathbb{E}|X_1| < \infty$. Then by SLLN, $\mathbb{E}X_1 = \mu$. □

Infinite mean The next result shows that the strong law of large number holds whenever EX_i exists.

Theorem 2.13. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i^+ = \infty$ and $\mathbb{E}X_i^- < \infty$. Let $S_n = X_1 + \cdots + X_n$ then*

$$\frac{S_n}{n} \to \infty \quad a.s.$$

Proof. Let $M > 0$ and $X_i^M = X_i \wedge M$. The X_i^M are i.i.d. with $\mathbb{E}\left|X_i^M\right| < \infty$, so if $S_n^M = X_1^M + \cdots + X_n^M$ then Theorem 2.10 implies $S_n^M/n \to \mathbb{E}X_i^M$. since $X_i \geqslant X_i^M$, it follows that

$$\liminf_{n\to\infty} \frac{S_n}{n} \geqslant \lim_{n\to\infty} \frac{S_n^M}{n} = \mathbb{E}X_i^M .$$

By monotone convergence theorem, $\mathbb{E}\left(X_i^M\right)^+ \uparrow \mathbb{E}X_i^+ = \infty$ as $M \uparrow \infty$, so $\mathbb{E}X_i^M = \mathbb{E}\left(X_i^M\right)^+ - \mathbb{E}\left(X_i^M\right)^- \uparrow \infty$, which implies the desired result. □

Theorem 2.14. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}|X_1| = \infty$ and let $S_n = X_1 + \cdots + X_n$. Then*

$$\limsup_{n\to\infty} \frac{|S_n|}{n} = \infty \quad a.s.$$

Proof. By Example 2.11, $\mathbb{E}|X_1| = \infty$ implies

$$\limsup_{n\to\infty} \frac{|X_n|}{n} = \infty \quad \text{a.s. ,}$$

since $|X_n| = |S_n - S_{n-1}| \leqslant |S_n| + |S_{n-1}|$, we have

$$\max\{|S_{n-1}|, |S_n|\} \geqslant \frac{|X_n|}{2},$$

it follows that

$$\limsup_{n\to\infty} \frac{|S_n|}{n} = \infty .$$ □

Applications The rest of this section is devoted to applications of the strong law of large numbers.

Example 2.14 (Renewal theory)**.** Let $\{X_n\}$ be i.i.d. r.v.'s with $X_1 > 0$. Let $T_n = X_1 + \cdots + X_n$ and think of T_n as the time of n th occurrence of some event. For a concrete situation, consider a diligent janitor who replaces a light bulb the instant it burns out. Suppose the first bulb is put in at time 0 and let X_i be the lifetime of the ith light bulb. In this interpretation, T_n is the time the n th light bulb burns out and

$$N(t) = \sum_{n=1}^{\infty} 1_{\{T_n \leqslant t\}}$$

is the number of light bulbs that have burnt out by time t.

Theorem. *If $\mathbb{E}X_1 = \mu \leqslant \infty$ then as $t \to \infty$, then as $t \to \infty$,*

$$\frac{N(t)}{t} \to \frac{1}{\mu} \quad \text{a.s.}$$

Proof. By SLLN, $\frac{T_n}{n} \to \mu$ a.s. From the definition of $N(t)$, it follows that $T_{N(t)} \leqslant t < T_{N(t)+1}$, so dividing through by $N(t)$ gives

$$\frac{T_{N(t)}}{N(t)} \leqslant \frac{t}{N(t)} \leqslant \frac{T_{N_{t+1}}}{N(t)+1} \cdot \frac{N(t)+1}{N(t)}$$

To take the limit, we note that since $T_n < \infty$ for all n, we have $N(t) \uparrow \infty$ as $t \to \infty$. Hence as $t \to \infty$,

$$\frac{T_{N(t)}}{N(t)} \to \mu\,, \ \frac{N(t)+1}{N(t)} \to 1\,, \ \text{a.s.}$$

From this it follows that

$$\frac{N(t)}{t} \to \frac{1}{\mu} \quad \text{a.s.}$$ □

Remark. The last argument shows that if $X_n \to X$ a.s. and $N(n) \to \infty$ a.s. then $X_N(n) \to X$ a.s. We should treat this with care because the analogous result for convergence in probability is false. If $X_n \in \{0, 1\}$ are independent with $\mathbb{P}(X = 1) = a_n \to 0$ and $\sum_n a_n = \infty$, then $X_n \to 0$ in probability, but if we let $N(n) = \inf\{m \geqslant n : X_m = 1\}$ then $X_{N(n)} = 1$ a.s.

Example 2.15 (Empirical distribution functions)**.** Let $X_1, X_2, \cdots$ be i.i.d. with distribution F and (for any given $\omega \in \Omega$) let

$$F_n(x, \omega) = \frac{1}{n} \sum_{m=1}^{n} 1_{\{X_m(\omega) \leqslant x\}}$$

$\{F_n(x)\}$ is the observed frequency of values that are $\leqslant x$, and is called the empirical process. By SLLN, we will show that, F_n converges uniformly to F almost surely.

Theorem (The Glivenko-Cantelli theorem). *As $n \to \infty$,*

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{a.s.}$$

Proof. Fix x and let $Y_n = 1_{\{X_n \leqslant x\}}$. Since the Y_n are i.i.d. with $\mathbb{E}Y_n = \mathbb{P}(X_n \leqslant x) = F(x)$, the strong law of large numbers implies that $F_n(x) = n^{-1} \sum_{m=1}^{n} Y_m \to F(x)$ a.s. By *Dini's theorem*, if F_n is a sequence of non-decreasing functions that converges pointwise to a bounded and continuous limit F then $\sup_x |F_n(x) - F(x)| \to 0$. However, the distribution function $F(x)$ may have jumps, so we have to work a little harder.

Again, fix x and let $Z_n = 1_{\{X_n < x\}}$. since the Z_n are i.i.d. with $\mathbb{E}Z_n = \mathbb{P}(X_n < x) = F(x^-) = \lim_{y \uparrow x} F(y)$, the strong law of large numbers implies that $F_n(x^-) = n^{-1} \sum_{m=1}^{n} Z_m \to F(x^-)$ a.s. Recall the proof of Dini's theorem, and we can prove the following lemma in the same way

Lemma. *$F_n(x)$ is a sequence of c.d.f.'s that converges pointwise to a c.d.f. F, and $F_n(x^-)$ converges pointwise to $F(x^-)$. Then $\sup_x |F_n(x) - F(x)| \to 0$.*

Then the Glivenko-Cantelli theorem holds. □

Example 2.16 (Shannon's theorem)**.** Let $X_1, X_2, \cdots \in \{1, \cdots, r\}$ be independent with $\mathbb{P}(X_i = k) = p(k) > 0$ for $1 \leqslant k \leqslant r$. Here we are thinking of $1, \cdots, r$ as the letters of an alphabet, and $X_1, X_2, \cdots$ are the successive letters produced by an information source. In this i.i.d. case, it is the proverbial monkey at a typewriter. Let $\pi_n(\omega) = p(X_1(\omega)) \cdots p(X_n(\omega))$ be

the probability of the realization we observed in the first n trials, or it is *likelihood function*. Since $\log \pi_n(\omega)$ is a sum of independent random variables, it follows from the strong law of large numbers that

$$-n^{-1} \log \pi_n(\omega) \to H := \mathbb{E} \log p(X_1) = -\sum_{k=1}^{r} p(k) \log p(k) \quad \text{a.s.}$$

The constant H is called the **entropy** of the source and is a measure of how random it is. The last result is the **asymptotic equipartition property**: for any $\epsilon > 0$, as $n \to \infty$

$$\mathbb{P}\left\{\exp(-n(H+\epsilon)) \leqslant \pi_n(\omega) \leqslant \exp(-n(H-\epsilon)\right\} \to 1 \ .$$

EXERCISE

EXERCISE 12. Show Lemma 2.11 by another method.

Sketch :

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbb{E}\left(Y_n^2\right)}{n^2} &= \sum_{n=1}^{\infty} \frac{1}{n^2} \mathbb{E}\left(X_1^2 1_{\{|X_1| \leqslant n\}}\right) \\
&= \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} \mathbb{E}\left(X_1^2 1_{\{k-1<|X_1| \leqslant k\}}\right) \\
&= \sum_{k=1}^{\infty} \sum_{n=k}^{n} \frac{1}{n^2} \mathbb{E}\left(X_1^2 1_{\{k-1<|X_1| \leqslant k\}}\right) \\
&\leqslant \sum_{k=1}^{\infty} \frac{C}{k} \mathbb{E}\left(X_1^2 1_{\{k-1<|X_1| \leqslant k\}}\right) \\
&\leqslant \sum_{k=1}^{\infty} C\, \mathbb{E}\left(|X_1| 1_{\{k-1<|X_1| \leqslant k\}}\right) = C\mathbb{E}\,|X_1|
\end{aligned}$$

as required.

EXERCISE 13 (Lazy janitor). Suppose the ith light bulb burns for an amount of time X_i and then remains burned out for time Y_i before being replaced. Suppose the X_i, Y_i are positive and independent, both of which have finite mean. Let $R(t)$ be the amount of time in $[0,t]$ that we have a working light bulb. Show that as $t \to \infty$,

$$\frac{R(t)}{t} \to \frac{\mathbb{E}X_1}{\mathbb{E}X_1 + \mathbb{E}Y_1} \quad \text{a.s.}$$

2.4 Random Series

In this section, we will pursue a second approach to the strong law of large numbers based on the convergence of random series. This approach has the advantage that it leads to estimates on the rate of convergence under moment assumptions, and to a negative result for the infinite mean case, which is stronger than the one in Theorem 2.14 . The first two results in this section are of considerable interest in their own right.

2.4.1 Zero-one Law

In probability theory, a zero-one law is a result that states that an event must have probability 0 or 1 and no intermediate value. We will show if $\{X_n\}$ are independent r.v.'s, then the event $\{\sum_{n=1}^{\infty} X_n \text{converges}\}$ has a zero-one law.

Kolmogorov's zero-one law To state the first result, we need some notation. Assume that $\{X_n\}$ are r.v.'s on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let

$$\mathcal{F}'_n \coloneqq \sigma\left(X_n, X_{n+1}, \cdots\right)$$

is the future after time n. Let

$$\mathcal{T} \coloneqq \bigcap_{n=1}^{\infty} \mathcal{F}'_n$$

is the remote future, called **tail σ-field**. The event in $\mathcal{T}$ is called **tail event** , and the measurable function [2] on $(\Omega, \mathcal{T})$ is called **tail measurable functions**. As usual, we turn two examples to help explain the definition.

Example 2.17. If $B_n \in \mathcal{B}$ then $\{X_n \in B_n \text{ i.o. }\} \in \mathcal{T}$, since for any $k \geqslant 0$,

$$\{X_n \in B_n \text{ i.o. }\} = \{X_{n+k} \in B_n \text{ i.o. }\}.$$

If we let $X_n = 1_{A_n}$ and $B_n = \{1\}$, this example becomes $\{A_n \text{ i.o. }\}$.

In the following two examples, let S_n denote $X_1 + \cdots + X_n$ for all n.

[2] means r.v. taking values in $[-\infty, \infty]$

Example 2.18. It is easy to check that $\{\lim_n X_n \text{ exists }\}$, $\{\sum_{n=1}^{\infty} X_n \text{ converges }\}$, and $\left\{\lim_n \frac{S_n}{n} \text{ exists }\right\}$ are tail events. Since for any $k \geqslant 0$,

$$\left\{\lim_n X_n \text{ exists }\right\} = \left\{\lim_n X_{n+k} \text{ exists }\right\},$$

$$\left\{\sum_{n=1}^{\infty} X_n \text{ converges }\right\} = \left\{\sum_{n=1}^{\infty} X_{n+k} \text{ converges }\right\}.$$

However, $\{X_n = 0 \text{ for all } n \geqslant 1\}$ and $\{\limsup_n S_n > 0\}$ are not tail event. For example, let $X_n \equiv 0$ for all $n \geqslant 2$, then $\{\limsup_n S_n > 0\} = \{X_1 > 0\}$.

Example 2.19.

$$\liminf_{n\to\infty} X_n\,,\ \limsup_{n\to\infty} X_n\,,\ \liminf_{n\to\infty} \frac{S_n}{n}\,,\ \limsup_{n\to\infty} \frac{S_n}{n}$$

are tail measurable functions. In general,

$$\limsup_n \frac{S_n}{c_n}$$

is tail measurable functions if $c_n \to \infty$. To see this, we observe that for any given $k \in \mathbb{N}_+$,

$$\left\{\limsup_{n\to\infty} \frac{S_n}{c_n} > a\right\} = \left\{\limsup_{n\to\infty} \frac{S_n - S_k}{c_n} > a\right\} \in \mathcal{F}'_{k+1}\,.$$

Kolmogorov's zero-one law specifies that a certain type of event, called a tail event, will either almost surely happen or almost surely not happen; that is, the probability of such an event occurring is zero or one.

Theorem 2.15 (Kolmogorov's 0-1 law). *$X_1, X_2 \cdots$ are independent. Then for any $A \in \mathcal{T}$, $\mathbb{P}(A) = 0$ or 1.*

Proof. We will show that A is independent of itself, that is,

$$\mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)\,,$$

so $\mathbb{P}(A) = \mathbb{P}(A)^2$,and hence $\mathbb{P}(A) = 0$ or 1 .

Take any $n \in \mathbb{N}_+$, let $\mathcal{F}_n \coloneqq \sigma(X_1, \cdots, X_n)$, then $\mathcal{F}_n$ is independent of $\mathcal{F}'_{n+1}$. So $\mathcal{F}_n$ is independent of $\mathcal{T}$. Therefore, $\cup_{n=1}^{\infty}\mathcal{F}_n$ is independent of $\mathcal{T}$. Note that $\cup_{n=1}^{\infty}\mathcal{F}_n$ is a π-systerm, actually it is a field, so $\sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$ is independent of $\mathcal{T}$. Since $\mathcal{T} \subset \sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$, $\mathcal{T}$ is independent with itself. □

Let f be a tail measurable function. Then for any $x \in \overline{\mathbb{R}}$, $\{f \leqslant x\} \in \mathcal{T}$ so $\mathbb{P}(f \leqslant x) = 0$ or 1. Let $C = \inf\{x \in \overline{\mathbb{R}} : \mathbb{P}(f \leqslant x) = 1\}$. It's easy to check that $\mathbb{P}(f = C) = 1$. Thus we get

Corollary 2.16. *Any tail measurable function equals to a constant C in $\overline{\mathbb{R}}$ almost surely.*

From the Kolmogorov's 0-1 law, we konw that if $\{X_n\}$ are independent r.v.'s, then the series $\sum_{n=1}^{\infty} X_n$ either a.s. converges or a.s. diverges, because $\{\sum_{n=1}^{\infty} X_n \text{ converges }\}$ is a tail event. Moreover, since $\liminf_n \frac{S_n}{n}$ and $\limsup_n \frac{S_n}{n}$ both are tail measurable functions, so they must be constant. And if $\lim_n \frac{S_n}{n}$ exists a.s., it must be constant. Thus we have

Corollary 2.17. *$\{X_n\}$ are independent r.v.'s, then the series $\sum_{n=1}^{\infty} X_n$ either a.s. converges or a.s. diverges. Moreover, if*

$$\lim_n \frac{S_n}{n}$$

exists a.s., it must be constant.

Hewitt-Sage 0-1 law. To state the motivation, let's see an example first.

Example 2.20. Let $\xi_1, \xi_2, \cdots$ be a sequence of independent Bernoulli random variables with $\mathbb{P}(\xi_n = 1) = p, \mathbb{P}(\xi_n = -1) = q, p + q = 1, n \geqslant 1$, and let $S_n = \xi_1 + \cdots + \xi_n$. It seems intuitively clear that in the symmetric case $\left(p = \frac{1}{2}\right)$ a "typical" path of the random walk $S_n, n \geqslant 1$, will visit the origin infinitely often, whereas when $p \neq \frac{1}{2}$, it will go off to infinity,that is

$$\mathbb{P}(S_n = 0 \text{ i.o. }) = \begin{cases} 1, & p = \frac{1}{2}. \\ 0, & p \neq \frac{1}{2}. \end{cases}$$

Let us observe again that $\{S_n = 0 \text{ i.o. }\}$ is not a tail event. Nevertheless, for a Bernoulli scheme, the probability of this event, just as for tail events, takes only the values 0 and 1.

The phenomenon in the preceding example is not accidental: it is a corollary of the Hewitt-Savage zero-one law, which for *i.i.d.* r.v.'s extends Kolmogorov's result to the class of *"permutable"* events (which includes the class of tail events). Let us give the essential definitions.

- A bijection π from $\mathbb{N}_+$ onto itself is said to be a **finite permutation**, if $\pi(n) \neq n$ for only finitely many n.
- Let $(X_n)_{n\geqslant 1}$ be a sequence of r.v.'s, if A is the event in $\sigma(X_n, n \geqslant 1)$, i.e., $A = \{(X_n) \in B\}$, where $B \in \mathcal{B}^{\mathbb{N}_+}$, then let $\pi(A)$ denotes the event $\{(X_{\pi_n}) \in B\}$. In other woeds,

  $$\pi : \sigma(X) \to \sigma(X)\,; \{(X_n) \in B\} \mapsto \{(X_{\pi_n}) \in B\}\,.$$

  is a bijection. An event A is called **permutable** or **symmetric** if $\pi(A) = A$ for every finite permutation π.
- The collection of permutable events is a σ -field, which is called the **exchangeable σ -field** and denoted by $\mathcal{E}$.

Example 2.21. Let $a \in \mathbb{R}$. Two examples of permutable events are

$$\text{(i) } \{S_n \leqslant a \text{ i.o. }\} \text{ and (ii) } \left\{\limsup_{n\to\infty} \frac{S_n}{n} \geqslant a\right\}.$$

Since in each case, the event is permutable because $S_n(\omega) = S_{\pi_n}(\omega)$ for sufficiently large n. The list of examples can be enlarged considerably by observing:

(iii) *Any events A in the tail σ-field $\mathcal{T}$ are permutable.*

To see this, observe that if $A \in \mathcal{F}'_{n+1}$, the occurrence of A is unaffected by a permutation of $X_1, \cdots, X_n$. (i) shows that the converse of (iii) is false.

The next result shows that for an i.i.d. sequence there is no difference between $\mathcal{E}$ and $\mathcal{T}$: They are both trivial .

Theorem 2.18 (Hewitt-Savage 0-1 law). *If $X_1, X_2, \cdots$ are i.i.d. and $A \in \mathcal{E}$ then $\mathbb{P}(A) \in \{0, 1\}$.*

REMARK. Note that Hewitt-Savage 0-1 law needs $\{X_n\}$ is i.i.d but Kolmogorov 0-1 law only needs $\{X_n\}$ is independent.

Proof. As in the proof of Kolmogorov's 0-1 law, we will show A is independent of itself, then $\mathbb{P}(A) \in \{0, 1\}$.

First, $X_1, X_2, \cdots$ are i.i.d., then the distributions of $(X_n)_{n\geqslant 1}$ coincides with $(X_{\pi_n})_{n\geqslant 1}$ for any finite permutation π. Hence, $\mathbb{P}(B) = \mathbb{P}(\pi B)$, for all $B \in \sigma(X)$. In words, *any finite permutation π is measure preserving isomorphism on $(\Omega, \sigma(X), \mathbb{P})$.*

Since $A \in \mathcal{E} \subset \sigma(X)$, we have learned in measure theory that there exists $\{A_n\}$ suth that $A_n \in \sigma(X_1, \cdots, X_n)$ so that $\mathbb{P}(A_n \Delta A) \to 0$. For $n = 1, 2, \cdots$, define

$$\pi_n(j) = \begin{cases} j + n & \text{if } 1 \leqslant j \leqslant n \\ j - n & \text{if } n + 1 \leqslant j \leqslant 2n \\ j & \text{if } j \geqslant 2n + 1 \end{cases}$$

is a sequence of finite permutations. Note that $A_n \in \sigma(X_1, \cdots, X_n)$, we have $\pi_n(A_n) \in \sigma(X_{\pi_n(1)}, \cdots, X_{\pi_n(n)}) = \sigma(X_{n+1}, \cdots, X_{2n})$, so A_n is independent of $\pi_n(A_n)$. Thus

$$\mathbb{P}\big(A_n \cap \pi_n A_n\big) = \mathbb{P}(A_n)\mathbb{P}\big(\pi_n A_n\big). \tag{2.9}$$

Since we have $\mathbb{P}(A_n) \to \mathbb{P}(A)$, we only need to show $\mathbb{P}\big(\pi_n A_n\big) \to \mathbb{P}(A)$ and $\mathbb{P}\big(A_n \cap \pi_n A_n\big) \to \mathbb{P}(A)$. Note that π_n is measure preserving, A is permutable

$$\begin{aligned} \mathbb{P}\big(\pi_n A_n \Delta A\big) &= \mathbb{P}\big(\pi_n A \Delta \pi_n(A_n))\big) \\ &= \mathbb{P}\left(\pi_n(A \Delta A_n)\right) = \mathbb{P}\left(A \Delta A_n\right) \to 0\,. \end{aligned}$$

So $\mathbb{P}\big(\pi_n A_n\big) \to \mathbb{P}(A)$. And

$$\mathbb{P}\bigg(A\Delta\big(A_n \cap \pi_n A_n\big)\bigg) \leqslant \mathbb{P}\left(A\Delta A_n\right) + \mathbb{P}\big(A\Delta\pi_n A_n\big) \to 0\,,$$

where we use $E\Delta(F\cap G) \subset (E\Delta F)\cup(E\Delta G)$. Thus $\mathbb{P}\big(A_n \cap \pi_n A_n\big) \to \mathbb{P}(A)$. Therefore, let $n \to \infty$ in (2.9), we get $\mathbb{P}(A) = \mathbb{P}(A)^2$. □

2.4.2 Convergence of Series

The center problem in this section is that if $\{X_n\}$ be independent r.v.'s , what conditions are given can make $\sum_{n=1}^{\infty} X_n$ converges with probability one? Frist, recall that in measure theory we have learned that, random variables $\{\xi_n\}$ converges to ξ a.s. if and only if

$$\lim_{k\to\infty} \mathbb{P}\big(\sup_{n\geqslant k} |\xi_n - \xi| > \epsilon\big) = 0\,.$$

When we dont konw what ξ is, we can use the form of Cauchy sequence to rewrite the conditon.

Lemma 2.19. *Let $\{X_n\}$ be r.v.'s and denote $S_n = \sum_{k=1}^{n} X_k$. Then $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\big(\sup_{k\geqslant n} |S_k - S_n| > \epsilon\big) = 0\,. \tag{2.10}$$

Proof. Sufficiency: for all $j \in \mathbb{N}_+$, there exists $N_j \in \mathbb{N}$ such that

$$\mathbb{P}\big(\sup_{k\geqslant N_j} |S_k - S_{N_j}| > \frac{1}{j}\big) \leqslant \frac{1}{2^j}\,.$$

By B-C lemma,

$$\mathbb{P}\big(\sup_{k\geqslant N_j} |S_k - S_{N_j}| > \frac{1}{j} \text{ i.o.}\big) = 0\,,$$

thus

$$\mathbb{P}\left(\liminf_{j\to\infty}\left\{\sup_{k\geqslant N_j} |S_k - S_{N_j}| \leqslant \frac{1}{j}\right\}\right) = 1\,,$$

which deduces that almost surely, $\{S_n\}$ is Cauchy sequence, hence it converges almost surely.

Necessity: Suppose $\{S_n\}$ converges to random bariable S almost surely, then for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\sup_{k\geqslant n} |S_k - S| > \epsilon\big) = 0\,.$$

Note that

$$\left\{\sup_{k\geqslant n} |S_k - S_n| > \epsilon\right\} \subset \left\{\sup_{k\geqslant n} |S_k - S| > \frac{\epsilon}{2}\right\}.$$

Thus

$$\lim_{n\to\infty} \mathbb{P}\big(\sup_{k\geqslant n} |S_k - S_n| > \epsilon\big) = 0\,. \qquad \square$$

Kolmogorov's two-series theorem We turn now to our results on convergence of series. We will give a convinient sufficient condition for the convergence of random series.

Lemma 2.20 (Kolmogorov's maximal inequality I). *Let $X_1, \cdots, X_n$ are independent with $\mathbb{E}X_i = 0$ and $\mathrm{Var}\,(X_i) < \infty$ for all $1 \leqslant i \leqslant n$. Then for any $a > 0$, we have*

$$\mathbb{P}\left(\max_{1\leqslant k\leqslant n} |S_k| \geqslant a\right) \leqslant \frac{\mathbb{E}S_n^2}{a^2} \tag{2.11}$$

Remark. Under the same hypotheses, Chebyshev's inequality gives only

$$\mathbb{P}\left(|S_n| \geqslant a\right) \leqslant \frac{\mathbb{E}S_n^2}{a^2}\,.$$

Proof. Let $A = \{\max_{1\leqslant k\leqslant n} |S_k| \geqslant a\}$ and

$$A_k = \{|S_k| \geqslant a \text{ but } |S_j| < a \text{ for } 1 \leqslant j < k\}\ ,\ k = 1, \cdots, n.$$

In words, we break things down according to the time that $|S_k|$ first exceeds

a, since $\{A_k\}$ are disjoint,

$$\begin{aligned}\mathbb{E}S_n^2 &= \sum_{k=1}^{n} \mathbb{E}S_n^2 I_{A_k} + \mathbb{E}S_n^2 I_{A^c} \\ &= \sum_{k=1}^{n} \mathbb{E}(S_n - S_k + S_k)^2 I_{A_k} + \mathbb{E}S_n^2 I_{A^c} \\ &= \sum_{k=1}^{n} \mathbb{E}S_k^2 I_{A_k} + 2\sum_{k=1}^{n} \mathbb{E}(S_n - S_k) I_{A_k} + \sum_{k=1}^{n} \mathbb{E}(S_n - S_k)^2 I_{A_k} + \mathbb{E}S_n^2 I_{A^c}\,.\end{aligned}$$

Note that $S_n - S_k$ is indeprndent of I_{A_k}, thus $\mathbb{E}(S_n - S_k)I_{A_k} = \mathbb{E}(S_n - S_k)\mathbb{E}I_{A_k} = 0$. And on A_k, we have $|S_k| \leqslant a$, thus $\mathbb{E}S_k^2 I_{A_k} \geqslant a^2\mathbb{P}(A_k)$. Therefore,

$$\begin{aligned}\mathbb{E}S_n^2 &= \sum_{k=1}^{n} \mathbb{E}S_k^2 I_{A_k} + \sum_{k=1}^{n} \mathbb{E}(S_n - S_k)^2 I_{A_k} + \mathbb{E}S_n^2 I_{A^c} \qquad (2.12)\\ &\geqslant \sum_{k=1}^{n} \mathbb{E}S_k^2 I_{A_k} \geqslant a^2 \sum_{k=1}^{n} \mathbb{P}(A_k) = a^2\mathbb{P}(A)\,. \qquad \square\end{aligned}$$

REMARK. In fact, The condition $\mathbb{E}X_i = 0$ implies that $(S_n)_{n\geqslant 0}$ is a *martingle*, and $(S_n^2)_{n\geqslant 0}$ is a *submartingle* : $E[S_n^2 \mid \mathcal{F}_{n-1}] = S_{n-1}^2 + \mathbb{E}X_n^2 \geqslant S_{n-1}^2$ for all n. This inequality has a generalization named *Dobb's ineqality*, which we will meet when discussing martingle.

We restate the proof above by *stopping times*. Let

$$\tau = \inf\left\{k \geqslant 0 : S_k^2 \geqslant a^2\right\}\,.$$

It' easy to check that τ is a stopping time with respect to $(\mathcal{F}_n)_{n\geqslant 0}$, where $\mathcal{F}_n = \sigma(X_1, \cdots, X_n)$. Then

$$\left\{\max_{1\leqslant k\leqslant n} |S_k| \geqslant a\right\} = \{\tau \leqslant n\}$$

Note that

$$\begin{aligned}
\mathbb{E}S_n^2 &= \sum_{k=1}^{n} \mathbb{E}S_n^2\, 1_{\{\tau=k\}} + \mathbb{E}S_n^2\, 1_{\{\tau>n\}} \\
&= \sum_{k=1}^{n} \mathbb{E}\left(\mathbf{E}\left[S_n^2\, 1_{\{\tau=k\}} \mid \mathcal{F}_k\right]\right) + \mathbb{E}S_n^2\, 1_{\{\tau>n\}} \\
&= \sum_{k=1}^{n} \mathbb{E}\left(1_{\{\tau=k\}}\, \mathbf{E}\left[S_n^2 \mid \mathcal{F}_k\right]\right) + \mathbb{E}S_n^2\, 1_{\{\tau>n\}} \qquad (2.13) \\
&= \sum_{k=1}^{n} \mathbb{E}S_k^2\, 1_{\{\tau=k\}} + \mathbb{E}S_n^2\, 1_{\{\tau>n\}} + \sum_{k=1}^{n} \mathbb{E}\left(S_n - S_k\right)^2 \mathbb{P}(\tau = k) \\
&= \mathbb{E}S_{\tau\wedge n}^2 + \sum_{k=1}^{n} \mathbb{E}\left(S_n - S_k\right)^2 \mathbb{P}(\tau = k)
\end{aligned}$$

On the other hand, we have

$$\mathbb{E}S_{\tau\wedge n}^2 = \mathbb{E}S_\tau^2\, 1_{\{\tau\leqslant n\}} + \mathbb{E}S_n^2\, 1_{\{\tau>n\}} \geqslant a^2\, \mathbb{P}(\tau \leqslant n)\,.$$

Thus

$$\mathbb{P}(\tau \leqslant n) \leqslant \frac{\mathbb{E}S_{\tau\wedge n}^2}{a^2} \leqslant \frac{\mathbb{E}S_n^2}{a^2}\,.$$

There is another way to show that $\mathbb{E}S_{\tau\wedge n}^2 \leqslant \mathbb{E}S_n^2$. Recall the proof of *Wald's indentity*,

$$\begin{aligned}
\mathbb{E}S_{\tau\wedge n}^2 &= \mathbb{E}\left(\sum_{k=1}^{n} X_k 1_{\{k\leqslant\tau\}}\right)^2 \\
&= \sum_{k=1}^{n} \mathbb{E}\left(X_k^2 1_{\{k\leqslant\tau\}}\right) + \sum_{1\leqslant i<j\leqslant n} \mathbb{E}\left(X_i X_j 1_{\{j\leqslant\tau\}}\right) \\
&= \sum_{k=1}^{n} \mathbb{E}X_k^2\, \mathbb{P}\left(\tau \geqslant k\right) + \sum_{1\leqslant i<j\leqslant n} \mathbb{E}X_j\, \mathbb{E}\left(X_i 1_{\{j\leqslant\tau\}}\right) \\
&= \sum_{k=1}^{n} \mathbb{E}X_k^2\, \mathbb{P}\left(\tau \geqslant k\right) \leqslant \sum_{k=1}^{n} \mathbb{E}X_k^2 = \mathbb{E}S_n^2\,.
\end{aligned}$$

Theorem 2.21 (Kolmogorov and Khinchin). *Let $\{X_n\}$ are independent with $\mathbb{E}X_i = 0$. If*

$$\sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty\,,$$

then series $\sum_{n=1}^{\infty} X_n$ converges a.s.

Proof. Take any $\epsilon > 0$, using Kolmogorov maximal ineaquality we have

$$\mathbb{P}(\sup_{k \geqslant n} |S_k - S_n| > \epsilon) \leqslant \frac{1}{\epsilon^2} \sum_{k=n}^{\infty} \mathrm{Var}(X_k)$$

By Lemma 2.19 the theorem holds. □

It's eary to get the following theorem.

Corollary 2.22 (Kolmogorov's two-series theorem). *Let $\{X_n\}$ are independent, If*

$$\sum_{n=1}^{\infty} \mathbb{E}(X_n) < \infty\,,\ \sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty\,,$$

then series $\sum_{n=1}^{\infty} X_n$ converges a.s.

Kolmogorov's three-series theorem Theorem 2.21 is sufficient for all of our applications, but our treatment would not be complete if we did not mention the last word on convergence of random series, which is a necessary and sufficient condition for convergence.

Lemma 2.23 (Kolmogorov's maximal inequality II). *Let $X_1, \cdots, X_n$ are independent with $\mathbb{E}X_i = 0$ and $\mathrm{Var}\,(X_i) < \infty$. If $\{X_n\}$ are uniformly bounded, i.e., there is a constant C such that $X_n \leqslant C$ a.s. for all n. Then for any $a > 0$,*

$$\mathbb{P}\left(\max_{1 \leqslant k \leqslant n} |S_k| \leqslant a\right) \leqslant \frac{(C+a)^2}{\mathbb{E}S_n^2}\,. \tag{2.14}$$

Proof. Let $A = \{\max_{1\leqslant k\leqslant n} |S_k| \geqslant a\}$ and

$$A_k = \{|S_k| \geqslant a \text{ but } |S_j| < a \text{ for } 1 \leqslant j < k\}\ , k = 1, \cdots, n.$$

In (2.12) we have proved that

$$\mathbb{E}S_n^2 = \sum_{k=1}^{n} \mathbb{E}S_k^2 I_{A_k} + \sum_{k=1}^{n} \mathbb{E}(S_n - S_k)^2 I_{A_k} + \mathbb{E}S_n^2 I_{A^c}\,.$$

Note that now, on A_k we have $|S_k| \leqslant (C+a)^2$, on A^c we have $S_n^2 \leqslant a$. Also, using the independence we have $\mathbb{E}(S_n - S_k)^2 I_{A_k} = \mathbb{E}(S_n - S_k)^2 \mathbb{P}(A_k) \leqslant \mathbb{E}(S_n^2)\mathbb{P}(A_k)$. Thus

$$\begin{aligned}
\mathbb{E}S_n^2 &\leqslant \sum_{k=1}^{n} (C+a)^2 \mathbb{P}(A_k) + \sum_{k=1}^{n} ES_n^2 \mathbb{P}(A_k) + a^2 \mathbb{P}(A^c) \\
&= \sum_{k=1}^{n} (C+a)^2 \mathbb{P}(A) + \mathbb{E}S_n^2 \mathbb{P}(A) + a^2 \mathbb{P}(A^c) \\
&\leqslant \mathbb{E}S_n^2 \mathbb{P}(A) + (C+a)^2
\end{aligned}$$

So $\mathbb{E}S_n^2 \mathbb{P}(A) \leqslant (C+a)^2$. □

Remark. Just like the Reamrk of Lemma 2.20, We restate the proof above by stopping times. In (2.13) we have proved that

$$\mathbb{E}S_n^2 = \mathbb{E}S_{\tau\wedge n}^2 + \sum_{k=1}^{n} \mathbb{E}\left(S_n - S_k\right)^2 \mathbb{P}(\tau = k)$$

Since $|X_n| \leqslant C$ for all n, and $|S_{\tau\wedge n-1}| < a$ we have $S_{\tau\wedge n}^2 \leqslant (C+a)^2$. On the other hand $\mathbb{E}\left(S_n - S_k\right)^2 \leqslant \mathbb{E}S_n^2$, thus

$$\mathbb{E}S_n^2 \leqslant (C+a)^2 + \mathbb{E}S_n^2 \sum_{k=1}^{n} \mathbb{P}(\tau = k) = (C+a)^2 + \mathbb{E}S_n^2\, \mathbb{P}(\tau \leqslant n)\,.$$

So

$$\mathbb{E}S_n^2\, \mathbb{P}(\tau > n) \leqslant (C+a)^2\,.$$

Theorem 2.24 (Kolmogorov's three-series theorem). *Let $\{X_n\}$ be independent r.v.'s. Take $A > 0$ and let $Y_n = X_n 1_{\{|X_n| \leqslant A\}}$. In order that $\sum_{n=1}^{\infty} X_n$ converges a.s.,it is necessary and sufficient that*

$$(i) \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty, \quad (ii) \sum_{n=1}^{\infty} \mathbb{E}Y_n < \infty, \quad (iii) \sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty \tag{2.15}$$

Proof. Sufficiency. By $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty$ and B-C lemma we have

$$\mathbb{P}(X_n \neq Y_n \text{ i.o.}) = 0$$

Thus it suffices to show $\sum_{n=1}^{\infty} Y_n$ converges a.s. . Note that $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty$ and Theorem 2.21 imply that

$$\sum_{n=1}^{\infty} Y_n - \mathbb{E}Y_n \text{ converges a.s. .}$$

Using $\sum_{n=1}^{\infty} \mathbb{E}Y_n < \infty$, now gives that $\sum_{n=1}^{\infty} Y_n$ converges a.s.

Necessity: $\sum_{n=1}^{\infty} X_n$ converges a.s. and B-C lemma imply $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty$ and $\mathbb{P}(X_n \neq Y_n \text{ i.o.}) = 0$, Thus $\sum_{n=1}^{\infty} Y_n$ converges a.s.

We use the following *symmetrization method*. Take $(Y_n')_{n \geqslant 1}$ is independent of $(Y_n)_{n \geqslant 1}$ such that thw two have same distribution. Since $\sum_{n=1}^{\infty} Y_n$ converges a.s.,$\sum_{n=1}^{\infty} Y_n'$ converges a.s. , $\sum_{n=1}^{\infty} Y_n - Y_n'$ converges a.s. Denote $T_n = \sum_{k=1}^{n} Y_k - Y_k'$, since $\{T_n\}$ converges a.s. when $n \to \infty$, we have

$$\mathbb{P}\left(\sup_{n \geqslant 1} |T_n| < \infty\right) = 1\,.$$

So there is $M > 0$ such that $\mathbb{P}(\sup_n |T_n| \leqslant M) > \frac{1}{2}$. Note that $|Y_n - Y_n'| \leqslant 2A$ for all n, by Lemma 2.23 we have

$$\mathbb{P}\left(\max_{1 \leqslant k \leqslant n} |T_k| \leqslant M\right) \leqslant \frac{(2A+M)^2}{\mathrm{Var}(T_n)}$$

for any given n. Since $\mathrm{Var}(T_n) = 2\sum_{k=1}^{n} \mathrm{Var}(Y_k)$, letting $n \to \infty$ we have

$$\frac{(2A+M)^2}{2\sum_{n=1}^{\infty} \mathrm{Var}(Y_n)} \geqslant \frac{1}{2}$$

Thus

$$\sum_{n=1}^{\infty} \mathrm{Var}\,(Y_n) < \infty\,.$$

Using Theorem 2.21 we have that $\sum_{n=1}^{\infty}(Y_n - \mathbb{E}Y_n)$ converges a.s., so

$$\sum_{n=1}^{\infty} \mathbb{E}Y_n < \infty\,. \qquad \square$$

EXERCISE

EXERCISE 14. Let $X_n \geqslant 0$ be independent for $n \geqslant 1$. The following are equivalent: (i) $\sum_{n=1}^{\infty} X_n < \infty$ a.s. (ii) $\sum_{n=1}^{\infty}\left[\mathbb{P}\,(X_n > 1) + \mathbb{E}\left(X_n 1_{(X_n \leqslant 1)}\right)\right] < \infty$ (iii) $\sum_{n=1}^{\infty} \mathbb{E}\,(X_n/\,(1+X_n)) < \infty$.
Hint: Use Kolmogorov three-series theorem.

EXERCISE 15. Let $X_1, X_2, \ldots$ be independent and let $S_{m,n} = X_{m+1} + \ldots + X_n$ for $n > m$. Then

(i) Show that

$$\mathbb{P}\left(\max_{m<j\leqslant n} |S_{m,j}| > 2a\right) \min_{m<k\leqslant n} \mathbb{P}\,(|S_{k,n}| \leqslant a) \leqslant \mathbb{P}\,(|S_{m,n}| > a) \quad (2.16)$$

(ii) Use (i) to prove a theorem of P. Lévy : If $\{S_n\}$ converges in probability then it also converges almost surely.

(iii) Use (i) to conclude that if $\frac{S_n}{n} \to 0$ in probability then

$$\frac{\max\limits_{1\leq m\leq n} S_m}{n} \to 0 \quad \text{in probability.}$$

Hint: For fixed m, let $\tau = \inf\{j \geqslant m : |S_{m,j}| > 2a\}$, then using $\{|S_{m,n}| > a\} \supset \cup_{j=1}^{n}\{|S_{m,n}| > a\,, \tau = j\}$ can prove (i). To show (ii), note that S_n is Cauchy sequence in probability and use Lemma 2.19.

EXERCISE 16. Lemma 2.20 has the following "one-sided" analogue. Under the same hypotheses, we have

$$\mathbb{P}\left(\max_{1\leq j\leq n} S_j \geqslant a\right) \leqslant \inf_{c\geqslant 0} \frac{\mathrm{Var}\,(S_n)+c^2}{(a+c)^2} \leqslant \frac{\mathrm{Var}\,(S_n)}{a^2+\mathrm{Var}\,(S_n)}\,.$$

This is due to A. W. Marshall.

Hint: note that for any c, $S_n + c$ is a submartingle.

2.5 Strong Law of Large Numbers (II)

The link between convergence of series and the strong law of large numbers is provided by

Lemma 2.25 (Kronecker's lemma). *$\{x_n\}$ and $\{b_n\}$ are sequence of real numbers. $b_n > 0$, and $b_n \uparrow \infty$. If $\sum_{n=1}^{\infty} \frac{x_n}{b_n}$ converges, let $S_n = x_1 + \cdots + x_n$, then*

$$\frac{S_n}{b_n} \to 0\,.$$

Proof. Note that

$$\frac{S_n}{b_n} = \frac{1}{b_n} \sum_{m=1}^{n} b_m \, \frac{x_m}{b_m}\,.$$

Let $T_n = \sum_{m=1}^{n} x_m/b_m$, and $T_0 = 0$, using *summation by parts*,

$$\begin{aligned} \frac{S_n}{b_n} &= \frac{1}{b_n} \sum_{m=1}^{n} b_m \frac{x_m}{b_m} = \frac{1}{b_n} \sum_{m=1}^{n} b_m (T_m - T_{m-1}) \\ &= T_n + \frac{1}{b_n} \sum_{m=1}^{n-1} T_m (b_m - b_{m+1}) \end{aligned}$$

By *Stolz–Cesàro theorem*, we konw

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{m=1}^{n-1} T_m (b_m - b_{m+1}) = \lim_{n\to\infty} \frac{T_{n-1}(b_{n-1} - b_n)}{b_n - b_{n-1}} = -\lim_{n\to\infty} T_n$$

Thus

$$\lim_{n\to\infty} \frac{S_n}{b_n} = \lim_{n\to\infty} T_n - \lim_{n\to\infty} T_n = 0.$$

□

Theorem 2.26 (SLLN for i.i.d. sequence). *Let $\{X_n\}$ be i.i.d. r.v.'s and $S_n = X_1 + \cdots + X_n$. Then there exists some $\mu \in \mathbb{R}$ such that*

$$\frac{S_n}{n} \to \mu \quad \text{a.s.}$$

if and only if $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$.

Proof. We only need to show sufficiency. Let $Y_n = X_n 1_{\{|X_n| \leqslant n\}}$ and $T_n = Y_1 + \cdots + Y_n$. It suffices to show

$$\frac{T_n}{n} \to \mu \quad \text{a.s.}$$

Let $Z_n = Y_n - \mathbb{E}Y_n$, then $\mathbb{E}Z_n = 0$. Since $\operatorname{Var}(Z_n) = \operatorname{Var}(Y_n) \leqslant \mathbb{E}Y_n^2$ and by Lemma 2.11

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var}(Z_n)}{n^2} \leqslant \sum_{n=1}^{\infty} \frac{\mathbb{E}Y_n^2}{n^2} < \infty$$

Applying Theorem 2.21, we conclude that $\sum_{n=1}^{\infty} \frac{Z_n}{n}$ converges a.s, so Lemma 2.25 implies

$$\frac{1}{n}\sum_{k=1}^{n} Z_k = \frac{T_n - \mathbb{E}T_n}{n} \to 0 \text{ a.s. .}$$

The dominated convergence theorem implies $\mathbb{E}Y_n \to \mu$ as $n \to \infty$. From this, it follows easily that $\mathbb{E}\frac{T_n}{n} \to \mu$ and hence the drsired result follows. □

From the proof of strong law of large number, we get the following theorem the proof is left as an exerise.

Theorem 2.27 (SLLN for independent r.v.'s). *Suppose $\{X_n\}$ is independent r.v.'s with $\mathbb{E}X_n^2 < \infty$ for all n. Let $S_n = X_1 + \cdots + X_n$. If $b_n \uparrow \infty$, and $\sum_{n=1}^{\infty} \frac{Var(X_n)}{b_n^2}$ converges, then as*

$$\frac{S_n - \mathbb{E}S_n}{b_n} \to 0, \quad a.s.$$

REMARK. When $\{X_n\}$ is independent and we don't konw if $\mathbb{E}X_n^2 < \infty$ or not, in order to use Theorem 2.27, we can truncate $\{X_n\}$. Besides, Theorem 2.27 can be used to estimate the rates convergence in SLLN for i.i.d sequence.

2.5.1 Rates of Convergence

As mentioned earlier, one of the advantages of the random series proof is that it provides estimates on the rate of convergence of $\frac{S_n}{n} \to \mu$.

Assume $\mathrm{Var}(X_1) = \sigma^2 < \infty$, by CLT we konw for all $a > 0$,

$$\mathbb{P}(-a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < a) = \mathbb{P}(-\frac{\sigma}{\sqrt{n}}a < \frac{S_n}{n} - \mu < \frac{\sigma}{\sqrt{n}}a) \to \Phi(a) - \Phi(-a)\,.$$

This means that $\frac{S_n}{n} = \mu + O(\frac{1}{\sqrt{n}})$ is NOT ture.

Theorem 2.28. *Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}X_1^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. For any $\epsilon > 0$ then*

$$\frac{S_n - n\mu}{n^{1/2}(\log n)^{1/2+\epsilon}} \to 0\,, \quad \text{a.s.} \tag{2.17}$$

or,

$$\frac{S_n}{n} = \mu + o(\frac{(\log n)^{1/2+\epsilon}}{n^{1/2}})\,.$$

REMARK. The *law of the iterated logarithm* claims that

$$\limsup_{n\to\infty} \frac{S_n - n\mu}{n^{1/2}(\log n)^{1/2}} = \sigma\sqrt{2} \quad \text{a.s.} \tag{2.18}$$

where σ is standard deviation of X_1, so the last result is not far from the best possible.

Proof. By subtracting μ from each random variable, we can and will suppose without loss of generality that $\mu = 0$. Let $b_n = n^{1/2}(\log n)^{1/2+\epsilon}$ for $n \geqslant 2$ and $b_1 > 0$. Note that

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}\,(X_n)}{b_n^2} = \sigma^2 \left(\frac{1}{b_1^2} + \sum_{n=2}^{\infty} \frac{1}{n(\log n)^{1+2\epsilon}} \right) < \infty\,.$$

Then we get the required result from Theorem 2.27. □

The next result due to Marcinkiewicz and Zygmund treats the situation in which $\mathbb{E}X_1^2 = \infty$, but $\mathbb{E}\,|X_1|^p < \infty$ for some $1 < p < 2$.

Theorem 2.29. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}\,|X_1|^p < \infty$, where $1 < p < 2$. If $S_n = X_1 + \cdots + X_n$ then*

$$\frac{S_n - n\mu}{n^{1/p}} \to 0 \quad \text{a.s.} \tag{2.19}$$

or,

$$\frac{S_n}{n} = \mu + o(\frac{1}{n^{1-\frac{1}{p}}})\ .$$

Proof. To use Theorem 2.27, we begin by truncation. Take truncation level $\{b_n\}$, let $Y_n = X_n 1_{\{|X_n| \leqslant b_n\}}$. In order that $\mathbb{P}(Y_n \neq X_n \text{ i.o.}) = 0$, by B-C lemma we need

$$\sum_{n=1}^{\infty} \mathbb{P}(Y_n \neq X_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geqslant b_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1|^p \geqslant b_n^p) < \infty . \tag{2.20}$$

Thus we let $b_n^p = n$ for each n, since $\mathbb{E}|X_1|^p < \infty$, (2.20) holds.

Put $T_n = Y_1 + \cdots + Y_n$, it suffices to show

$$\frac{T_n - \mathbb{E}T_n}{n^{1/p}} \to 0 \text{ and } \frac{\mathbb{E}T_n - n\mu)}{n^{1/p}} \to 0 .$$

Note that

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{\mathrm{Var}(Y_n)}{n^{2/p}} \leqslant \sum_{n=1}^{\infty} \frac{EY_n^2}{n^{2/p}} &= \sum_{n=1}^{\infty} \int \frac{1}{n^{2/p}} 1_{\{|y| \leqslant n^{1/p}\}} y^2 dF(y) \\ &= \int \sum_{n=1}^{\infty} n^{-2/p} 1_{\{|y|^p \leqslant n\}} y^2 dF(y) \end{aligned}$$

To bound the integral, we note that for $n \geqslant 2$ comparing the sum with the integral of $x^{-2/p}$

$$\sum_{n \geqslant |y|^p}^{\infty} n^{-2/p} \leqslant C \int_{|y|^p}^{\infty} x^{-2/p} dx \leqslant C' |y|^{p-2} .$$

where C and C' are constant only related to p. Thus

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(Y_n)}{n^{2/p}} \leqslant C' \int |y|^p dF(y) < \infty .$$

By Theorem 2.27, we have

$$\frac{T_n - ET_n}{n^{1/p}} \to 0, \text{ a.s. } .$$

Next we will estimate

$$\frac{\mathbb{E}T_n - n\mu}{n^{1/p}} .$$

$\mathbb{E}Y_n - \mu = -\mathbb{E}\left(X_1 1_{\{|X_1|^p > n\}}\right)$, so

$$\frac{|ET_n - n\mu|}{n^{1/p}} \leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{n} \mathbb{E}\left(|X_1| 1_{\{|X_1|^p > k\}}\right) .$$

Note that $\mathbb{E}|X_1|^p < \infty$,

$$\begin{aligned}\frac{|ET_n - n\mu|}{n^{1/p}} &\leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{n} k^{-1+1/p}\, \mathbb{E}\left(|X_1| \cdot |X_1|^{p-1} 1_{\{|X_1|^p > k\}}\right) . \\ &\leqslant \frac{1}{n^{1/p}} \sum_{k=1}^{n} k^{-1+1/p}\, \mathbb{E}\left(|X_1|^p 1_{\{|X_1|^p > k\}}\right) .\end{aligned}$$

By *stolz theorem*, or Lebesgue domainted convergence theorem,

$$\frac{1}{n^{1/p}} \sum_{k=1}^{n} k^{-1+1/p}\, \mathbb{E}\left(|X_1|^p 1_{\{|X_1|^p > k\}}\right) \to 0 .$$

So

$$\frac{\mathbb{E}T_n - n\mu}{n^{1/p}} \to 0 . \qquad \square$$

REMARK. The converse of Theorem 2.29 is much easier. Let $p > 0$. If $\frac{S_n}{n^{1/p}} \to 0$ a.s. then $\mathbb{E}\,|X_1|^p < \infty$.

2.5.2 Infinite Mean

Theorem 2.30. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}\,|X_1| = \infty$. Let $S_n = X_1 + \cdots + X_n$. Suppose $\{a_n\}$ is a sequence of positive numbers with $\frac{a_n}{n}$ increasing. Then with probability one*

$$\limsup_{n\to\infty} \frac{|S_n|}{a_n} = \begin{cases} 0 \ , \ \text{if} \ \displaystyle\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > a_n\right) < \infty . \\ \infty \ , \ \text{if} \ \displaystyle\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > a_n\right) = \infty . \end{cases} \tag{2.21}$$

Proof. Case 1 : if $\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > a_n\right) < \infty$, we show

$$\frac{|S_n|}{a_n} \to 0 \quad \text{a.s.}$$

To use Theorem 2.27, we begin by truncation. Take truncation level $\{b_n\}$ and let $Y_n = X_n 1_{\{|X_n| \leqslant b_n\}}$ for each n, to make

$$\mathbb{P}\left(Y_n \neq X_n \text{ i.o.}\right) = 0 ,$$

by B-C lemma, that is

$$\sum_{n=1}^{\infty} \mathbb{P}\left(Y_n \neq X_n\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| > b_n\right) < \infty$$

Clearly, it suffices to let $b_n = a_n$. Put $T_n = Y_1 + \cdots + Y_n$, we only need to show

$$\frac{|T_n|}{a_n} \to 0 \quad \text{a.s.}$$

To do this, we compute

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{\operatorname{Var}\left(Y_n\right)}{a_n^2} \leqslant \sum_{n=1}^{\infty} \frac{\mathbb{E} Y_n^2}{a_n^2} &= \sum_{n=1}^{\infty} \frac{1}{a_n^2} \sum_{m=1}^{n} \mathbb{E}\, X_1^2\, 1_{\{a_{m-1}<|X_1|\leqslant a_m\}} \\ &= \sum_{m=1}^{\infty} \mathbb{E}\, X_1^2\, 1_{\{a_{m-1}<|X_1|\leqslant a_m\}} \sum_{n=m}^{\infty} \frac{1}{a_n^2}\end{aligned}$$

Since $\frac{a_n}{n} \geqslant \frac{a_m}{m}$ when $n \geqslant m$, we have

$$\sum_{n=m}^{\infty} \frac{1}{a_n^2} \leqslant \frac{m^2}{a_m^2} \sum_{n=m}^{\infty} \frac{1}{n^2} \leqslant 2\, \frac{m}{a_m^2}\,.$$

Thus

$$\begin{aligned}&\sum_{n=1}^{\infty} \frac{\mathbb{E} Y_n^2}{a_n^2} \leqslant 2 \sum_{m=1}^{\infty} \frac{m}{a_m^2}\, \mathbb{E}\, X_1^2\, 1_{\{a_{m-1}<|X_1|\leqslant a_m\}} \\ &\leqslant 2 \sum_{m=1}^{\infty} m\mathbb{P}(a_{m-1} < |X_1| \leqslant a_m) = 2 \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > a_n) < \infty\,.\end{aligned}$$

Thus by Theorem 2.27,

$$\frac{T_n - \mathbb{E} T_n}{a_n} \to 0 \quad \text{a.s.}$$

Now we only need to show

$$\frac{\mathbb{E} T_n}{a_n} \to 0\,.$$

To begin, note that since $\mathbb{E}\,|X_1| = \infty$, $\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > a_n\right) < \infty$, and $\frac{a_n}{n}$ increasing, we must have $\frac{a_n}{n} \uparrow \infty$.

To estimate $\frac{\mathbb{E}T_n}{a_n}$ now, we observe that

$$\begin{aligned}\left|\frac{\mathbb{E}T_n}{a_n}\right| &\leqslant \frac{1}{a_n}\sum_{m=1}^{n}\mathbb{E}\,|X_1|\,1_{\{|X_1|\leqslant a_m\}} = \frac{1}{a_n}\sum_{m=1}^{n}\sum_{k=1}^{m}\mathbb{E}\,|X_1|\,1_{\{a_{k-1}<|X_1|\leqslant a_k\}}\\ &= \frac{1}{a_n}\sum_{k=1}^{n}(n-k+1)\mathbb{E}\,|X_1|\,1_{\{a_{k-1}<|X_1|\leqslant a_k\}}\\ &\leqslant \frac{n}{a_n}\sum_{k=1}^{n}\mathbb{E}\,|X_1|\,1_{\{a_{k-1}<|X_1|\leqslant a_k\}} \leqslant \frac{n}{a_n}\sum_{k=1}^{n}a_k\,\mathbb{P}\left(a_{k-1}<|X_1|\leqslant a_k\right).\end{aligned}$$

Using $\dfrac{na_k}{a_n}\leqslant k$, then as $n\to\infty$,

$$\sum_{k=1}^{n}k\,\mathbb{P}\left(a_{k-1}<|X_1|\leqslant a_k\right)\to\sum_{k=1}^{\infty}k\,\mathbb{P}\left(a_{k-1}<|X_1|\leqslant a_k\right)=\sum_{k=1}^{\infty}\mathbb{P}\left(|X_1|>a_{k-1}\right). \tag{2.22}$$

doesn't tend wo zero. However, we can divide the sum into two parts : for any given $N\in\mathbb{N}_+$, when $n>N$, we have

$$\left|\frac{\mathbb{E}T_n}{a_n}\right|\leqslant\frac{n}{a_n}Na_N+\frac{n}{a_n}\sum_{k=N}^{n}a_k\mathbb{P}(a_{k-1}<|X_1|\leqslant a_k),$$

and

$$\begin{aligned}\frac{n}{a_n}\sum_{k=N}^{n}a_k\mathbb{P}(a_{k-1}<|X_1|\leqslant a_k)&\leqslant\sum_{k=N}^{n}k\,\mathbb{P}(a_{k-1}<|X_1|\leqslant a_k)\\ &\leqslant\sum_{k=N}^{\infty}k\,\mathbb{P}\left(a_{k-1}<|X_1|\leqslant a_k\right).\end{aligned}$$

By (2.22), as $N\to 0$,

$$\sum_{k=N}^{\infty}k\,\mathbb{P}\left(a_{k-1}<|X_1|\leqslant a_k\right)\to 0.$$

From

$$\left|\frac{\mathbb{E}T_n}{a_n}\right|\leqslant\frac{n}{a_n}Na_N+\sum_{k=N}^{\infty}k\,\mathbb{P}\left(a_{k-1}<|X_1|\leqslant a_k\right),$$

let $n \to 0$ first, since N is fixed,

$$\limsup_{n\to\infty} \left| \frac{\mathbb{E}T_n}{a_n} \right| \leqslant \sum_{k=N}^{\infty} k\,\mathbb{P}\left(a_{k-1} < |X_1| \leqslant a_k\right) .$$

Then let $N \to \infty$ we have the desired result.

Case 2: Suppose $\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > a_n\right) = \infty$. Recall the proof in Theorem 2.14, we guess if there holds

$$\limsup_{n\to\infty} \frac{|X_n|}{a_n} = \infty\,.$$

If ture, since

$$\max\left\{|S_{n-1}|\,, |S_n|\right\} \geqslant \frac{|X_n|}{2},$$

the desired result follows.

It suffices to show that for any given $k \in \mathbb{N}_+$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > ka_n\right) = \infty\,.$$

Note that $\frac{a_n}{n}$ increasing, we have $a_{kn} \geqslant ka_n$. Using this and $\{a_n\}$ increasing, then

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > ka_n\right) \geqslant \sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| > a_{kn}\right) \geqslant \frac{1}{k} \sum_{m=k}^{\infty} \mathbb{P}\left(|X_1| > a_m\right) = \infty\,. \quad \square$$

2.6 Large Deviations

Motivation To give an example about standard normal distribution, we analysis the probability $\mathbb{P}(N(0,1) > x)$ first

Lemma 2.31. *For $x > 0$.*

$$\left(x^{-1} - x^{-3}\right) \exp\left(-x^2/2\right) \leqslant \int_x^\infty \exp\left(-y^2/2\right) dy \leqslant x^{-1} \exp\left(-x^2/2\right) \tag{2.23}$$

Proof. Changing variables $y = x + z$ and using $\exp\left(-z^2/2\right) \leq 1$ gives

$$\int_x^\infty \exp\left(-y^2/2\right) dy \leq \exp\left(-x^2/2\right) \int_0^\infty \exp(-xz) dz = x^{-1} \exp\left(-x^2/2\right) .$$

For the other direction, we observe,

$$\int_x^\infty \left(1 - 3y^{-4}\right) \exp\left(-y^2/2\right) dy = \left(x^{-1} - x^{-3}\right) \exp\left(-x^2/2\right) . \qquad \square$$

Now we give the example:

Example 2.22. Let $X_1, X_2, \ldots$ be i.i.d. with distribution $N(0,1)$. Then, $S_n = X_1 + \cdots + X_n \sim N(0, n^2)$, for every $a > 0$

$$\mathbb{P}\left(S_n \geqslant na\right) = \mathbb{P}(X_1 \geqslant a\sqrt{n}) = (1 + \epsilon_n) \frac{1}{a\sqrt{2\pi n}} e^{-\frac{a^2}{2}n}$$

where $\epsilon_n \to 0$ by Lemma 2.31. Taking logarithms, we get

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(S_n \geqslant na\right) = -\frac{a^2}{2} , \qquad \text{for every } a > 0.$$

2.6.1 Exponenitial convergence rate

Let $X_1, X_2, \cdots$ be i.i.d. r.v.'s with finite expected value μ, and $S_n = X_1 + \cdots + X_n$. In this section, for some $a > \mu$, we will investigate the rate at which

$$\mathbb{P}\left(S_n > na\right) \to 0$$

We will ultimately conclude that if the moment-generating function $\phi(\theta) = \mathbb{E}e^{\theta X_1} < \infty$ for some $\theta > 0$, then $\mathbb{P}(S_n \geqslant na) \to 0$ exponentially rapidly and we will identify the limit

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant na) \ .$$

Our first step is to prove that the limit exists.

Lemma 2.32. *For any $a \in \mathbb{R}$, the sequence $\left\{\frac{1}{n}\log \mathbb{P}(S_n \geqslant na)\right\}$ converges.*

Proof. This is based on an observation that will be useful several times below.

$$\begin{aligned}\mathbb{P}(S_{n+m} \geqslant na + ma) &\geqslant \mathbb{P}(S_m \geqslant ma, S_{n+m} - S_m \geqslant na) \\ &= \mathbb{P}(S_m \geqslant ma)\, \mathbb{P}(S_n \geqslant na) \ ,\end{aligned}$$

where using that S_m and $S_{n+m} - S_m$ are independent and $S_{n+m} - S_m$ are identicaly distributed with S_n. Let $\gamma_n \coloneqq \frac{1}{n}\log \mathbb{P}(S_n \geqslant na)$, then we have $\gamma_{m+n} \geqslant \gamma_m + \gamma_n$ for any $n, m \in \mathbb{N}_+$. From mathematical analysis, we konw $\gamma_n/n \to \sup_m \gamma_m/m$ as $n \to \infty$. □

We denote $\gamma(a)$ as the limit of sequence above, i.e.,

$$\gamma(a) \coloneqq \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \geqslant na) = \sup_{n\geqslant 1} \frac{1}{n} \log \mathbb{P}(S_n \geqslant na) \ . \tag{2.24}$$

Obviously, $\gamma(a) \leqslant 0$. It follows from the formula for the limit that

$$\mathbb{P}(S_n \geqslant na) \leqslant e^{n\gamma(a)} \tag{2.25}$$

The last conclusions is valid for any distribution but it is useless if $\gamma(a) = 0$.

Theorem 2.33. *If there exists some $a > \mu$ such that $\gamma(a) < 0$, then there exists $\theta > 0$ such that $\mathbb{E}e^{\theta X_1} < \infty$.*

Proof. Suppose $\gamma(a) < -\beta < 0$, then for all $n \in \mathbb{N}$,

$$\mathbb{P}\left(S_n \geqslant na\right) \leqslant e^{-n\beta}$$

Pick any $\epsilon > 0$, we have

$$\begin{aligned}\mathbb{P}(S_n \geqslant na) &\geqslant \mathbb{P}(S_{n-1} \geqslant (\mu-\epsilon)n,\, X_n \geqslant (a-\mu+\epsilon)n) \\ &= \mathbb{P}(S_{n-1} \geqslant (\mu-\epsilon)n)\mathbb{P}(X_1 \geqslant (a-\mu+\epsilon)n)\,.\end{aligned}$$

By law of large number, there is some $N = N_\epsilon$ such that for all $n \geqslant N$,

$$\mathbb{P}(S_{n-1} \geqslant (\mu-\epsilon)n) \geqslant 1/2\,.$$

Thus when $n \geqslant N$,

$$\mathbb{P}(X_1 \geqslant (a-\mu+\epsilon)n) \leqslant 2P(S_n \geqslant na) \leqslant 2e^{-\beta n}\,.$$

To complete the proof, we need the following lemma,

Lemma 2.34. *Let ξ be a random varible. If there exists some $\beta > 0$ such that*

$$\mathbb{P}(\xi \geqslant n) \ll e^{-\beta n} \text{ for sufficiently large } n$$

Then there exists some $\theta > 0$ such that $\mathbb{E}e^{\theta\xi} < \infty$.

Proof of lemma 2.34. Note that

$$\mathbb{E}e^{\theta\xi} = \int_0^\infty \mathbb{P}(e^{\theta\xi} > t)dt = \int_0^\infty \mathbb{P}(\xi > \theta^{-1}\log t)dt\,.$$

When t is sufficiently large,

$$\mathbb{P}(\xi > \theta^{-1}\log t) \leqslant \mathbb{P}\left(\xi > \left[\frac{1}{\theta}\log t\right]\right) \leqslant Ce^{-\beta\left[\frac{1}{\theta}\log t\right]} \leqslant Ct^{-\frac{\beta}{2\theta}}\,.$$

Pick $\theta \in (0, \beta/2)$, then we have $\mathbb{E}e^{\theta\xi} < \infty$. □

Let $\xi = X_1/(a-\mu+\epsilon)$. From Lemma 2.34, there exists some $\theta > 0$ such that $\mathbb{E}e^{\theta X_1} < \infty$. □

Therefore, to make $\gamma(a) < 0$, we will suppose:

$$\phi(\theta) = \mathbb{E}\exp(\theta X_1) < \infty \text{ for some } \theta_0 > 0 \quad (\mathrm{H}_1 : \text{Cramér's condition})$$

in the rest of this section, where ϕ is called **moment-generating function**. Then,

- Denote $I := \{\theta \geqslant 0 : \phi(\theta) < \infty\}$, then I is an interval. Indeed, let $\theta_+ = \sup\{\theta : \phi(\theta) < \infty\}$ and there holds $\phi(\theta) < \infty$ for all $\theta \in [0, \theta_+)$, since when $\phi(\theta_0) < \infty$, for any $0 \leqslant \theta < \theta_0$ we have

  $$\phi(\theta) = \mathbb{E}\exp(\theta X) \leqslant \mathbb{E}\exp(\theta_0 X) + 1 = \phi(\theta_0) + 1 < \infty .$$

  Thus $I = [0, \theta_+)$ or $I = [0, \theta_+]$. Both the two cases are possible, see Example 2.24 and Example 2.25.

- (H_1 : Cramér's condition) holds iff there exists some $\theta_0 > 0$ such that

  $$\mathbb{P}(X_1 > x) \ll e^{-\theta_0 x} \text{ as } x \to \infty .$$

Lemma 2.35. *For moment-generating function ϕ, we have*

(i) $\phi \in C(I)$.

(ii) $\phi \in C^\infty[0, \theta_+)$, $\phi^{(n)}(\theta) = \mathbb{E}(X_1^n e^{\theta X_1})$, *for all* $\theta \in [0, \theta_+)$.

(iii) $\phi'(0) = \mu$ *and* $\phi''(\theta) \geqslant 0$, *for all* $\theta \in [0, \theta_+)$.

Proof. We only prove (ii). Take any $\theta \in [0, \theta_+)$ and fix it, for any $h \neq 0$ such that $\theta + h \in I$,

$$\frac{\phi(\theta + h) - \phi(\theta)}{h} = \int \frac{\exp(hx) - 1}{h} \exp(\theta x)\, dF(x) ,$$

where F is the c.d.f. of X_1. Note that for any $h \neq 0$ and $x \in \mathbb{R}$, there exists some $\xi \in (0, 1)$ such that

$$\frac{\exp(hx) - 1}{h} = x \exp(\xi h x)$$

Since $\theta + h \in I$, $\xi \in (0,1)$,$\theta + \xi h \in I$. Then we can pick some $\theta_0 \in I$ such that $\theta + \xi h < \theta_0$ when $|h|$ sufficiently small, so

$$\frac{\exp(hx) - 1}{h} \exp(\theta x) = x \exp((\theta + \xi h)x) \ll \exp(\theta_0 x), \text{ as } x \to \infty .$$

Using Lebesgue domainted convergence theorem (Note that only when $\theta \in [0, \theta_+)$ can we use this theorem), we have

$$\begin{aligned} \lim_{\substack{h \to 0 \\ \theta + h \in I}} \frac{\phi(\theta + h) - \phi(\theta)}{h} &= \int \lim_{\substack{h \to 0 \\ \theta + h \in I}} \frac{\exp(hx) - 1}{h} \exp(\theta x)\, dF(x) \\ &= \int x \exp(\theta x)\, dF(x) . \end{aligned}$$

Thus we have

$$\phi'(\theta) = \int x \exp(\theta x)\, dF(x) , \quad \text{for all } \theta \in I .$$

By induction, we have

$$\phi^{(n)}(\theta) = \int x^n \exp(\theta x)\, dF(x) , \quad \text{for all } \theta \in I . \tag{2.26}$$

So we get the desired results. □

REMARK. If $\theta_+ < \infty$ and $\phi(\theta_+) < \infty$, it could happen $\phi'(\theta) \uparrow \infty$ as $\theta \uparrow \theta_+$, see Example 2.25. It's obvious that $\phi'(\theta_+)$ exists iff $\lim_{\theta \uparrow \theta_+} \phi'(\theta) < \infty$.

We use Chebyshev's inequality to give an upper bound for $\gamma(a)$. For any $\theta \in I$, by Chebyshev's inequality,

$$\mathbb{P}\left(S_n \geqslant na\right) \leqslant \frac{\mathbb{E}e^{\theta S_n}}{e^{\theta na}} = \frac{\phi(\theta)^n}{e^{\theta na}} ,$$

Define $\kappa(\theta) \coloneqq \log \phi(\theta)$ is the **logarithmic moment-generating function** of X_1, then

$$\frac{1}{n} \log \mathbb{P}\left(S_n \geqslant na\right) \leqslant -\left[a\theta - \kappa(\theta)\right] , , \text{ for all } n . \tag{2.27}$$

Theorem 2.36. *X_1 satisfying (H_1 : Cramér's condition), then*

$$\gamma(a) \leqslant -\sup_{\theta\in I}(a\theta - \kappa(\theta)) < 0\,.$$

Proof. From (2.27) we have

$$\gamma(a) \leqslant -\sup_{\theta\in I}(a\theta - \kappa(\theta)).$$

So we only need to prove that $\sup_{\theta\in I}(a\theta - \kappa(\theta)) > 0$. By Lemma 2.35, $\kappa \in C^\infty(I)$, and $\kappa'(0) = \phi'(0)/\phi(0) = \mu < a$. Thus for small θ we have $\kappa'(x) < a$ when $x \in [0, \theta]$, then

$$a\theta - \kappa(\theta) = \int_0^\theta a - \kappa'(x)dx > 0\,. \qquad \square$$

The rest of this section will identify the precise value of $\gamma(a)$. In fact, we will show that

$$\gamma(a) = -\sup_{\theta\in I}(a\theta - \kappa(\theta))\,.$$

2.6.2 Precise value of $\gamma(a)$

First we try to find the supremum of $a\theta-\kappa(\theta)$ on I. $(a\theta-\kappa(\theta))' = a-\kappa'(\theta)$, so if things are nice, the maximum occurs when $\kappa'(\theta) = a$. To get this, we should compute the second derivative of κ:

$$\kappa''(\theta) = \frac{d}{d\theta}\frac{\phi'(\theta)}{\phi(\theta)} = \frac{\phi''(\theta)\phi(\theta) - \phi'(\theta)^2}{\phi(\theta)^2} \geqslant 0\,.$$

Indeed, from (2.26), we have

$$\phi'(\theta) = \int x\exp(\theta x)\,dF(x)\,,\ \ \phi''(\theta) = \int x^2\exp(\theta x)\,dF(x)\,.$$

By C-B-S inequality,

$$\left(\int x\exp(\theta x)\,dF(x)\right)^2 \leqslant \int x^2\exp(\theta x)\,dF(x)\int \exp(\theta x)\,dF(x)\,,$$

and the equality holds if anf only if $X_1 = \mu$ a.s. If we assume :

X_1 is not constant random variable, (H_2 : nondegenerate condition)

Then

$$\kappa''(\theta) > 0\,, \quad \text{for all } \theta \in I\,.$$

so $\kappa'(\theta)$ is strictly increasing, and $a\theta - \kappa(\theta)$ is *concave*. Since $\kappa'(0) = \mu$, this shows that for each $a > \mu$ there is at most one $\theta_a > 0$ that solves $\kappa'(\theta_a) = a$, and θ_a maximizes $a\theta - \kappa(\theta)$. Before discussing the existence of θ_a, we will consider some examples.

Example 2.23 (Standard normal distribution)**.**

$$\begin{aligned}\phi(\theta) &= \int e^{\theta x}(2\pi)^{-1/2}\exp\left(-x^2/2\right)dx \\ &= \exp\left(\theta^2/2\right)\int (2\pi)^{-1/2}\exp\left(-(x-\theta)^2/2\right)dx\end{aligned}$$

The integrand in the last integral is the density of a normal distribution with mean θ and variance 1, so any $\theta \geqslant 0$,

$$\begin{aligned}\phi(\theta) &= \exp\left(\theta^2/2\right) \\ \kappa(\theta) &= \theta^2/2 \\ \kappa'(\theta) &= \theta\end{aligned}$$

In this case : $\mu = 0$, $I = [0, \infty)$ and $\kappa'(\theta) \in [0, \infty)$ when $\theta \in I = [0, \infty)$.

Example 2.24 (Exponential distribution with parameter λ)**.** If $0 \leqslant \theta < \lambda$,

$$\phi(\theta) = \int_0^\infty e^{\theta x}\lambda e^{-\lambda x}dx = \frac{\lambda}{\lambda - \theta}\,.$$

and if $\theta \geqslant \lambda$, $\phi(\theta) = \infty$. Thus $\theta_+ = \lambda$, and when $\theta \in [0, \lambda)$,

$$\begin{aligned}\kappa(\theta) &= \log\frac{\lambda}{\lambda - \theta} \\ \kappa'(\theta) &= \phi'(\theta)/\phi(\theta) = 1/(\lambda - \theta)\end{aligned}$$

In this case : $\mu = 1/\lambda$, $I = [0, \lambda)$ and $\kappa'(\theta) \in [1/\lambda, \infty)$ when $\theta \in I$.

Example 2.25 (Perverted exponential)**.** Let

$$g(x) = \begin{cases} Cx^{-3}e^{-x}\,, & x \geqslant 1\,. \\ 0\,, & x < 1\,. \end{cases}$$

Choose C so that g is a probability density. So,

$$\phi(\theta) = \int_1^\infty Cx^{-3}e^{(\theta-1)x}dx < \infty$$

if and only if $\theta \leqslant 1$, and when $\theta \leqslant 1$, we have,

$$\kappa'(\theta) = \frac{\int_1^\infty x^{-2}e^{(\theta-1)x}dx}{\int_1^\infty x^{-3}e^{(\theta-1)x}dx}$$

When $\theta \in [0,1]$,

$$\kappa'(\theta) \leqslant \kappa'(1) = \frac{\phi'(1)}{\phi(1)} = \frac{\int_1^\infty x^{-2}dx}{\int_1^\infty x^{-3}dx} = 2\,.$$

In this case : $I = [0,1]$, $\kappa'(\theta) \in [\mu, 2]$ when $\theta \in I$.

But if we let

$$h(x) = \begin{cases} Cx^{-2}e^{-x}\,, & x \geqslant 1\,. \\ 0\,, & x < 1\,. \end{cases}$$

So,

$$\phi(\theta) = \int_1^\infty Cx^{-2}e^{(\theta-1)x}dx < \infty$$

if and only if $\theta \leqslant 1$. When $\theta < 1$, we have,

$$\kappa'(\theta) = \frac{\int_1^\infty x^{-1}e^{(\theta-1)x}dx}{\int_1^\infty x^{-2}e^{(\theta-1)x}dx}$$

When $\theta \uparrow 1$,

$$\kappa'(\theta) \uparrow \frac{\int_1^\infty x^{-1}dx}{\int_1^\infty x^{-2}dx} = \infty\,.$$

In this case : $I = [0,1]$, $\kappa'(\theta) \in [\mu, \infty)$ when $\theta \in [0,1)$.

Example 2.26 (Coin filps)**.** $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Then for any $\theta \geqslant 0$, we have

$$\phi(\theta) = \frac{e^{\theta} + e^{-\theta}}{2}$$
$$\kappa(\theta) = \log \frac{e^{\theta} + e^{-\theta}}{2}$$
$$\kappa'(\theta) = \frac{e^{\theta} - e^{-\theta}}{e^{\theta} + e^{-\theta}}$$

In this case : $\mu = 0$, $I = [0, \infty)$ and $\kappa'(\theta) \in [0, 1)$ when $\theta \in I$.

Example 2.27 (**R.v. with upper bound**)**.** If $x_o = \sup\{x : F(x) < 1\} < \infty$, i.e. x_o is the essential upper bound of X_1, and X_1 is not a constant r.v., then $\theta_+ = \infty$, and $\kappa'(\theta) \uparrow x_o$ as $\theta \uparrow \infty$.
In this case : $I = [0, \infty)$ and $\kappa'(\theta) \in [\mu, x_o)$ when $\theta \in I$.

Proof. For any $\theta > 0$ there holds $Ee^{\theta X_1} \leqslant e^{\theta x_o} + 1 < \infty$, so $\theta_+ = \infty$.

$$\kappa'(\theta) = \frac{\phi'(\theta)}{\phi(\theta)} = \frac{\int_{(-\infty, x_o]} x e^{\theta x} dF(x)}{\int_{(-\infty, x_o]} e^{\theta x} dF(x)} \leqslant x_o \,.$$

On the other hand, for given $\epsilon > 0$,

$$\kappa'(\theta) \geqslant \frac{(x_o - \epsilon) \int_{(x_o - \epsilon, x_o]} e^{\theta x} dF(x)}{\int_{(-\infty, x_o - \epsilon]} e^{\theta x} dF(x) + \int_{(x_o - \epsilon, x_o]} e^{\theta x} dF(x)} \,.$$

since

$$\frac{\int_{(-\infty, x_o - \epsilon]} e^{\theta x} dF(x)}{\int_{(x_o - \epsilon, x_o]} e^{\theta x} dF(x)} \leqslant \frac{e^{\theta(x_o - \epsilon)}}{e^{\theta(x_o - \epsilon/2)}(F(x_o) - F(x_0 - \epsilon/2))} \to 0 \,, \quad \text{as } \theta \to \infty$$

Thus $\lim_{\theta \to \infty} \kappa'(\theta) \geqslant x_o - \epsilon$ for all $\epsilon > 0$, so $\kappa'(\theta) \uparrow x_o$. □

Theorem 2.37. *Suppose in addition to (H_1 : Cramér's condition) and (H_2 : nondegenerate condition) that there is a $\theta_a \in (0, \theta_+)$ so that $\kappa'(\theta_a) = a$, i.e., $a\theta - \kappa(\theta)$ arrive it's maximal value in the interior of I, then*

$$\gamma(a) = -\sup_{\theta \in I}(a\theta - \kappa(\theta)) = -a\theta_a + \kappa(\theta_a) \,.$$

Proof. By Theorem 2.36, we only need to prove

$$\gamma(a) \geqslant -a\theta_a + \kappa(\theta_a) .$$

To show this, we use the method of an exponential size-biasing of the distribution $\mu := P_{X_1}$ of X_1 which turns the atypical values that are of interest here into typical values. That is, for some fixed $\theta \in I$ (we will choose it later), we define the *Cramér transform* $\hat{\mu} \in \mathcal{M}_1(\mathbb{R})$ of μ by

$$\hat{\mu}(dx) = \phi(\theta)^{-1} e^{\theta x} \mu(dx), \text{ for } x \in \mathbb{R}.$$

Let $\hat{X}_1, \hat{X}_2, \ldots$ be independent and identically distributed with $\mathbb{P}_{\hat{X}_1} = \hat{\mu}$. Define $\hat{S}_n := \hat{X}_1 + \ldots + \hat{X}_n$, then

$$\begin{aligned} \mathbb{P}(S_n \geqslant na) &= \int_{\{x_1+\ldots+x_n \geqslant na\}} \mu(dx_1) \cdots \mu(dx_n) \\ &= \int_{\{x_1+\ldots+x_n \geqslant na\}} \phi(\theta)^n e^{-\theta(x_1+\cdots+x_n)} \hat{\mu}(dx_1) \cdots \hat{\mu}(dx_n) \\ &= \phi(\theta)^n \mathbb{E}\left[\exp(-\theta \hat{S}_n) 1_{\{\hat{S}_n \geqslant na\}}\right] . \end{aligned} \tag{2.28}$$

For any $\nu > a$, we have

$$\begin{aligned} \mathbb{P}(S_n \geqslant na) &= \phi(\theta)^n \mathbb{E}\left[\exp(-\theta \hat{S}_n) 1_{\{\hat{S}_n \geqslant na\}}\right] \\ &\geqslant \phi(\theta)^n \mathbb{E}\left[\exp(-\theta \hat{S}_n) 1_{\{na \leqslant \hat{S}_n \leqslant n\nu\}}\right] \\ &\geqslant \phi(\theta)^n e^{-n\nu\theta} \mathbb{P}(na \leqslant \hat{S}_n \leqslant n\nu) \end{aligned} \tag{2.29}$$

Thus

$$\frac{1}{n} \log \mathbb{P}(S_n \geqslant na) \geqslant -\nu\theta + \kappa(\theta) + \frac{1}{n} \log \mathbb{P}(na \leqslant \hat{S}_n \leqslant n\nu) \tag{2.30}$$

If we can show that as $n \to \infty$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(na \leqslant \hat{S}_n \leqslant n\nu) = 0 \tag{2.31}$$

From (2.30) we have $\gamma(a) \geqslant -\nu\theta + \kappa(\theta)$. Then assume we can let $\theta \to \theta_a$ and $\nu \to a$, then we get the required result.

To prove (2.31), we want to use law of large number: if $a < \mathbb{E}\hat{X}_1 < \nu$, then we get $\mathbb{P}(na \leqslant \hat{S}_n \leqslant n\nu) \to 1$, which deduces (2.31). We compute that

$$\mathbb{E}\hat{X}_1 = \int x\hat{\mu}(dx) = \phi(\theta)^{-1} \int xe^{\theta x}\mu(dx) = \phi'(\theta)/\phi(\theta) = \kappa'(\theta).$$

So we pick $\theta \in (\theta_a, \theta_+)$ arbitrarily at beginning, and choose $\nu > \theta$ arbitrarily. We get for any $\theta_a < \theta < \nu < \theta_+$,

$$\gamma(a) \geqslant -\nu\theta + \kappa(\theta)\,.$$

And we let $\theta \downarrow \theta_a$ and $\nu \downarrow a$, then $\gamma(a) \geqslant -a\theta_a + \kappa(\theta_a)$. □

Turning now to the problematic values for which we can't solve $\kappa'(\theta) = a$ in $(0, \theta_+)$.

Lemma 2.38. *If $x_o = \sup\{x : F(x) < 1\} < \infty$, and X_1 is not a constant r.v., then*

(i) $a = x_o$, $\gamma(a) = \gamma(x_o) = \log \mu\{x_o\} = -\sup_{\theta\in I}(a\theta - \kappa(\theta))$.

(ii) $a > x_o$, $\gamma(a) = -\infty = -\sup_{\theta\in I}(a\theta - \kappa(\theta))$.

Proof. For $a = x_o$ is trivial: for all $n \in \mathbb{N}_+$, we have

$$n^{-1} \log \mathbb{P}\left(S_n \geqslant nx_o\right) = \log \mathbb{P}\left(X_1 = x_o\right).$$

Thus $\gamma(x_o) = \log \mu\{x_o\}$. Then we show that $-\sup_{\theta\in I}(x_o\theta - \kappa(\theta)) = \log \mu\{x_o\}$.

Since

$$\kappa(\theta) - x_o\theta = \log \frac{\phi(\theta)}{e^{\theta x_o}}\,,$$

and $\kappa' < x_o$, as $\theta \to \infty$, by Lebesgue domainted convergence theorem,

$$\frac{\phi(\theta)}{e^{\theta x_o}} = \int_{(-\infty, x_o]} e^{\theta(x-x_o)}\mu(dx) \downarrow \log \mu\{x_o\}\,.$$

Thus $-\sup_{\theta\in I}(x_o\theta - \kappa(\theta)) = \log \mu\{x_o\}$.

Obviously, when $a > x_o$, $\gamma(a) = -\infty$ and $-\sup_{\theta\in I}(a\theta - \kappa(\theta)) = -\infty$. □

EXERCISE 17. show that as $a \uparrow x_o, \gamma(a) \downarrow \log \mathbb{P}(X_i = x_o)$.

Therefore, we noly need to discuss the case that X_1 does not have an upper bound, i.e., $x_o = \infty$.

Lemma 2.39. *If X_1 does not have an upper bound, $\theta_+ = \infty$, then $\kappa'(\theta) \uparrow \infty$ as $\theta \uparrow \infty$. So for all $a > \mu$, the condition of Theorem2.37 satisfied.*

Proof. For any $M > 0$,

$$\kappa'(\theta) = \frac{\int x e^{\theta x} dF(x)}{\int e^{\theta x} dF(x)} \geqslant \frac{M \int_{(M,\infty]} e^{\theta x} dF(x)}{\int_{(-\infty,M]} e^{\theta x} dF(x) + \int_{(M,\infty]} e^{\theta x} dF(x)}.$$

since

$$\frac{\int_{(-\infty,M]} e^{\theta x} dF(x)}{\int_{(M,\infty]} e^{\theta x} dF(x)} \leqslant \frac{e^{\theta M}}{e^{2\theta M)}(1 - F(2M))} \to 0, \quad \text{as } \theta \to \infty$$

Thus $\lim_{\theta\to\infty} \kappa'(\theta) \geqslant M$ for all $M > 0$, so $\kappa'(\theta) \uparrow \infty$. □

Assume $\theta_+ < \infty$, if $\kappa'(\theta) \uparrow \infty$ as $\theta \uparrow \theta_+$. Then for all $a > \mu$, the condition of Theorem2.37 satisfied.

Now the only case that remains is

Theorem 2.40. *Suppose X_1 doesn't have an upper bound, $\theta_+ < \infty$, and $\kappa'(\theta)$ increases to a finite limit a_0 as $\theta \uparrow \theta_+$. Then $\kappa(\theta_+) < \infty$ and for $a \geqslant a_0$,*

$$\gamma(a) = -\sup_{\theta \in I}(a\theta - \kappa(\theta)) = -a\theta_+ + \kappa(\theta_+)$$

Particularly, $\gamma(a)$ is linear for $a \geqslant a_0$.

Proof. **Step 1**. Using Cauchy's convergence test, we have $\kappa(\theta_+) < \infty$. By Lemma 2.35, $\kappa'(\theta_+) = a_0$. Letting $\theta = \theta_+$ in (2.27) shows that

$$\gamma(a) \leqslant -a\theta_+ + \kappa(\theta_+).$$

To get the other direction we use the Cramér transform $\hat{\mu} \in \mathcal{M}_1(\mathbb{R})$ of μ choosing parameter θ_+, that is,

$$\hat{\mu}(dx) = \phi(\theta_+)^{-1} e^{\theta_+ x} \mu(dx), \text{ for } x \in \mathbb{R}.$$

From (2.30) in the proof of Theorem 2.37, we see that if $\kappa'(\theta_+) = a_0 \leqslant a < \nu$,

$$\frac{1}{n}\log \mathbb{P}\left(S_n \geqslant na\right) \geqslant -\nu\theta_+ + \kappa\left(\theta_+\right) + \frac{1}{n}\log \mathbb{P}(na \leqslant \hat{S}_n \leqslant n\nu)\,, \tag{2.32}$$

and we still try to show

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}(na \leqslant \hat{S}_n \leqslant n\nu) = 0\,. \tag{2.33}$$

If (2.33) holds, let $n \to \infty$ in (2.32), we have $\gamma(a) \geqslant -\nu\theta_+ + \kappa(\theta_+)$ for any $\nu > a$. Let $\nu \downarrow a$, we get $\gamma(a) \geqslant -a\theta_+ + \kappa(\theta_+)$.

Step 2. Now we only need to prove (2.33). Note that for $0 < \epsilon < (\nu - a)/2$,

$$\begin{aligned}\mathbb{P}\left(\hat{S}_n \in [an, \nu n]\right) \geqslant\ & \mathbb{P}\left(\hat{S}_{n-1} \in [(a_0 - \epsilon)\, n, (a_0 + \epsilon)\, n]\right) \\ & \times \mathbb{P}\left(\hat{X}_1 \in [(a - a_0 + \epsilon)\, n, (v - a_0 - \epsilon)\, n]\right)\,.\end{aligned}$$

Let $\delta_1 \coloneqq a - a_0 + \epsilon < \delta_2 \coloneqq v - a_0 - \epsilon$. By law of large number, for sufficiently large n we have

$$\mathbb{P}\left(\hat{S}_{n-1} \in [(a_0 - \epsilon)\, n, (a_0 + \epsilon)\, n]\right) \geqslant \frac{1}{2}\,.$$

Thus, to prove (2.33), we only need to show that

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}\left(\hat{X}_1 \in [\delta_1 n, \delta_2 n]\right) = 0\,. \tag{2.34}$$

Step 3. Now we only need to prove (2.34). If not ture, there exists some $\beta > 0$ such that

$$\mathbb{P}\left(\hat{X}_1 \in [\delta_1 n, \delta_2 n]\right) \leqslant e^{-\beta n}\,, \quad \text{for large } n\,.$$

Thus if n sufficiently large,

$$\mathbb{P}\left(\hat{X}_1 \geqslant \delta_1 n\right) \leqslant \sum_{k=n}^{\infty} e^{-\beta k} = \frac{1}{1 - e^{-\beta}} e^{-\beta n}$$

Thus, by Lemma 2.34, there exists some $\theta_0 > 0$ such taht

$$\mathbb{E}e^{\theta_0 \hat{X}_1} = \int e^{\theta_0 x}\hat{\mu}(dx) = \frac{1}{\phi(\theta_+)}\int e^{(\theta_0+\theta_+)x}\mu(dx) < \infty\,.$$

So $\phi(\theta_0 + \theta_+) < \infty$, which contradicts $\theta_+ = \sup\{\theta : \phi(\theta) < \infty\}$. □

Combine all the case, we finally get the desired result.

Theorem 2.41. *X_1 satisfies (H_1 : Cramér's condition) and (H_2 : nondegenerate condition). Then we have*

$$\gamma(a) = -\sup_{\theta \in I}(a\theta - \kappa(\theta)) .$$

To get a feel for what the answers look like, we consider our examples.

Standard normal distribution (Example 2.23)

$$\kappa(\theta) = \theta^2/2 \quad \kappa'(\theta) = \theta$$

for $a > 0$,

$$\theta_a = a$$
$$\gamma(a) = -\frac{a^2}{2}$$

Exponential distribution (Example 2.24) with $\lambda = 1$,

$$\kappa(\theta) = -\log(1-\theta) \quad \kappa'(\theta) = \frac{1}{1-\theta}$$

for a > 1

$$\theta_a = 1 - \frac{1}{a}$$
$$\gamma(a) = -a + 1 + \log a$$

Coin flips (Example 2.26)

$$\phi(\theta) = \frac{e^\theta + e^{-\theta}}{2}, \quad \kappa(\theta) = \log\frac{e^\theta + e^{-\theta}}{2}, \quad \kappa'(\theta) = \frac{e^\theta - e^{-\theta}}{e^\theta + e^{-\theta}}$$

for $0 < a < 1$,

$$\theta_a = \frac{\log(1+a) - \log(1-a)}{2}$$
$$\phi(\theta_a) = \frac{e^{\theta_a} + e^{-\theta_a}}{2} = \frac{1}{\sqrt{(1+a)(1-a)}}$$
$$\gamma(a) = \frac{(1+a)\log(1+a) + (1-a)\log(1-a)}{2}$$

2.7 Midterm exam

1. Let $X_1, X_2, \ldots$ be i.i.d. with $P(X_i > x) = e^{-x}$, let $M_n = \max_{1 \le m \le n} X_m$. Show that

(a. 15pt) $\limsup_{n\to\infty} X_n / \log n = 1$ a.s.

(b. 15pt) $M_n / \log n \to 1$ a.s.

2. Let $\{c_n\}$ be positive real numbers between 0 and 2π, $\{\Theta_n\}$ be i.d.d. random variables uniformly distributed on $[0, 2\pi]$. For wach n define I_n be a random arc on the unit circle, centered at $(\cos\Theta_n, \sin\Theta_n)$ with arc length c_n.

(a. 15pt) Show that for fixed θ_0,

$$P((\cos\theta_0, \sin\theta_n) \in I_n \text{ i.o.})$$

has to equal to 0 or 1. Give the necessary and sufficient condition such probability=1.

(b. 10pt) show that the bi-variate mapping

$$(\theta, \omega) \mapsto 1_{\{(\cos\theta, \sin\theta) \in I_n \text{ i.o.}\}}$$

is measurable with respect to $([0, 2\pi] \times \Omega), \mathcal{B}([0, 2\pi]) \times \mathcal{F})$.

(c. 5pt) What can you say (almost surely) about the Lebesgue measure of the set of points on the unit circle covered by infinite many arcs?

3. Consider $\{X_n\}_{n=1}^{\infty}$ be an independent sequence of random exponential variables with mean n, respectively.

(a. 10pt) Show that $\sum_n X_n = \infty$ a.s.

(b. 20pt) For each $k \geqslant 1$, define indicator function $Y_k = 1_{\{X_k > k^{-1}\}}$ and $Z_n = \sum_{k=1}^{n} Y_k$. Calculate

$$\lim_{n\to\infty} \frac{1}{n} \log P(Z_n < \frac{n}{10})$$

(c. 10pt) Let $S_n = \sum_{k=1}^{n} X_k$. Prove that for all sufficiently large n

$$P\big(S_n < \frac{\log n}{20}\big) < \exp(-\frac{1}{n^4})\,.$$

Chapter 3

Centerl Limit Theorems

3.1 Weak convergence

3.1.1 Motivation

If a sequence of r.v.'s $\{X_n\}$ tends to a limit, the corresponding sequence of p.m. 's $\{\mu_n\}$ ought to tend to a limit in some sense. Is it true that $\lim_n \mu_n(A)$ exists for all $A \in \mathcal{B}$ or at least for all intervals A ? The answer is NO from trivial examples.

Example 3.1. Let $X_n = c_n$ where the c_n's are constants tending to zero. Then $X_n \to 0$ deterministically.

(i) For any interval I such that $0 \notin \bar{I}$, we have $\lim_n \mu_n(I) = 0 = \mu(I)$; for any interval such that $0 \in I^\circ$, where I° is the interior of I, we have $\lim_n \mu_n(I) = 1 = \mu(I)$.

(ii) But if $\{c_n\}$ oscillates between strictly positive and strictly negative values, and $I = (a, 0)$ or $(0, b)$, where $a < 0 < b$, then $\mu_n(I)$ oscillates between 0 and 1, while $\mu(I) = 0$. On the other hand, if $I = (a, 0]$ or $[0, b)$, then $\mu_n(I)$ oscillates as before but $\mu(I) = 1$.

Observe that $\{0\}$ is the sole *atom* of μ and it is the root of the trouble.

Example 3.2. Instead of the point masses concentrated at c_n, we may consider, e.g., r.v.'s $\{X_n\}$ having uniform distributions over intervals (c_n, c_n') where $c_n < 0 < c_n'$ and c_n, $c_n' \to 0$. Then again $X_n \to 0$ a.e., but $\mu_n((a,0))$ may not converge at all, or converge to any number between 0 and 1.

Example 3.3. Let X have distribution μ, and $X_n = X + 1/n$ has distribution

$$F_n(x) = P(X + 1/n \leqslant x) = F(x - 1/n)$$

As $n \to \infty, F_n(x) \to F(x-) = \lim_{y \uparrow x} F(y)$ so convergence only occurs at continuity points.

The three example above show that, we should restrict our attention to "continuity" points.

Example 3.4. If $X_n \to X$ in probability, then for every continuty piont x of F, we have

$$F_n(x) \to F(x)\,.$$

To see this, note that for any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}(X_n \leqslant x) &= \mathbb{P}(X_n \leqslant x, |X_n - X| \leqslant \epsilon) + \mathbb{P}(X_n \leqslant x, |X_n - X| > \epsilon) \\ &\leqslant \mathbb{P}(X \leqslant x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon)\,,\end{aligned}$$

letting $n \to \infty$, then $\epsilon \to 0$, we have

$$\limsup_{n\to\infty} F_n(x) \leqslant F(x)\,.$$

On the other hand,

$$\begin{aligned}\mathbb{P}(X_n > x) &= \mathbb{P}(X_n > x\,, |X_n - X| \leqslant \epsilon) + \mathbb{P}(X_n > x, |X_n - X| > \epsilon) \\ &\leqslant \mathbb{P}(X > x - \epsilon) + \mathbb{P}(|X_n - X| > \epsilon)\,,\end{aligned}$$

letting $n \to \infty$, then $\epsilon \to 0$, we have

$$F(x^-) \leqslant \liminf_{n\to\infty} F_n(x)\,.$$

Thus, when F is continuous at x,

$$\lim_{n\to\infty} \mathbb{P}\left(f_n \leqslant x\right) = F(x)\,.$$

An interval I is called a *continuity interval* of μ if $\mu(\partial I) = 0$. Using this term, the Example 3.4 is saying that, if $X_n \xrightarrow{\mathbb{P}} X$, then the conresponding distributions μ_n and μ satisfying

$$\mu_n(a, b] \to \mu(a, b]\,, \tag{3.1}$$

for every continuity interval $(a, b]$ of μ.

3.1.2 Weak convergence

Weak convergence In order to describe the type of convergnece in (3.1), we introduce the definitions of weak convergence between probability measures and convergence in distribution between random variables.

Definition 3.1.

- μ_n and μ are p.m.'s on $(\mathbb{R}, \mathcal{B})$. $\{\mu_n\}$ is said to **converge weakly** to μ, written $\mu_n \Rightarrow \mu$ or $\mu_n \xrightarrow{w} \mu$, if

  $$\mu_n(a, b] \to \mu(a, b]$$

  for any continuty interval $(a, b]$ of μ.

- F_n and F are c.d.f.'s on $\mathbb{R}$. $\{F_n\}$ is said to **converge weakly** to F, written $F_n \Rightarrow F$ or $F_n \xrightarrow{w} F$, if

  $$F_n(x) \to F(x)$$

  for all x that are continuity points of F.

- X_n and X are r.v's, $\{X_n\}$ is said to **converge in distribution** to X, written $X_n \Rightarrow X$, or $X_n \xrightarrow{d} X$, if their c.d.f.'s $\{F_n\}$ converge weakly to F, or equivalently, their distributions $\{\mu_n\}$ converge weakly to μ.

We give a really important equivalent definition of weak convergence.

Theorem 3.1. *μ_n and μ are p.m.'s on $(\mathbb{R}, \mathcal{B})$. Then $\mu_n \Rightarrow \mu$ if and only if*

$$\int f\, d\mu_n \to \int f d\mu\,,\ \forall f \in C_b(\mathbb{R})\,. \tag{3.2}$$

Proof. Take any $f \in C_b(\mathbb{R})$.

Step 1. we shall show that for each $(a, b]$, a finite continuous interval of μ, there holds

$$\int_{(a,b]} f \, \mathrm{d}\mu_n \to \int_{(a,b]} f \, \mathrm{d}\mu.$$

Since f is uniformly continuous on $[a, b]$, given $\epsilon > 0$, there exists δ, depending on ϵ, such taht for each $\xi, \eta \in [a, b]$, $|\xi - \eta| < \delta$, we have

$$|f(\xi) - f(\eta)| < \epsilon.$$

Take a partition of the interval, $a = x_0 < x_1 < \cdots < x_k = b$ such that each x_j isn't an atom of μ, and $\sup_{1 \leqslant j \leqslant k} |x_j - x_{j-1}| < \delta$. Let

$$\phi \coloneqq \sum_{j=1}^{k} f(x_j) I_{(x_{j-1}, x_j]} \,,$$

then $|f(x) - \phi(x)| < \epsilon$ for all $x \in (a, b]$. By the definition of weak convergence we have

$$\int_{(a,b]} \phi \, \mathrm{d}\mu_n \to \int_{(a,b]} \phi \, \mathrm{d}\mu \,.$$

Thus

$$\left| \int_{(a,b]} f \, \mathrm{d}(\mu_n - \mu) \right| \leqslant \left| \int_{(a,b]} (f - \phi) \, \mathrm{d}\mu_n \right| + \left| \int_{(a,b]} \phi \, \mathrm{d}(\mu_n - \mu) \right| + \left| \int_{(a,b]} (f - \phi) \, \mathrm{d}\mu \right| ,$$

letting $n \to \infty$, and note that μ_n, μ are p.m.'s we have

$$\limsup_{n \to \infty} \left| \int_{(a,b]} f \, \mathrm{d}\mu_n - \int_{(a,b]} f \, \mathrm{d}\mu \right| \leqslant 2\epsilon \,.$$

Since ϵ is arbitary, we get the desired result.

Step 2. Now we deal with the infinite interval. Given any $\epsilon > 0$, there exists a continuty interval of μ, depending on ϵ, denoted by $(a, b]$, such that

$$\mu\big(\mathbb{R} \backslash (a, b]\big) < \epsilon.$$

Since $\mu_n \Rightarrow \mu$, there are some positive integer N, depending on ϵ, such that for all $n \geqslant N$,

$$\mu_n\big(\mathbb{R} \backslash (a, b]\big) < \epsilon.$$

Thus

$$\left|\int_{\mathbb{R}} f \,\mathrm{d}(\mu_n - \mu)\right| \leqslant \left|\int_{(a,b]} f \,\mathrm{d}(\mu_n - \mu)\right| + \left|\int_{\mathbb{R}\setminus(a,b]} f \,\mathrm{d}(\mu_n - \mu)\right| ,$$

letting $n \to \infty$, and note that f is bounded

$$\limsup_{n\to\infty} \left|\int_{\mathbb{R}} f \,\mathrm{d}(\mu_n - \mu)\right| \leqslant 2\|f\|_\infty \epsilon$$

Since ϵ is arbitary, the thoerem follows. □

REMARK. Denoted by $\mathcal{M}_1(\mathbb{R})$ all the p.m.'s on $(\mathbb{R}, \mathcal{B})$. $\mathcal{M}_1(\mathbb{R})$ is a subset of $C_b(\mathbb{R})^*$, and μ_n converge weakly to μ is equivalent to μ_n converge to μ with respect to the weak* topology on $C_b(\mathbb{R})^*$.

Corollary 3.2. *$X_n \Rightarrow X$ if and only if for every $f \in C_b(\mathbb{R})$, we have*

$$\mathbb{E}f\left(X_n\right) \to \mathbb{E}f\left(X\right) . \tag{3.3}$$

Method of a single probability space If X_n and X are r.v.'s so that $X_n \xrightarrow{\text{a.s.}} X$, we can see the corresponding sequence of p.m.'s $\mu_n \Rightarrow \mu$ by Example 3.6 or by Corollary 3.2. More unexpectedly, in a certain sense there is a converse result, the precise formulation and application we now turn to.

Theorem 3.3 (Skorokhod). *μ_n and μ are p.m.'s on $(\mathbb{R}, \mathcal{B})$ and $\mu_n \Rightarrow \mu$. We can find random variables X_n and X with distributions μ_n and μ, so that*

$$X_n \to X \quad \text{a.s.} \tag{3.4}$$

Proof. We have learned that for a distribution function F, then the generalized inverse of F, defined by

$$F^{\leftarrow}(t) \coloneqq \inf\{x : F(x) \geqslant t\} = \sup\{x : F(x) < t\} ,$$

for all $t \in (0,1)$ is a r.v. on probability space $((0,1), \mathcal{B} \cap (0,1), \lambda)$ with distribution function F, where λ is the Lebesgue measure.

Let F_n and F be the distribution functions corresponding to the o.m.'s μ_n and μ. It suffices to show that, $F_n^{\leftarrow} \xrightarrow{\text{a.s.}} F^{\leftarrow}$.

To see this, let t be a continuty point of $F^{\leftarrow}$, we give a proof by contradiction. If $\lim_n F_n^{\leftarrow}(t) = F^{\leftarrow}(t)$ not ture, then

$$\text{either } \limsup_{n\to\infty} F_n^{\leftarrow}(t) > F^{\leftarrow}(t) \text{ or } \liminf_{n\to\infty} F_n^{\leftarrow}(t) < F^{\leftarrow}(t) \text{ or both.}$$

If the first inequality holds, take a continuty point y of F, such that

$$F^{\leftarrow}(t) < y < \limsup_{n\to\infty} F_n^{\leftarrow}(t)$$

Then there exists $\{n_k\}$ so that $F_{n_k}(y) < t$, and $t \leqslant F(y)$. Letting $k \to \infty$ we have $F(y) = t$. But this contradicts that $F^{\leftarrow}$ is continuous at t : we can find sufficiently small $\epsilon > 0$ that $F^{\leftarrow}(t+\epsilon) < y$, then $F(y) \geqslant t + \epsilon$.

If the second inequality holds, take a continuty point y of F, such that

$$\liminf_{n\to\infty} F_n^{\leftarrow}(t) < y < F^{\leftarrow}(t)$$

Then there exists $\{n_k\}$ so that $F_{n_k}(y) \geqslant t$, and $t < F(y)$. Letting $k \to \infty$ we have $t < F(y) = t$, which is a contradiction ! □

Theorem 3.4 (Continuous mapping theorem). *Let g be a measurable function on $(\mathbb{R}, \mathcal{B})$, and denote by D_g all the discontinuty point of g. If $X_n \Rightarrow X$ and $\mathbb{P}(X \in D_g) = 0$, then*

$$g(X_n) \Rightarrow g(X).$$

If in addition g is bounded then $\mathbb{E}\, g(X_n) \to \mathbb{E}\, g(X_\infty)$.

Proof. First, note that D_g is always a Borel set, thus measurable. By the method of a single probability space, there are r.v.'s Y_n and Y such that

$$Y_n \overset{d}{=} X_n,\ Y \overset{d}{=} X \quad \text{and} \quad Y_n \to Y \text{ , a.s.}$$

For any $f \in C_b(\mathbb{R})$, note that $D_{fog} \subset D_g$, then $\mathbb{P}(Y \in D_{fog}) = \mathbb{P}(X \in D_{fog}) = 0$ and it follows that

$$f(g(Y_n)) \to f(g(Y)) ,$$

by the bounded convergence theorem, we have

$$\mathbb{E}f(g(Y_n)) \to \mathbb{E}f(g(Y)) \quad \text{a.s.}$$

and it follows from Corollary 3.2 that $g(X_n) \Rightarrow g(X)$. □

Alternative definitions of weak convergence The next result provides a number of useful alternative definitions of weak convergence.

Theorem 3.5. *μ_n and μ are p.m.'s on $(\mathbb{R}, \mathcal{B})$. The following statements are equivalent:*

(i) $\mu_n \Rightarrow \mu$.

(ii) *For all closed sets F,* $\limsup \mu_n(F) \leqslant \mu(F)$.

(iii) *For all open sets G,* $\liminf \mu_n(G) \geqslant \mu(G)$.

(iv) *For all Borel sets A with $\mu(\partial A) = 0$,* $\lim \mu_n(A) = \mu(A)$.

REMARK. To help remember the directions of the inequalities in (ii) and (iii), consider the special case in which $\mu_n = \delta_{x_n}$ and $\mu = \delta_x$. In this case, if $x_n \notin F$ and $x_n \to x \in \partial F$, then $\mu_n(F) = 0$ for all n but $\mu(F) = 1$. Letting $K = G^c$ gives an example for (iii).

Proof. Clearly, (ii) is equivalent to (iii) and (iv) implies (i) are obvious.

(i) implies (ii): We give two proofs. The first one uses the method of a single probability space. Take Y_n and Y with distribution μ_n, μ, and $Y_n \to Y$ a.s. Since F is closed

$$\limsup_{n\to\infty} 1_{\{Y_n \in F\}} \leqslant 1_{\{Y \in F\}} , \text{ a.s. } ,$$

so Fatou's lemma implies

$$\limsup_{n\to\infty} \mu_n(F) \leqslant \mu(F)\,.$$

The second one uses Theorem 3.1. For any positive integer k, define f_k by

$$f(x) \coloneqq (1 - kd(x, F)) \vee 0\,, \text{ for all } x\,.$$

By Theorem 3.1, $\int f_k \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$. thus

$$\limsup_{n} \mu_n(F) \leqslant \int f_k \mathrm{d}\mu.$$

Letting $k \to \infty$, by bounded convergence theorem, (ii) holds.

(ii) and (iii) imply (iv): Note that $\mu(\partial A) = 0$ means

$$\mu(A^\circ) = \mu(A) = \mu(\bar{A})$$

then by (ii), (iii) we get the desired result. □

3.1.3 Vague convergence

Example 3.5. Let $a_n \to -\infty$, $b_n \to +\infty$ and

$$X_n = \begin{cases} a_n & \text{with probability } \alpha \\ 0 & \text{with probability } 1-\alpha-\beta \\ b_n & \text{with probability } \beta \end{cases}$$

Then $X_n \to X$ where

$$X = \begin{cases} +\infty & \text{with probability } \alpha \\ 0 & \text{with probability } 1-\alpha-\beta \\ -\infty & \text{with probability } \beta \end{cases}$$

For any finite interval (a, b) containing 0 we have

$$\lim_n \mu_n(a, b) = \lim_n \mu_n\{0\} = 1 - \alpha - \beta$$

In this situation it is said that masses of amount α and β "have wandered off to $+\infty$ and $-\infty$ respectively." The remedy here is obvious: we should consider measures on the extended line $\overline{\mathbb{R}} = [-\infty, +\infty]$, with possible atoms at $\{+\infty\}$ and $\{-\infty\}$.

We give the appropriate definitions which take into account the troubles discussed above. A measure μ on $(\mathbb{R}, \mathcal{B})$ with $\mu(\mathbb{R}) \leqslant 1$ will be called a **subprobability measure (s.p.m.)**.

Definition 3.2. A sequence $\{\mu_n, n \geqslant 1\}$ of s.p.m.'s is said to **converge vaguely** to an s.p.m. μ, denoted by $\mu_n \xrightarrow{v} \mu$, if

$$\mu_n(a, b] \to \mu(a, b]$$

for any continuty interval $(a, b]$ of μ.

REMARK. We should point that, the limit of a vague convergence sequence is unique.

As in Theorem 3.1, there is a similar result. First, we denote by $C_c(\mathbb{R})$ the class of continuous functions on $\mathbb{R}$ having compact support, $C_0(\mathbb{R})$ the class of continuous functions on $\mathbb{R}$ vanishing at infty , i.e.,

$$\lim_{|x|\to\infty} f(x) = 0\,.$$

Theorem 3.6. *Let $\{\mu_n\}$ and μ be s.p.m's. Then $\mu_n \xrightarrow{v} \mu$ if and only if*

$$\int f(x)\mu_n(dx) \to \int f(x)\mu(dx) \tag{3.5}$$

for each $f \in C_c(\mathbb{R})$ (or $C_0(\mathbb{R})$).

Proof. Using the same method in the proof of Theorem 3.1. □

REMARK. We should be careful when $f \in C_b(\mathbb{R})$, since $\int f d\mu_n \to \int f d\mu$ may not hold . For an example, let $f \equiv 1$ and the sequence $\{\mu_n\}$ consists only of strict p.m.'s, but the sequential vague limit may not be so, see Example 3.6.

Sequential compactness of vague convergence We will discuss the sequential compactness of vague convergence. Surprisingly, $\mathcal{M}_{\leqslant 1}$, all the s.p.m.'s on $(\mathbb{R}, \mathcal{B})$, is sequentially compact with respect to vague convergence.

Theorem 3.7. *Given any sequence of s.p.m.'s* $\{\mu_n\}$ *, there is a subsequence that converges vaguely to an s.p.m* μ*.*

Proof. Here it is convenient to consider the *subdistribution function(s.d.f.)* F_n defined as follows:

$$F_n(x) = \mu_n(-\infty, x]\,, \quad \text{for all } x \in \mathbb{R}\,.$$

F_n is increasing, right continuous and $F_n(-\infty) = 0\,,\ F_n(\infty) = \mu_n(\mathbb{R}) \leqslant 1$.

Let D be a countable dense set of $\mathbb{R}$, and let $\{r_k, k \geqslant 1\}$ be an enumeration of it. The sequence of numbers $\{F_n(r_1), n \geqslant 1\}$ is bounded, hence by the Bolzano-Weierstrass theorem there is a subsequence $\{F_{1k}, k \geqslant 1\}$ of the given sequence such that the limit

$$\lim_{k\to\infty} F_{1k}(r_1) = \ell_1$$

exists; clearly $0 \leqslant \ell_1 \leqslant 1$. Next, the sequence of numbers $\{F_{1k}(r_2), k \geqslant 1\}$ is bounded, hence there is a subsequence $\{F_{2k}, k \geqslant 1\}$ of $\{F_{1k}, k \geqslant 1\}$ such that

$$\lim_{k\to\infty} F_{2k}(r_2) = \ell_2$$

where $0 \leqslant \ell_2 \leqslant 1$. since $\{F_{2k}\}$ is a subsequence of $\{F_{1k}\}$, it converges also at r_1 to ℓ_1. Continuing, we obtain

$$\begin{array}{ll}
F_{11}, F_{12}, \cdots, F_{1k}, \cdots & \text{converging at } r_1 \\
F_{21}, F_{22}, \cdots, F_{2k}, \cdots & \text{converging at } r_1, r_2 \\
\cdots\cdots\cdots\cdots\cdots & \\
F_{j1}, F_{j2}, \cdots, F_{jk}, \cdots & \text{converging at } r_1, r_2, \cdots, r_j \\
\cdots\cdots\cdots\cdots\cdots &
\end{array}$$

Now consider the diagonal sequence $\{F_{kk}, k \geqslant 1\}$. We assert that it converges at every $r_j, j \geqslant 1$. To see this let r_j be given. Apart from the first $j-1$ terms, the sequence $\{F_{kk}, k \geqslant 1\}$ is a subsequence of $\{F_{jk}, k \geqslant 1\}$, which converges at r_j and hence $\lim_{k\to\infty} F_{kk}(r_j) = \ell_j$, as desired.

We have thus proved the existence of an infinite subsequence $\{n_k\}$ and a function G defined and increasing on D such that

$$G(r) \coloneqq \lim_{k\to\infty} F_{n_k}(r)\,, \text{ for all } r \in D\,.$$

From G we define a function F on $\mathbb{R}$ as follows:

$$F(x) \coloneqq \inf_{r>x} G(r)\,, \text{ for all } x\,.$$

Clearly, F is increasing and right continuous. In fact, for $x < y < r$, we have $F(x) \leqslant F(y) \leqslant G(r)$, letting $r \to x$, by the definition of $F(x)$, we have $F(y) \to F(x)$, thus F is right continuous.

To complete the proof, we need to show that any continuity point x of F,

$$\lim_{k\to\infty} F_{n_k}(x) = F(x)\,.$$

Note that for r, $s \in D$, $r < x < s$, we have $F_{n_k}(r) \leqslant F_{n_k}(x) \leqslant F_{n_k}(s)$, then

$$F(r) \leqslant G(r) \leqslant \liminf_{k\to\infty} F_{n_k}(x) \leqslant \limsup_{k\to\infty} F_{n_k}(x) \leqslant G(s)\,,$$

letting $r \uparrow x$ and $s \downarrow x$, since F is continuous at x, the desired result follows. □

Corollary 3.8. *If every vaguely convergent subsequence of the sequence of s.p.m.'s $\{\mu_n\}$ converges to the same μ, then $\mu_n \xrightarrow{v} \mu$.*

Proof. We use arguement by controdiction. If $\{\mu_n\}$ doesn't vaugely converge to μ, there exists an interval $(a, b]$, $\epsilon > 0$ and a sequence $\{\mu_{n_k}\}$ such that

$$|\,\mu_{n_k}(a, b] - \mu(a, b]\,| > \epsilon\,.$$

However, there is a subsequence of $\{\mu_{n_k}\}$, denoted by $\{\mu_{n_{k_p}}\}$, vaguely convergent to μ, which is a contradiction. □

Corollary 3.9. *If $\{\mu_n\}$ is a sequence of s.p.m.'s such that for every $f \in C_c(\mathbb{R})$,*

$$\lim_n \int f(x)\mu_n(dx)$$

exists, then $\{\mu_n\}$ converges vaguely.

Proof. Since every vaguely convergent subsequence of the sequence of s.p.m.'s $\{\mu_n\}$ converges to the same μ. □

Tightness Theorem 3.7 deal with s.p.m's, even if the given sequence $\{\mu_n\}$ consists only of strict p.m's, the sequential vague limit may not be so.

Example 3.6. The limit may not be a distribution function. For example, let $\mu_n = \frac{1}{3}\delta_n + \frac{1}{3}\delta_{-n} + \frac{1}{3}\nu$ where δ_a is the point mass at a and ν is a p.m., then μ_n converges vaguely to $\frac{1}{3}\nu$, or

$$\mu_n(x, y] \to \frac{1}{3}\,\nu(x, y]$$

for any interval $(x, y]$. In words, an amount of mass $1/3$ escapes to ∞, and mass $1/3$ escapes to $-\infty$.

So there raises a natural question: When can we conclude that no mass is lost in the limit in Theorem 3.7 ?

Definition 3.3. A family of p.m.'s $\{\mu_\alpha\}$ is called **tight**, if for any $\epsilon > 0$, there exists a compact set K, depending on ϵ, so that

$$\inf_\alpha \mu_\alpha(K) > 1 - \epsilon\,, \quad \text{or equivalently,} \quad \sup_\alpha \mu_\alpha(K^c) < \epsilon\,.$$

Theorem 3.10. *$\{\mu_\alpha, \alpha \in A\}$ is a family of p.m.'s. In order that every sequence of them contains a subsequence which converges weakly to a p.m., it is necessary and sufficient that $\{\mu_\alpha, \alpha \in A\}$ is tight.*

Proof. Sufficiency. Suppose $\{\mu_n\}$ contained in $\{\mu_\alpha, \alpha \in A\}$ is weakly convergent to μ, since $\{\mu_n\}$ is tight, for any given $\epsilon > 0$, there exist a continuty interval $[-M, M]$, depending on ϵ, so that

$$\inf_n \mu_n[-M, M] > 1 - \epsilon\,.$$

Letting $n \to \infty$ then we get $\mu\,[-M, M] > 1-\epsilon$. Since ϵ is arbitray, $\mu(\mathbb{R}) = 1$.

Necessity. Conversely, if $\{\mu_\alpha\}$ is not tight, then there exists $\epsilon > 0$, a sequence of compact intervals K_n increasing to $\mathbb{R}$, and a sequence $\{\mu_n\}$ from the family such that

$$\mu_n\,(K_n) \leqslant 1 - \epsilon\,, \text{ for all } n\,.$$

Without loss of generality, assume μ_n converges weakly to μ. For any given finite continuty inetrval $(a, b]$ of μ, we have

$$\mu(a, b] = \lim_{n\to\infty} \mu_n(a, b] \leqslant \limsup_{n\to\infty} \mu_n(K_n) \leqslant 1 - \epsilon\,.$$

Since $(a, b]$ is arbitary, $\mu(\mathbb{R}) \leqslant 1 - \epsilon$. □

The following sufficient condition for tightness is often useful.

Theorem 3.11. *If there is a $\phi \geqslant 0$ so that $\phi(x) \to \infty$ as $|x| \to \infty$ and*

$$\sup_n \int \phi(x)\mu_n(dx) < \infty$$

Then $\{\mu_n\}$ is tight.

Proof. Note that

$$\mu([-M, M]^c) \leqslant \frac{\sup_n \int \phi\, d\mu_n}{\inf_{|x|\geqslant M} \phi(x)}$$ □

3.1.4 Examples

An example of weak convergence that we have seen earlier are:

Example 3.7. Let $X_1, X_2, \cdots$ be i.i.d. with distribution F. The Glivenko-Cantelli theorem implies that for almost every ω,

$$F_n(y) = n^{-1} \sum_{m=1}^{n} 1_{(X_m(\omega)\leqslant y)} \to F(y) \text{ for all } y$$

This examples convergence occurred for all y, even though in the second case the distribution function could have discontinuities.

Example 3.8 (Waiting for rare events)**.** Let X_p be the number of trials needed to get a success in a sequence of independent trials with success probability p. Then $\mathbb{P}(X_p \geqslant n) = (1-p)^{n-1}$ for $n = 1, 2, 3, \cdots$ and as $p \to 0$,

$$\mathbb{P}(pX_p > x) \approx \big(1 - \frac{x}{p}\big)^{x/p} \to e^{-x}\,, \text{ for all } x \geqslant 0\,.$$

In words, pX_p converges weakly to an exponential distribution.

Example 3.9 (Birthday problem)**.** Let $X_1, X_2, \cdots$ be independent and uniformly distributed on $\{1, \cdots, N\}$, and let

$$T_N = \min\{n : X_n = X_m \text{ for some } m < n\}$$

Then

$$\mathbb{P}(T_N > n) = \prod_{m=2}^{n} \left(1 - \frac{m-1}{N}\right)$$

When $N = 365$ this is the probability that two people in a group of size n do not have the same birthday (assuming all birthdays are equally likely). Using Exercise 3.1.1, it is easy to see that

$$\mathbb{P}\left(\frac{T_N}{\sqrt{N}} > x\right) \approx \prod_{m=2}^{\sqrt{N}x} \left(1 - \frac{m-1}{N}\right) \to \exp\left(-x^2/2\right) \text{ for all } x \geqslant 0\,.$$

Taking $N = 365$ and noting $22/\sqrt{365} = 1.1515$ and $(1.1515)^2/2 = 0.6630$, this says that This answer is 2 smaller than the true probability 0.524 .

Before giving our next example, we need a simple result

Theorem (Scheffé). *Suppose $f_n \in L^1$, and $f_n \to f$ a.e. (or in measure) as $n \to \infty$. If $\|f_n\|_1 \to \|f\|_1$, then $f_n \to f$ in L^1.*

Thus if μ_n and μ are p.m.'s on $(\mathbb{R}, \mathcal{B})$ with density f_n and f and $f_n \xrightarrow{\text{a.e.}} f$, then $\mu_n \to \mu$ in total variation distance. Obviously, $\mu_n \xrightarrow{w} \mu$.

Example 3.10 (Central order statistic)**.** Put $(2n+1)$ points at random in $(0,1)$, i.e., with locations that are independent and uniformly distributed. Let V_{n+1} be the $(n+1)$ th largest point. It is easy to see that V_{n+1} has density function

$$f_{V_{n+1}}(x) = (2n+1)\binom{2n}{n} x^n (1-x)^n$$

To see this, there are $2n+1$ ways to pick the observation that falls at x, then we have to pick n indices for observations $< x$, which can be done in $\binom{2n}{n}$ ways. Once we have decided on the indices that will land $< x$ and $> x$, the probability the corresponding random variables will do what we want

is $x^n(1-x)^n$, and the probability density that the remaining one will land at x is 1. If you don't like the previous sentence compute the probability $X_1 < x-\epsilon, X_n < x-\epsilon, x-\epsilon < X_{n+1} < x+\epsilon, X_{n+2} > x+\epsilon, \cdots X_{2n+1} > x+\epsilon$ then let $\epsilon \to 0$.

To compute the density function of $Y_n = 2\left(V_{n+1} - 1/2\right)\sqrt{2n}$, we simply change variables $x = 1/2 + y/2\sqrt{2n}, dx = dy/2\sqrt{2n}$ to get

$$\begin{aligned} f_{Y_n}(y) &= (2n+1)\binom{2n}{n}\left(\frac{1}{2}+\frac{y}{2\sqrt{2n}}\right)^n\left(\frac{1}{2}-\frac{y}{2\sqrt{2n}}\right)^n\frac{1}{2\sqrt{2n}} \\ &= \binom{2n}{n}2^{-2n}\cdot\left(1-y^2/2n\right)^n\cdot\frac{2n+1}{2n}\cdot\sqrt{\frac{n}{2}} \end{aligned}$$

Using Stirling's approximation we have

$$f_{Y_n}(y) \to (2\pi)^{-1/2}\mathrm{e}^{-y^2/2}, \text{ as } n \to \infty \tag{3.6}$$

Here and in what follows we write $P\left(Y_n = y\right)$ for the density function of Y_n. Using Scheffé's theorem now, we conclude that Y_n converges weakly to a standard normal distribution. The result in (3.6) is a *local centerl limit theorem.*

EXERCISE

EXERCISE 18. Using Theorem 3.3 (or Corollary 3.2) to show Example 3.4. Using the method of a single probability space, Theorem 3.3, to prove Theorem 3.1.

EXERCISE 19 (Converging together lemma). Suppose $X_n \Rightarrow X$ and $Y_n \Rightarrow c$, where c is a constant then show that (i) $X_n + Y_n \Rightarrow X + c$. and, (ii) $X_n Y_n \Rightarrow cX$.
Hint: See the proof of Example 3.4.

EXERCISE 20. Let $X_n, 1 \le n \le \infty$, be integer valued. Show that $X_n \Rightarrow X_\infty$ if and only if $\mathbb{P}\left(X_n = m\right) \to \mathbb{P}\left(X_\infty = m\right)$ for all m.

EXERCISE 21. Suppose $Y_n \geqslant 0$, $\mathbb{E}Y_n^\alpha \to 1$ and $\mathbb{E}Y_n^\beta \to 1$ for some $0 < \alpha < \beta$. Show that $Y_n \to 1$ in probability.

Hint: Note that $\{\mathbb{E}Y_n^\beta\}$ is uniformly bounded implies that (1) $\{Y_n\}$ is tight by Theorem 3.11 and, (2) $\{Y_n^\gamma\}$ is uniformly integrable, for any $0 < \gamma < \beta$.

3.2 Characteristic functions

This long section is divided into three parts.

- In the first part, we show that the characteristic function

$$\varphi(t) = \mathbb{E}\exp(itX)$$

  determines the distribution μ of X, and we give recipes for computing μ from φ.

- In the second part, we relate weak convergence of distributions to the behavior of the corresponding characteristic functions.

- In the third part, we relate the behavior of $\varphi(t)$ at 0 to the moments of X.

3.2.1 Definition, Inversion Formula

The definition of the characteristic function requires taking the expected value of a complex-valued random variable. If Z is complex valued random variable, we define

$$\mathbb{E}Z = \mathbb{E}(\operatorname{Re} Z) + i\,\mathbb{E}(\operatorname{Im} Z)\,. \tag{3.7}$$

Clearly, Z is integrable iff $|Z|$ is integrable, and in this case we have $|\mathbb{E}Z| \leqslant \mathbb{E}|Z|$.

Definition 3.4. If X is a random variable we define it's **characteristic function (ch.f.)** by

$$\varphi(t) = \mathbb{E}\,\mathrm{e}^{itX} = \mathbb{E}\cos tX + i\,\mathbb{E}\sin tX\,. \tag{3.8}$$

If μ is a p.m. on $(\mathbb{R}, \mathcal{B})$, it's **characteristic function (ch.f.)** is defined by

$$\varphi(t) = \int \mathrm{e}^{itx}\mu(dx) = \int \cos(tx)\mu(dx) + i\int \sin(tx)\mu(dx)\,. \tag{3.9}$$

Proposition 3.12. *All characteristic functions have the following properties:*

(i) $|\varphi(t)| \leqslant 1$ *for all* t *and* $\varphi(0) = 1$.

(ii) $\varphi(-t) = \overline{\varphi(t)}$ *for all* t.

(iii) φ *is uniformly continuous on* $\mathbb{R}$.

(iv) $\mathbb{E}e^{it(aX+b)} = e^{itb}\varphi(at)$, *for each* $a, b \in \mathbb{R}$.

(v) If $\{\varphi_n, n \geqslant 1\}$ *are ch.f.'s,* $\lambda_n \geqslant 0$, $\sum_{n=1}^{\infty} \lambda_n = 1$, *then* $\sum_{n=1}^{\infty} \lambda_n \varphi_n$ *is a ch.f. Briefly, a convex combination of ch.f.'s is a ch.f.*

Proof. (i), (ii), (iv) is clear. To show (iii), note that

$$\begin{aligned} |\varphi(t+h) - \varphi(t)| &= \left| E\left(e^{i(t+h)X} - e^{itX}\right)\right| \\ &\leqslant E\left|e^{i(t+h)X} - e^{itX}\right| = E\left|e^{ihX} - 1\right|, \end{aligned}$$

so uniform convergence follows from the bounded convergence theorem.

For if $\{\mu_n, n \geqslant 1\}$ are the corresponding p.m.'s, then $\sum_{n=1}^{\infty} \lambda_n \mu_n$ is a p.m. whose ch.f. is $\sum_{n=1}^{\infty} \lambda_n f_n$ □

The main reason for introducing characteristic functions is the following:

Theorem 3.13. *If* X_1 *and* X_2 *are independent and have ch.f.'s* φ_1 *and* φ_2 *then* $X_1 + X_2$ *has ch.f.* $\varphi_1(t)\varphi_2(t)$.

Proof.

$$\mathbb{E}e^{it(X_1+X_2)} = E\left(e^{itX_1}e^{itX_2}\right) = \mathbb{E}e^{itX_1}\mathbb{E}e^{itX_2} = \varphi_1(t)\varphi_2(t),$$

since e^{itX_1} and e^{itX_2} are independent. □

The inversion formula

Theorem 3.14 (The inversion formula). *Let $\varphi(t) = \int e^{itx}\mu(dx)$ where μ is a p.m. If $a < b$ then*

$$\lim_{T\to\infty}(2\pi)^{-1}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt = \mu(a,b) + \frac{1}{2}\mu\{a,b\} \tag{3.10}$$

REMARK. The existence of the limit is part of the conclusion. If $\mu = \delta_0$, a point mass at $0, \varphi(t) \equiv 1$. In this case, if $a = -1$ and $b = 1$, the integrand is $(2\sin t)/t$ and the integral does not converge absolutely.

Proof. Let

$$I_T = \int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt = \int_{-T}^{T}\int_{\mathbb{R}}\frac{e^{-ita}-e^{-itb}}{it}e^{itx}\mu(dx)dt$$

The integrand may look bad near $t = 0$ but if we observe that

$$\frac{e^{-ita}-e^{-itb}}{it} = \int_a^b e^{-ity}dy$$

we see that the modulus of the integrand is bounded by $b - a$. since μ is a probability measure and $[-T, T]$ is a finite interval it follows from Fubini's theorem, $\cos(-x) = \cos x$, and $\sin(-x) = -\sin x$ that

$$\begin{aligned} I_T &= \int_{\mathbb{R}}\mu(dx)\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}e^{itx}dt \\ &= \int_{\mathbb{R}}\mu(dx)\int_{-T}^{T}\frac{\sin(t(x-a))}{t} - \frac{\sin(t(x-b))}{t}dt\,. \end{aligned}$$

There appers certain "*Dirichlet integrals*".

Lemma. *(i) For any $x \geqslant 0$,*

$$\int_0^{\pi}\frac{\sin t}{t}\,dt \geqslant \int_0^{x}\frac{\sin t}{t}\,dt \geqslant 0\,. \tag{3.11}$$

(ii)

$$\int_0^{\infty}\frac{\sin x}{x}dx = \frac{\pi}{2}.\ [1] \tag{3.12}$$

[1] This integral is not absolutely convergent, and so the integral is not even defined in the sense of Lebesgue integration, but it is defined in the sense of the generalized Riemann integral

Proof. The inequality (i) is proved by partitioning the interval $[0, \infty)$ with positive multiples of π so as to convert the integral into a series of alternating signs and decreasing moduli.

The integral in (ii) is a standard exercise in contour integration, note that

$$\begin{aligned}\int_0^\infty \frac{\sin x}{x} dx &= \int_0^\infty \sin x dx \left[\int_0^\infty e^{-xu} du\right] \\ &= \int_0^\infty \left[\int_0^\infty e^{-xu} \sin x dx\right] du \\ &= \int_0^\infty \frac{du}{1+u^2} = \frac{\pi}{2}\,.\end{aligned}$$

□

Introducing $R(\theta, T) = \int_{-T}^{T} (\sin \theta t)/t dt$, we can write I_T as

$$I_T = \int \{R(x-a, T) - R(x-b, T)\}\mu(dx) \tag{3.13}$$

As $T \to \infty$, we have $R(\theta, T) \to \pi \operatorname{sgn} \theta$ and

$$R(x-a, T) - R(x-b, T) \to \begin{cases} 2\pi & a < x < b \\ \pi & x = a \text{ or } x = b \\ 0 & x < a \text{ or } x > b \end{cases}$$

Since $|R(\theta, T)| \leqslant 2\int_0^\pi \frac{\sin t}{t}\, dt$, the bounded convergence theorem with (3.13) implies

$$(2\pi)^{-1} I_T \to \mu(a, b) + \frac{1}{2}\mu\{a, b\}$$

proving the desired result. □

REMARK. 逆转公式可以看为某种意义上傅里叶逆变换的推广.

A trivial consequence of the inversion formula are

Corollary 3.15. *X and $-X$ have the same distribution if and only if ϕ is real-valued.*

The inversion formula is simpler when ϕ is integrable, but as the next result shows this only happens when the underlying measure is nice.

Theorem 3.16. *If $\int |\varphi(t)|dt < \infty$ then μ has bounded continuous density*

$$f(y) = \frac{1}{2\pi}\int e^{-ity}\varphi(t)dt \tag{3.14}$$

Proof. As we observed in the proof of the inversion formular,

$$\left|\frac{e^{-ita} - e^{-itb}}{it}\right| = \left|\int_a^b e^{-ity}dy\right| \leqslant |b-a|$$

so the integral in Theorem 3.14 converges absolutely in this case and

$$\mu(a,b) + \frac{1}{2}\mu(\{a,b\}) = \frac{1}{2\pi}\int \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt \leqslant \frac{(b-a)}{2\pi}\int |\varphi(t)|dt\,.$$

The last result implies μ has no point masses and

$$\begin{aligned}\mu(x, x+h) &= \frac{1}{2\pi}\int \frac{e^{-itx} - e^{-it(x+h)}}{it}\varphi(t)dt \\ &= \frac{1}{2\pi}\int \left(\int_x^{x+h} e^{-ity}dy\right)\varphi(t)dt \\ &= \int_x^{x+h}\left(\frac{1}{2\pi}\int e^{-ity}\varphi(t)dt\right)dy\end{aligned}$$

by Fubini's theorem, so the distribution μ has density function

$$f(y) = \frac{1}{2\pi}\int e^{-ity}\varphi(t)dt$$

The dominated convergence theorem implies f is continuous and the proof is complete. □

We should emphasize that the converse of Theorem 3.16 is false, a counterexample is given in Example 3.16. But now, let us see a application of Theorem 3.16 first.

Example 3.11. Suppose $X_1, \ldots, X_n$ are independent and uniformly distributed on $(-1, 1)$ then for $n \geqslant 2$, $X_1 + \cdots + X_n$ has density

$$f(x) = \frac{1}{\pi}\int_0^\infty \left(\frac{\sin t}{t}\right)^n \cos tx\, dt$$

Although it is not obvious from the formula, f is a polynomial in each interval $(k, k+1)$, $k \in \mathbf{Z}$ and vanishes on $[-n, n]^c$

Examples for ch.f. The next order of business is to give some examples.

Example 3.12 (Coin flips.)**.** If $\mathbb{P}(X=1)=\mathbb{P}(X=-1)=1/2$ then

$$\mathbb{E}e^{itX}=\frac{e^{it}+e^{-it}}{2}=\cos t\,. \tag{3.15}$$

Example 3.13 (Poisson distribution.)**.** The ch.f. of Poisson(λ) is

$$\varphi(t)=\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k e^{itk}}{k!}=\exp\left(\lambda\left(e^{it}-1\right)\right) \tag{3.16}$$

Example 3.14 (Uniform distribution)**.** The ch.f. of uniform(a,b) is

$$\varphi(t)=\frac{e^{itb}-e^{ita}}{it(b-a)}\,. \tag{3.17}$$

In the special case $a=-c, b=c$, the ch.f. is

$$\varphi(t)=\frac{\sin ct}{it(b-a)ct}\,. \tag{3.18}$$

Once you recall that $\int_a^b e^{\lambda x}dx=\left(e^{\lambda b}-e^{\lambda a}\right)/\lambda$ holds for complex λ, this is immediate.

Example 3.15 (Normal distribution)**.** The ch.f. of $N(\mu,\sigma^2)$ is

$$\varphi(t)=\exp\left(i\mu t-\frac{\sigma^2}{2}t^2\right)\,. \tag{3.19}$$

To see this, it's suffices to compute the ch.f. of $N(0,1)$,

$$\int e^{itx}(2\pi)^{-1/2}e^{-x^2/2}dx=e^{-t^2/2}\int(2\pi)^{-1/2}e^{-(x-it)^2/2}dx$$

The integral is 1 since the integrand is the normal density with mean it and variance 1. [2]

By using the ch.f.'s, we can see that if $X_i, i=1,2$ are independent and have normal distributions with mean μ_i and variance σ_i^2, then X_1+X_2 has a normal distribution with mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$.

[2] 或者可用复积分严格的计算出为 1

Example 3.16 (Exponential distribution)**.** The ch.f. of exponential(λ) is

$$\varphi(t) = \frac{\lambda}{\lambda - it} \tag{3.20}$$

To see this, it suffices to compute the ch.f. of exponential(1). Integrating gives

$$\int_0^\infty e^{itx} e^{-x} dx = \left. \frac{e^{(it-1)x}}{it-1} \right|_0^\infty = \frac{1}{1-it}\,.$$

Besides, we point that $\int |\varphi(t)| dt = \infty$, although the exponential distribution has a density.

Example 3.17 (Gamma distribution)**.** The density of Gamma(α, λ) is given by

$$\lambda \mathrm{e}^{-\lambda x} \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} 1_{\{x>0\}}\,,$$

and the ch.f. of Gamma(α, λ) is

$$\varphi(t) = \left(\frac{\lambda}{\lambda - it} \right)^\alpha \tag{3.21}$$

Example 3.18 (Bilateral exponential)**.** The density of bilateral exponential distribution is given by

$$\frac{1}{2} e^{-|x|}\,,\ x \in (-\infty, \infty)\,.$$

Then it has ch.f.

$$\varphi(t) = \frac{1}{(1+t^2)}\,.$$

To see this,

$$\frac{1}{2} \int e^{itx} e^{-|x|} dx = \frac{1}{2(1-it)} + \frac{1}{2(1+it)} = \frac{1}{(1+t^2)}$$

Example 3.19 (Cauchy distribution)**.** The density of bilateral exponential distribution is given by

$$\frac{1}{\pi(1+x^2)}\,,\ x \in (-\infty, \infty)\,.$$

Then it has ch.f.

$$\phi(t) = \mathrm{e}^{-|t|}\,. \tag{3.22}$$

To see this, note that by Example 3.18 and Theorem 3.16 we have

$$\frac{1}{2\pi}\int \frac{1}{1+s^2}e^{-isy}ds = \frac{1}{2}e^{-|y|}\,.$$

Now let $s = x, y = -t$ and multiply each side by (3.22).

Besides, suppose $X_1, X_2, \cdots$ are independent and have the Cauchy distribution, then

$$\frac{X_1 + \cdots + X_n}{n}$$

has the same distribution as X_1, thus SLLN is not valid in this case.

Futher reading* We will give a necessary and sufficient condition that a complex-valued function ϕ is a ch.f. First, we can show that if $\varphi(t)$ is the ch.f. of a random variable X, then

(i) $\varphi(0) = 1$.

(ii) $\varphi(t)$ is a continuous function in $t \in R_1$

(iii) $\varphi(t)$ is positively defined function, i.e., a quadratic form

$$\sum_{1\leqslant j,l\leqslant n} c_j\bar{c}_l\varphi\left(t_j - t_l\right) \geqslant 0$$

for any complex numbers $c_1, \ldots, c_n$ and real $t_1, \ldots, t_n$ and $n \geqslant 1$.

Theorem (Bochner). *A complex-valued function $\phi(t)$ defined on a real line is a characteristic function of some random variable X, if and only if it possesses the properties (i)-(iii) formulated above.*

EXERCISE

EXERCISE 22. (i) Imitate the proof of Theorem 3.14 to show that, for any $a \in \mathbb{R}$,

$$\mu\{a\} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita}\varphi(t)t\,. \tag{3.23}$$

(ii) If $\mathbb{P}(X \in h\mathbb{Z}) = 1$ where $h > 0$ then its ch.f. has period $\frac{2\pi}{h}$, and

$$\mathbb{P}(X = x) = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} e^{-itx}\varphi(t)dt \quad \text{for } x \in h\mathbf{Z}$$

(iii) So if $\mathbb{P}(X \in b + h\mathbb{Z}) = 1$, the inversion formula in (ii) is valid for $x \in b + h\mathbb{Z}$.

EXERCISE 23. Suppose X and Y are independent and have ch.f. φ and distribution μ. Apply Exercise 3.3 .2 to $X - Y$ and use Exercise 2.1 .5 to get

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |\varphi(t)|^2 dt = \mathbb{P}(X - Y = 0) = \sum_x \mu(\{x\})^2$$

REMARK. The last result implies that if $\varphi(t) \to 0$ as $t \to \infty, \mu$ has no point masses. The Riemann-Lebesgue Lemma shows that if μ has a density, $\varphi(t) \to 0$ as $t \to \infty$.

EXERCISE 24. (i) Suppose that the family of measures $\{\mu_i, i \in I\}$ is tight, show that their ch.f.'s φ_i are equicontinuous, i.e., if $\epsilon > 0$ we can pick $\delta > 0$ so that if $|h| < \delta$ then $|\varphi_i(t + h) - \varphi_i(t)| < \epsilon$.

(ii) Suppose $\mu_n \Rightarrow \mu$. Use equicontinuity to conclude that the ch.f.'s $\varphi_n \to \varphi$ uniformly on compact sets.

3.2.2 Lévy's continuity theorem

Our next step toward the central limit theorem is to relate convergence of characteristic functions to weak convergence.

Theorem 3.17. *Let μ, μ_n be probability measures with ch.f. φ and φ_n. If $\mu_n \Rightarrow \mu$ then*

$$\varphi_n(t) \to \varphi(t)\,, \text{ for all } t\,.$$

Proof. e^{itx} is bounded and continuous so if $\mu_n \Rightarrow \mu$ then Theorem 3.1 implies $\varphi_n(t) \to \varphi(t)$. □

Theorem 3.18 (Lévy's continuity theorem). *Let μ_n be probability measures with ch.f. φ_n. If*

$$\varphi_n(t) \to \varphi(t)\,, \text{ for all } t$$

and, φ is continuous at 0. Then there exists a probability measure μ with characteristic function φ, so that

$$\mu_n \Rightarrow \mu\,.$$

Proof. **Step 1**. We need the following lemma, which shows that the smoothness of the characteristic function at 0 is related to the decay of the measure μ at ∞.

Lemma 3.19. *φ is the ch.f. of the distribution μ, then for any $\delta > 0$,*

$$\frac{1}{2\delta}\int_{-\delta}^{\delta} \varphi(t)\,dt = \int_{\mathbb{R}} \frac{\sin \delta x}{\delta x}\,\mu(dx)\,. \tag{3.24}$$

Proof. To see this, note that

$$\begin{aligned}\frac{1}{2\delta}\int_{-\delta}^{\delta} \varphi(t)\,dt &= \frac{1}{2\delta}\int_{-\delta}^{\delta} dt \int_{\mathbb{R}} e^{itx}\,\mu(dx) = \int_{\mathbb{R}} \mu(dx)\frac{1}{2\delta}\int_{-\delta}^{\delta} e^{itx}\,dt \\ &= \int_{\mathbb{R}} \mu(dx)\int_{-\delta}^{\delta} \frac{\cos tx}{2\delta}\,dt = \int_{\mathbb{R}} \frac{\sin \delta x}{\delta x}\,\mu(dx)\,.\end{aligned}$$

□

Step 2. By using the preceeding lemma, we can show that $\{\mu_n\}$ is tight. For given $M > 0$,

$$\begin{aligned}\left|\frac{1}{2\delta}\int_{-\delta}^{\delta} \varphi_n(t)dt\right| &\leqslant \int_{\mathbb{R}} \left|\frac{\sin \delta x}{\delta x}\right| \mu_n(dx) \\ &\leqslant \mu_n[-M, M] + \frac{1}{\delta M}\,.\end{aligned}$$

Letting $n \to \infty$, by domainted convergence theorem we have

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi_n(t)dt \to \frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi(t)dt\,.$$

Thus

$$\liminf_{n\to\infty}\mu_n[-M,M] + \frac{1}{\delta M} \geqslant \frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi(t)dt$$

Since φ is continuous at 0, for any $\epsilon > 0$, there exists some $\delta > 0$, depending on ϵ, so that

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi(t)dt > \phi(0) - \epsilon = 1 - \epsilon\,.$$

So there exists M, depending on ϵ,

$$\liminf_{n\to\infty}\mu_n[-M,M] \geqslant 1 - 2\epsilon$$

Thus $\{\mu_n\}$ is tight.

Step 3. To complete the proof now, we observe that if

$$\mu_{n(k)} \Rightarrow \mu\,,$$

By Theorem 3.17, we have that μ has ch.f. φ. Thus every subsequence has a further subsequence that converges to μ. So $\{\mu_n\}$ weakly converges to μ. □

REMARK. Levy 连续定理的关键在于证明 $\{\mu_n\}$ 的胎紧性, 而 $\{\mu_n\}$ 的胎紧性证明的关键是 *Lemma 3.19*, 它说明了特征函数在原点附近的取值已经蕴含了测度在无穷远处的分布的信息.

Example 3.20. Suppose that $X_n \Rightarrow X$ and X_n has a Gaussian distribution with parameters 0 and σ_n^2. Using Theorem 3.17, we can find a $\sigma \geqslant 0$ such atht $\sigma_n^2 \to \sigma^2$, by Lévy's continuity theorem, X is Gaussian distribution with parameters 0 and σ^2.

EXERCISE

EXERCISE 25. If Y_n are r.v.'s with ch.f.'s φ_n then $Y_n \Rightarrow 0$ if and only if there is a $\delta > 0$ so that $\varphi_n(t) \to 1$ for $|t| \leqslant \delta$.
Hint : Use Lemma 3.19 to deduce that $\{Y_n\}$ is tight, and $Y_n \Rightarrow 0$.

EXERCISE 26. Let $X_1, X_2, \cdots$ be independent. If $S_n = X_1 + \cdots + X_n$ converges in distribution then it converges in probability (and hence a.s. by Exercise 15).
Hint: The last exercise implies that if $m, n \to \infty$ then $S_m - S_n \to 0$ in probability. Now use Exercise 25.

3.2.3 Moments and Derivatives

We have pointed that the smoothness of the characteristic function at 0 is related to the decay of the measure μ at ∞, in the proof of Theorem 3.18. In fact, we have a more detailed description about this.

Theorem 3.20. *If $\int |x|^n \mu(dx) < \infty$ then its characteristic function φ has a continuous derivative of order n given by*

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(dx)\,. \tag{3.25}$$

Proof. We prove 3.25 by intduction. Assume $\varphi^{(k)}$ is continuous on $\mathbb{R}$ and

$$\varphi^{(k)}(t) = \int (ix)^k e^{itx} \mu(dx)\,.$$

where $0 \leqslant k < n$. For any given $h \in \mathbb{R}/\{0\}$,

$$\frac{\varphi^{(k)}(t+h) - \varphi^{(k)}(t)}{h} = \int (ix)^k e^{itx} \frac{e^{ihx} - 1}{h} \mu(dx)\,.$$

By Lagrange mean-value theorem, there exists some θ_1, $\theta_2 \in (0, 1)$, depending on h, so that

$$\frac{e^{ihx} - 1}{h} = ix[\cos(\theta_1 hx) + i \sin(\theta_2 hx)]$$

Thus

$$\left|(ix)^k e^{itx}\frac{e^{ihx}-1}{h}\right| \leqslant 2|x|^{k+1}$$

By Lebesgue domainted convergence theorem,

$$\varphi^{(k+1)}(t) = \lim_{h\to 0}\frac{\varphi^{(k)}(t+h)-\varphi^{(k)}(t)}{h} = \int (ix)^{k+1}e^{itx}\mu(dx)\,.$$

Clearly, $\phi(n)$ is continuous. Thus we get the desired result. □

The result in Theorem 3.20 shows that if $\mathbb{E}|X|^n < \infty$, then its characteristic function is n times differentiable at 0, and $\varphi^n(0) = \mathbb{E}(iX)^n$. Expanding φ in a Taylor series about 0 leads to

$$\varphi(t) = \sum_{m=0}^{n}\frac{\mathbb{E}(itX)^m}{m!} + o\left(t^n\right)$$

Example 3.21. Usesing Theorem 3.20 and the series expansion for $e^{-t^2/2}$, it's easy to show that the standard normal distribution has

$$\mathbb{E}X^{2n} = \frac{(2n)!}{2^n n!} = (2n-1)(2n-3)\cdots 3\cdot 1 =: (2n-1)!!$$

Lemma 3.21.

$$\left|e^{ix} - \sum_{m=0}^{n}\frac{(ix)^m}{m!}\right| \leqslant \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\right)$$

REMARK. The first term on the right is the usual order of magnitude we expect in the correction term. *The second is better for large $|x|$ and will help us prove the central limit theorem without assuming finite third moments.*

Proof. Integrating by parts. □

Taking expected values gives

$$\left|\mathbb{E}e^{itX} - \sum_{m=0}^{n}\mathbb{E}\frac{(itX)^m}{m!}\right| \leqslant \mathbb{E}\min\left(2|tX|^n, |tX|^{n+1},\right) \tag{3.26}$$

where in the second step we have dropped the denominators to make the bound simpler. In the next section, the following special case will be useful.

Theorem 3.22. *If* $\mathbb{E}|X|^2 < \infty$ *then*

$$\varphi(t) = 1 + i\,\mathbb{E}X\,t - \frac{\mathbb{E}\left(X^2\right)}{2}t^2 + \varepsilon\,(t)\ , \tag{3.27}$$

where $\varepsilon(t) = o(t^2)$ *as* $t \to 0$.

Proof. The error term $\varepsilon(t)$ is $\leqslant t^2\mathbb{E}\left[2|X|^2 \wedge |t||X|^3\right]$. The variable in parentheses is smaller than $2|X|^2$ and converges to 0 as $t \to 0$, so the desired conclusion follows from the dominated convergence theorem. □

Remark. The point of the estimate in (3.27) which involves the minimum of two terms rather than just the first one which would result from a naive application of Taylor series, is that we get the conclusion in (3.27) under the assumption $\mathbb{E}|X|^2 < \infty$, i.e., we do not have to assume $\mathbb{E}|X|^3 < \infty$.

EXERCISE

Exercise 27. Let $X_1, X_2, \cdots$ be i.i.d. with characteristic function φ.

(i) If $\varphi'(0) = ia$ and $S_n = X_1 + \cdots + X_n$ then $S_n/n \to a$ in probability.

(ii) If $S_n/n \to a$ in probability then $\varphi(t/n)^n \to e^{iat}$ as $n \to \infty$ through the integers.

(iii) Use (ii) and the uniform continuity of ϕ to show that $(\varphi(h) - 1)/h \to -ia$ as $h \to 0$ through the positive reals.

Thus the weak law holds if and only if $\varphi'(0)$ exists. This result is due to E.J.G. Pitman (1956), with a little help from John Walsh who pointed out that we should prove (iii).

The next exercise shows that the existence of second derivatives implies the existence of second moments.

Exercise 28. If

$$\limsup_{h\downarrow 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} > -\infty,$$

then $\mathbb{E}|X|^2 < \infty$.
Sketch: $\left(e^{ihx} - 2 + e^{-ihx}\right)/h^2 = -2(1-\cos hx)/h^2 \leqslant 0$ and $2(1-\cos hx)/h^2 \to x^2$ as $h \to 0$ so Fatou's lemma and Fubini's theorem imply

$$\begin{aligned}\int x^2 \mu(dx) &\leqslant 2 \lim_{h\to 0} \frac{1-\cos hx}{h^2} \mu(dx) \\ &= -\limsup_{h\to 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} < \infty\,,\end{aligned}$$

which proves the desired result.

Exercise 29. Show that if $\lim_{t1 0(\varphi(t)-1)/t^2} = c > -\infty$ then $\mathbb{E}X = 0$ and $E|X|^2 = -2c <$ œ. In particular, if $\varphi(t) = 1 + o\left(t^2\right)$ then $\varphi(t) \equiv 1$.

3.3 Central limit theorems

We are now ready for the main bussiness of the chapter. We will first prove the central limit theorem for

3.3.1 i.i.d. sequences

Theorem 3.23. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = \mu, \mathrm{Var}\,(X_i) = \sigma^2 \in (0, \infty)$. If $S_n = X_1 + \cdots + X_n$ then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1)\,. \tag{3.28}$$

Proof. By considering $X_i' = X_i - \mu$, it suffices to prove the result when $\mu = 0$. By Theorem 3.22

$$\varphi_{X_1}(t) = \mathbb{E}\, e^{itX_1} = 1 - \frac{\sigma^2 t^2}{2} + \varepsilon(t)\,,$$

where $\varepsilon(t) = o(t^2)$ as $t \to 0$. So

$$\varphi_n(t) \coloneqq \mathbb{E} \exp\left(it\frac{S_n}{\sigma\sqrt{n}}\right) = \left(1 - \frac{t^2}{2n} + \varepsilon\big(\frac{t}{\sigma\sqrt{n}}\big)\right)^n$$

Recall that in complex analysis we have learned that, for any complex number z with $|z| \leqslant 1/2$

$$\log(1+z) = z + \theta|z|^2\,, \quad [3]$$

where $\theta = \theta(z)$ with $|\theta| \leqslant 1$. Therefore,

$$\begin{aligned}\log\varphi_n(t) &= n\log\left(1 - \frac{t^2}{2n} + \varepsilon\big(\frac{t}{\sigma\sqrt{n}}\big)\right)\\ &= -\frac{t^2}{2} + n\varepsilon\big(\frac{t}{\sigma\sqrt{n}}\big) + \theta\left|\frac{t^2}{2n} - \varepsilon\big(\frac{t}{\sigma\sqrt{n}}\big)\right|^2\end{aligned}$$

For fixed t, letting $n \to \infty$, since $\varepsilon(t) = o(t^2)$ as $t \to 0$, $n\varepsilon\big(\frac{t}{\sigma\sqrt{n}}\big) \to 0$, thus

$$\log\varphi_n(t) \to -\frac{t^2}{2}\,, \text{ as } n \to \infty\,,$$

which with Theorem 3.18 and Example 3.15 completes the proof. □

[3] In this text, log is the *principal value* of the logarithm ($\log z = \log|z| + i\arg z, -\pi < \arg z \leqslant \pi$).

Rates of convergence* We give a result about the rates of convergence in CLT, the proof can be found in *Probability : Theory and examples* by Durret.

Theorem. *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = 0, \mathbb{E}X_i^2 = \sigma^2$, and $\mathbb{E}\,|X_i|^3 = \rho < \infty$. If $F_n(x)$ is the distribution of $\frac{S_n}{\sigma\sqrt{n}}$ and then*

$$|F_n(x) - \Phi(x)| \leqslant 3\frac{\rho}{\sigma^3\sqrt{n}}\,, \text{ for each } n, x \tag{3.29}$$

Applications To get a feel for what the central limit theorem says, we will look at some concrete cases.

Example 3.22 (Roulette(轮盘赌博))**.** A roulette wheel has slots numbered $1 - 36$ (18 red and 18 black) and two slots numbered 0 and 00 that are painted green. Players can bet \$1 that the ball will land in a red (or black) slot and win \$1 if it does. If we let X_i be the winnings on the i th play then $X_1, X_2, \cdots$ are i.i.d. with $\mathbb{P}(X_i = 1) = 18/38$ and $\mathbb{P}(X_i = -1) = 20/38$.

$$\mathbb{E}X_i = -1/19 \quad \text{and} \quad \mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 1 - (1/19)^2 = 0.9972$$

We are interested in

$$\mathbb{P}(S_n \geqslant 0) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \geqslant \frac{-n\mu}{\sigma\sqrt{n}}\right)$$

Taking $n = 361 = 19^2$ and replacing σ by 1 to keep computations simple,

$$\frac{-n\mu}{\sigma\sqrt{n}} = \frac{361 \cdot (1/19)}{\sqrt{361}} = 1$$

So the central limit theorem and our table of the normal distribution in the back of the book tells us that

$$\mathbb{P}(S_n \geqslant 0) \approx \mathbb{P}(N(0,1) \geqslant 1) = 1 - 0.8413 = 0.1587$$

In words, after 361 spins of the roulette wheel the casino will have won \$19 of your money on the average, but there is a probability of about 0.16 that you will be ahead.

Example 3.23 (Coin flips)**.** Let $X_1, X_2, \cdots$ be i.i.d. with

$$\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = 1/2\,.$$

If $X_i = 1$ indicates that a heads occured on the i th toss then $S_n = X_1 + \cdots + X_n$ is the total number of heads at time n.

$$\mathbb{E}X_i = 1/2 \quad \text{and} \quad \mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 1/2 - 1/4 = 1/4$$

So the central limit theorem tells us

$$\frac{S_n - n/2}{\sqrt{n/4}} \Rightarrow N(0,1).$$

Our table of the normal distribution tells us that

$$\mathbb{P}(N(0,1) > 2) = 1 - 0.9773 = 0.0227$$

so $\mathbb{P}(|N(0,1)| \leqslant 2) = 1 - 2(0.0227) = 0.9546$, or plugging into the central limit theorem

$$0.95 \approx \mathbb{P}\left(\frac{S_n - n/2}{\sqrt{n/4}} \in [-2,2]\right) = \mathbb{P}\left(S_n - n/2 \in [-\sqrt{n}, \sqrt{n}]\right)$$

Taking $n = 10,000$ this says that 95% of the time the number of heads will be between 4900 and 5100.

Example 3.24 (Normal approximation to the binomial)**.** Let $X_1, X_2, \cdots$ and S_n be as in the previous example. To estimate $\mathbb{P}(S_{16} = 8)$ using the central limit theorem, we regard 8 as the interval $[7.5, 8.5]$. since $\mu = 1/2$, and $\sigma\sqrt{n} = 2$ for $n = 16$

$$\begin{aligned}\mathbb{P}(|S_{16} - 8| \leqslant 0.5) &= \mathbb{P}\left(\frac{|S_n - n\mu|}{\sigma\sqrt{n}} \leqslant 0.25\right) \\ &\approx \mathbb{P}(|N(0,1)| \leqslant 0.25) = 2(0.5987 - 0.5) = 0.1974\end{aligned}$$

Even though n is small, this agrees well with the exact probability

$$\binom{16}{8} 2^{-16} = \frac{13 \cdot 11 \cdot 10 \cdot 9}{65,536} = 0.1964$$

The computations above motivate the histogram correction, which is important in using the normal approximation for small n. For example, if we are going to approximate $\mathbb{P}(S_{16} \leqslant 11)$, then we regard this probability as $\mathbb{P}(S_{16} \leqslant 11.5)$. One obvious reason for doing this is to get the same answer if we regard $\mathbb{P}(S_{16} \leqslant 11) = 1 - \mathbb{P}(S_{16} \geqslant 12)$.

Example 3.25 (Normal approximation to the Poisson)**.** Let Z_λ have a Poisson distribution with mean λ. If $X_1, X_2, \cdots$ are independent and have Poisson distributions with mean 1, then $S_n = X_1 + \cdots + X_n$ has a Poisson distribution with mean n. since $\operatorname{Var}(X_i) = 1$, the central limit theorem implies:

$$\frac{S_n - n\mu}{\sqrt{n}} \Rightarrow N(0,1), \text{ as } n \to \infty$$

To deal with values of λ that are not integers, let N_1, N_2, N_3 be independent Poisson with means $[\lambda], \lambda - [\lambda]$, and $[\lambda] + 1 - \lambda$. If we let $S_{[\lambda]} = N_1, Z_\lambda = N_1 + N_2$ and $S_{[\lambda]+1} = N_1 + N_2 + N_3$ then $S_{[\lambda]} \leqslant Z_\lambda \leqslant S_{[\lambda]+1}$ and using the limit theorem for the S_n it follows that

$$\frac{Z_\lambda - \lambda}{\sqrt{\lambda}} \Rightarrow N(0,1), \text{ as } \lambda \to \infty.$$

A counterexample* We pointed that pairwise independence is good enough for the strong law of large numbers (see Theorem 2.10). But it is not good enough for the central limit theorem. Here is a counterexample.

Example 3.26. Let $\xi_1, \xi_2, \cdots$ be i.i.d. with $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = 1/2$. We will arrange things so that for $n \geqslant 1$,

$$S_{2^n} = \xi_1 (1 + \xi_2) \cdots (1 + \xi_{n+1}) = \begin{cases} \pm 2^n & \text{with prob } 2^{-n-1} \\ 0 & \text{with prob } 1 - 2^{-n} \end{cases}$$

Clearly, $S_{2^n}/\sqrt{2^n}$ doesn't weakly converges to $N(0,1)$. To do this, note that

$$\begin{aligned} S_{2^{n+1}} - S_{2^n} &= X_{2^n+1} + \cdots + X_{2^n+2^n} \\ &= \xi_1 (1 + \xi_2) \cdots (1 + \xi_{n+1}) \xi_{n+2} = S_{2^n} \xi_{n+2} \\ &= (X_1 + \cdots + X_{2^n}) \xi_{n+2} \end{aligned}$$

Thus, we let $X_1 = \xi_1, X_2 = \xi_1\xi_2$, and

$$X_{2^n+j} = X_j\, \xi_{n+2}\,,$$

where $n \geqslant 1$ and $j \in \{1, \cdots, 2^n\}$. Then each X_m is a product of a different set of ξ_j 's so they are pairwise independent. In fact, it suffices to show ξ_1 is independent of $\xi_1\xi_2$, and $\xi_1\xi_2$ is independent of $\xi_1\xi_3$, which is clear.

EXERCISE

EXERCISE 30. Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = 0$, $\mathrm{Var}\,(X_i) \in (0, \infty)$, and let $S_n = X_1 + \cdots + X_n$.

(i) Use the c.l.t. and Kolmogorov's zero-one law to conclude that

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} = \infty\,, \text{ a.s.} \tag{3.30}$$

(ii) Use an argument by contradiction to show that $S_n/\sqrt{n}$ does not converge in probability.

EXERCISE 31. Let $X_1, X_2, \cdots$ be i.i.d. with $X_i \geqslant 0, \mathbb{E}X_i = 1$, and $\mathrm{Var}\,(X_i) = \sigma^2 \in (0, \infty)$. Show that

$$\sqrt{S_n} - \sqrt{n} \Rightarrow \frac{\sigma}{2} N(0,1)\,.$$

EXERCISE 32 (Self-normalized sums). Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = \sigma^2 \in (0, \infty)$. Then

$$\sum_{m=1}^{n} X_m \Big/ \left(\sum_{m=1}^{n} X_m^2\right)^{1/2} \Rightarrow N(0,1)\,.$$

EXERCISE 33 (Random index central limit theorem). Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = \sigma^2 \in (0, \infty)$, and let $S_n = X_1 + \cdots + X_n$. Let N_n be a sequence of nonnegative integer-valued random variables and a_n a sequence of integers with $a_n \to \infty$ and $N_n/a_n \to 1$ in probability. Show that

$$\frac{S_{N_n}}{\sigma\sqrt{a_n}} \Rightarrow N(0,1)\,.$$

Hint : Use Kolmogorov's maximal inequality to conclude that if $Y_n = S_{N_n}/\sigma\sqrt{a_n}$ and $Z_n = S_{a_n}/\sigma\sqrt{a_n}$, then $Y_n - Z_n$ converges to 0 in probability.

EXERCISE* (A central limit theorem in renewal theory). Let $Y_1, Y_2, \cdots$ be i.i.d. positive random variables with $\mathbb{E}Y_i = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2 \in (0, \infty)$. Let $S_n = Y_1 + \cdots + Y_n$ and $N_t = \sup\{m : S_m \leqslant t\}$. Apply the previous exercise to $X_i = Y_i - \mu$ to prove that as $t \to \infty$

$$(\mu N_t - t) / \left(\sigma^2 t/\mu\right)^{1/2} \Rightarrow \chi$$

3.3.2 Triangular Arrays

Our next step is to generalize the central limit theorem for triangular Arrays. The theorem says that a sum of a large number of small independent effects has approximately a normal distribution.

Theorem 3.24 (The Lindeberg-Feller theorem). *For each n, $X_{n,m}$, $1 \leqslant m \leqslant n$, are independent random variables with $\mathbb{E}X_{n,m} = 0$. Suppose*

(i) $\sum_{m=1}^{n} \mathbb{E}X_{n,m}^2 \to \sigma^2 > 0$, *and*

(ii) For each $\epsilon > 0$,

$$\sum_{m=1}^{n} \mathbb{E}\left|X_{n,m}\right|^2 1_{\{|X_{n,m}|>\epsilon\}} \to 0 \quad \text{as } n \to \infty\,. \tag{Lindeberg's Condition}$$

Let $S_n = X_{n,1} + \cdots + X_{n,n}$, then

$$S_n \Rightarrow N(0, \sigma^2)\,. \tag{3.31}$$

Proof. Let $\varphi_{n,m}(t) = \mathbb{E}e^{itX_{n,m}}$, and $\sigma_{n,m}^2 = \mathbb{E}X_{n,m}^2$. By Theorem 3.18, it suffices to show that

$$\prod_{m=1}^{n} \varphi_{n,m}(t) \to \exp\left(-\frac{\sigma^2}{2}t^2\right)$$

By Theorem 3.22, $\varphi_{n,m}(t) = 1 - \frac{\sigma_{n,m}^2}{2}t^2 + \varepsilon_{n,m}(t)$, and

$$|\varepsilon_{n,m}(t)| \leqslant t^2\,\mathbb{E}\,[\,2|X_{n,m}|^2 \wedge |t|\,|X_{n,m}|^3\,]\,.$$

Note that

$$\begin{aligned}\sum_{m=1}^{n}\log\varphi_{n,m}(t) &= \sum_{m=1}^{n}\log\left(1-\frac{\sigma_{n,m}^2}{2}t^2+\varepsilon_{n,m}(t)\right)\\ &= -\sum_{m=1}^{n}\frac{\sigma_{n,m}^2}{2}t^2+\sum_{m=1}^{n}\varepsilon_{n,m}(t)+\sum_{m=1}^{n}\theta_{n,m}\left|\frac{\sigma_{n,m}^2}{2}t^2-\varepsilon_{n,m}(t)\right|^2\,,\end{aligned}$$

The first term on the left-hand side, is convergent to $-(\sigma^2/2)t^2$ as $n\to\infty$. Since $|\theta_{n,m}|\leqslant 1$, so it suffices to show that , for any given t,

$$\sum_{m=1}^{n}|\varepsilon_{n,m}(t)|\to 0 \quad\text{and,}\quad \sum_{m=1}^{n}\left|\frac{\sigma_{n,m}^2}{2}t^2-\varepsilon_{n,m}(t)\right|^2\to 0\,.$$

First term. Take $0<\epsilon<\frac{2}{|t|}$, then

$$\begin{aligned}|\varepsilon_{n,m}(t)| &\leqslant t^2\,\mathbb{E}\,[\,2|X_{n,m}|^2\wedge|t|\,|X_{n,m}|^3\,]\\ &\leqslant |t|^3\,\mathbb{E}|X_{n,m}|^3 1_{\{|X_{n,m}|\leqslant\epsilon\}}+2t^2\mathbb{E}\,|X_{n,m}|^2\,1_{\{|X_{n,m}|>\epsilon\}}\\ &\leqslant \epsilon|t|^3\sigma_{n,m}^2+2t^2\mathbb{E}\,|X_{n,m}|^2\,1_{\{|X_{n,m}|>\epsilon\}}\end{aligned}$$

Hence

$$\sum_{m=1}^{n}|\varepsilon_{n,m}(t)|\leqslant\epsilon|t|^3\sum_{m=1}^{n}\sigma_{n,m}^2+2t^2\sum_{m=1}^{n}\mathbb{E}\,|X_{n,m}|^2\,1_{\{|X_{n,m}|>\epsilon\}}$$

Clearly, by (i) and Lindeberg's condition, we have

$$\sum_{m=1}^{n}|\varepsilon_{n,m}(t)|\to 0\,.$$

Second term. Note that

$$\sum_{m=1}^{n}\left|\frac{\sigma_{n,m}^2}{2}t^2-\varepsilon_{n,m}(t)\right|^2\leqslant\max_{1\leqslant m\leqslant n}\left\{\sigma_{n,m}^2t^2+|\varepsilon_{n,m}(t)|\right\}\sum_{m=1}^{n}\left(\frac{\sigma_{n,m}^2}{2}t^2+|\varepsilon_{n,m}(t)|\right)\,,$$

and for any $\delta > 0$

$$\max_{1\leqslant m\leqslant n} \sigma_{n,m}^2 \leqslant \delta^2 + \max_{1\leqslant m\leqslant n} \mathbb{E}\left(|X_{n,m}|^2 1_{\{|X_{n,m}|>\delta\}}\right),$$

We have

$$\max_{1\leqslant m\leqslant n} \sigma_{n,m}^2 \to 0\,.$$

Thus

$$\sum_{m=1}^{n} \left|\frac{\sigma_{n,m}^2}{2}t^2 - \varepsilon_{n,m}(t)\right|^2 \to 0\,. \qquad \square$$

REMARK. 这里我们的三角阵第 n 行恰有 n 个相互独立的随机变量, 实际上有 k_n 个也可以, 其中 $k_n \to \infty$. 换言之, 条件改为 $\sum m = 1^{k_n} X_{n,m}^2 \to \sigma^2$ 和

$$\sum_{m=1}^{k_n} \mathbb{E}\,|X_{n,m}|^2 1_{\{|X_{n,m}|>\epsilon\}} \to 0$$

仍然有 $S_n = X_{n,1} + \cdots X_{n,k_n} \Rightarrow N(0,\sigma^2)$, 证明无须改变.

A useful corollary is the CLT for thue sum of independent (but not identically distributed) random variables.

Corollary 3.25 (CLT for independent r.v.'s). *$\{X_n\}$ is a sequence of independent r.v.'s. Let $S_n = X_1 + \cdots + X_n$. If*

(i) $\{X_n\}$ is uniformly bounded, that is , there exists some $M > 0$ so that almost surely, $|X_n| \leqslant M$ for all n, and

(ii) $\sum_{n=1}^{\infty} \mathrm{Var}\,(X_n) = \infty$,

then

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}\,(S_n)}} \Rightarrow N(0,1)\,.$$

Proof. Given n, define

$$\xi_{n,m} = \frac{X_m - \mathbb{E}X_m}{\sqrt{\mathrm{Var}\,(S_n)}},\ \text{for } 1 \leqslant m \leqslant n\,.$$

Clearly, $\mathbb{E}\xi_{n,m} = 0$, $\sum_{m=1}^{n} \mathbb{E}\xi_{n,m}^2 = 1$. Note that $|\xi_{n,m}| \leqslant \frac{2M}{\sqrt{\mathrm{Var}(S_n)}}$, thus given any $\epsilon > 0$, for sufficiently large n we have

$$\sum_{m=1}^{n} \mathbb{E}\,|\xi_{n,m}|^2\, 1_{\{|\xi_{n,m}|>\epsilon\}} = 0\,.$$

By Lindeberg-Feller's theorem, the desired result follows. □

Applications To get a feel for what the central limit theorem says, we will look at some concrete cases.

Example 3.27 (Cycles in a random permutation and record values)**.** Continuing the analysis of Example 2.4 and Example 2.13, let $X_1, X_2, \cdots$ be independent with

$$\mathbb{P}\,(X_m = 1) = \frac{1}{m}\,, \text{ and } \mathbb{P}\,(X_m = 0) = 1 - \frac{1}{m}$$

then we have

$$\mathbb{E}X_m = \frac{1}{m}\,, \text{ and } \mathrm{Var}\,(X_m) = \frac{1}{m} - \frac{1}{m^2}\,.$$

Let $S_n = X_1 + \cdots + X_n$, by Corollary 3.25, we have

$$\frac{S_n - \sum_{m=1}^{n} \frac{1}{m}}{\sqrt{\log n}} \Rightarrow N(0,1)$$

Since $\sum_{m=1}^{n} \frac{1}{m} \sim \log n$ we have

$$\frac{S_n - \log n}{\sqrt{\log n}} \Rightarrow N(0,1)\,.$$

Example 3.28 (The converse of the Kolmogorov's three-series theorem)**.** Recall the Kolmogorov's three-series theorem: $\{X_n\}$ is independent r.v.'s. Take $A > 0$ and let $Y_n = X_n 1_{\{|X_n|\leqslant A\}}$. In order that $\sum_{n=1}^{\infty} X_n$ converges almost surely, it is necessary and sufficient that

$$(i)\ \sum_{n=1}^{\infty} \mathbb{P}\,(|X_n| > A) < \infty,\ (ii)\ \sum_{n=1}^{\infty} \mathbb{E}Y_n < \infty,\ (iii)\ \sum_{n=1}^{\infty} \mathrm{Var}\,(Y_n) < \infty\,.$$

In this example, we will use CLT to show the necessity.

Proof. Recall the proof of necessity in Theorem 2.24, (i) is clear, and (ii) follows if (iii) holds. The difficulty is to show (iii). We use an arguement by contradiction to show (iii).

Suppose next that the sum in (iii) is infinite. Let $T_n = Y_1 + \cdots Y_n$, thus $\mathrm{Var}(T_n) \to \infty$. By Corollary 3.25 we have,

$$\frac{T_n - \mathbb{E}T_n}{\sqrt{\mathrm{Var}(T_n)}} \Rightarrow N(0,1)\,.$$

Since $\sum_{n=1}^{\infty} X_n$ converges a.s., the probability of $\{X_n \neq Y_n \text{ i.o. }\}$ is zero, so

$$\frac{T_n}{\sqrt{\mathrm{Var}(T_n)}} \to 0 \quad \text{a.s.}$$

hence

$$\frac{\mathbb{E}T_n}{\sqrt{\mathrm{Var}(T_n)}} \Rightarrow N(0,1)\,,$$

which is a contradiction. □

EXERCISE In the next two problems $X_1, X_2, \cdots$ are independent and $S_n = X_1 + \cdots + X_n$

Exercise 34. Suppose $\mathbb{E}X_i = 0, \mathbb{E}X_i^2 = 1$ and $\mathbb{E}\,|X_i|^{2+\delta} \leqslant M$ for some $\delta > 0$ and $M < \infty$. Show that

$$\frac{S_n}{\sqrt{n}} \Rightarrow N(0,1)\,.$$

Exercise 35 (Lyapunov's Theorem). Let $\alpha_n = \sqrt{\mathrm{Var}\,(S_n)}$. If there is a $\delta > 0$ so that

$$\lim_{n\to\infty} \frac{1}{\alpha_n^{2+\delta}} \sum_{m=1}^{n} \mathbb{E}\,|X_m - \mathbb{E}X_m|^{2+\delta} = 0$$

then

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}\,(S_n)}} \Rightarrow N(0,1)\,.$$

Note that the previous exercise is a special case of this result.

3.3.3 Sufficient conditions of CLT*

Lindeberg's condition For each n, let $X_{n,m}$, $1 \leqslant m \leqslant n$, be independent random variables with $\mathbb{E}X_{n,m} = 0$, let $S_n = X_{n1} + \cdots + X_{nn}$. We have shown that Lindeberg's condition implies that

$$\max_{1 \leqslant m \leqslant n} \mathbb{E}X_{n,m}^2 \to 0, \quad n \to \infty \tag{3.32}$$

Remarkably, subject to this condition the validity of the Central Limit Theorem automatically implies Lindeberg's condition.

Theorem 3.26. *Suppose* $\sum_{m=1}^{n} \mathbb{E}X_{n,m}^2 \to \sigma^2 > 0$, *and (3.32) holds. Then , then*

$$S_n \Rightarrow N(0, \sigma^2)\,. \tag{3.33}$$

if and only if Lindeberg's condition holds.

Proof. Sufficiency is Lindeberg-Feller's theorem.

To show the necessity, denote by $\sigma_{n,m}^2$ the variance of $X_{n,m}$, and $\varphi_{n,m}$ the ch.f. of $X_{n,m}$, $\mu_{n,m}$ the distribution of $X_{n,m}$.

Since $S_n \Rightarrow N(0,1)$, we have, $\prod_{m=1}^{n} \varphi_{n,m}(t) \to \exp\left(-\frac{1}{2}t^2\right)$ for any fixed $t \in \mathbb{R}$. Taking the logarithm, it is

$$\sum_{m=1}^{n} \log \varphi_{n,m}(t) \to -\frac{\sigma^2}{2}t^2\,. \tag{3.34}$$

Lindeberg's condition is, for any $\epsilon > 0$,

$$\sum_{m=1}^{n} \int_{\{|x|>\epsilon\}} x^2\, \mu_{n,m}(dx) \to 0\,. \tag{3.35}$$

Step 1. We show that (3.34) implies

$$\sum_{m=1}^{n} \varphi_{n,m}(t) - 1 \to -\frac{\sigma^2}{2}t^2\,. \tag{3.36}$$

By Theorem 3.22, $\varphi_{n,m}(t) = 1 - \frac{\sigma_{n,m}^2}{2}t^2 + \varepsilon_{n,m}(t)$, and $|\varepsilon_{n,m}(t)| \leqslant 2t^2\sigma_{n,m}^2$, and (3.32),

$$\max_{1 \leqslant m \leqslant n} |\varphi_{n,m}(t) - 1| \to 0 \quad \text{as } n \to \infty\,.$$

Note that for large n

$$\sum_{m=1}^{n} |\log \varphi_{n,m}(t) - (\varphi_{n,m}(t) - 1)| \leqslant \sum_{m=1}^{n} |\varphi_{n,m}(t) - 1|^2$$
$$\leqslant \max_{1\leqslant m\leqslant n} |\varphi_{n,m}(t) - 1| \sum_{m=1}^{n} |\varphi_{n,m}(t) - 1| \,.$$

Therefore,

$$\sum_{m=1}^{n} |\log \varphi_{n,m}(t) - (\varphi_{n,m}(t) - 1)| \to 0 \quad \text{as } n \to \infty \,,$$

which deduces (3.36).

Step 2. By (3.36), we have

$$\frac{\sigma^2}{2} t^2 + \sum_{m=1}^{n} \int (e^{itx} - 1)\, \mu_{n,m}(dx) + \to 0 \,.$$

Note that $\sum_{m=1}^{n} \sigma_{n,m}^2 \to \sigma^2$, we have

$$\sum_{m=1}^{n} \frac{\sigma_{n,m}^2}{2} t^2 + \int (e^{itx} - 1)\, \mu_{n,m}(dx)$$
$$= \sum_{m=1}^{n} \int e^{itx} - 1 + \frac{t^2}{2} x^2 \, \mu_{n,m}(dx) \to 0 \,.$$

Taking the real part, we have

$$\sum_{m=1}^{n} \int \cos tx - 1 + \frac{t^2}{2} x^2 \, \mu_{n,m}(dx) \to 0 \,.$$

Since t is arbitary, let $t = 1$, then

$$\sum_{m=1}^{n} \int \cos x - 1 + \frac{1}{2} x^2 \, \mu_{n,m}(dx) \to 0 \,.$$

Step 3. One can show that easily, by calculus,

(i) $\cos x - 1 + \frac{1}{2} x^2 \geqslant 0$ for all x, and

(ii) for any given $\epsilon > 0$, there exists $C > 0$, depending on ϵ, so that

$$\frac{\cos x - 1}{x^2} + \frac{1}{2} > C \quad \text{for all } |x| \geqslant \epsilon$$

Thus

$$\begin{aligned}\sum_{m=1}^{n} \int \cos x - 1 + \frac{1}{2}x^2 \, \mu_{n,m}(dx) &\geqslant \sum_{m=1}^{n} \int_{\{|x|>\epsilon\}} \cos x - 1 + \frac{1}{2}x^2 \, \mu_{n,m}(dx) \\ &\geqslant \sum_{m=1}^{n} \int_{\{|x|>\epsilon\}} Cx^2 \, \mu_{n,m}(dx) \,,\end{aligned}$$

and Lindeberg's condition follows. □

Other conditions Firstly, we point that, even if the variance of $X_{n,m}$ is infinity, CLT may also hold.

Example 3.29 (Infinite variance)**.** Suppose $\xi_1, \xi_2, \cdots$ are i.i.d. r.v.'s such that $\mathbb{P}(|\xi_1| > x) = x^{-2}$ for $x \geqslant 1$, and $\mathbb{P}(\xi_1 > x) = \mathbb{P}(\xi_1 < -x)$. Then

$$\mathbb{E}\,|\xi_1|^2 = \int_0^\infty 2x\mathbb{P}(|\xi_1| > x)\, dx = \infty \,.$$

Let

$$X_{n,m} = \frac{\xi_m}{\sqrt{n \log n}} \,, 1 \leqslant m \leqslant n \,.$$

Clearly $\mathbb{E}X_{n,m} = 0$ and $\mathbb{E}X_{n,m}^2 = \infty$. However, let $S_n = X_{n,1} + \cdots + X_{n,n}$, there holds $S_n \Rightarrow N(0,1)$.

To see this, we truncate ξ_m by level c_n which is chosen later, let

$$Y_{n,m} = \xi_m 1_{\{|\xi_m| \leqslant c_n\}} \,, 1 \leqslant m \leqslant n \,.$$

To ensure $\sum_{m=1}^{n} \mathbb{P}(Y_{n,m} \neq \xi_m) = n\,\mathbb{P}(|\xi_1| > c_n) \to 0$, let $c_n = \sqrt{n} \log\log n$, we want the variance of $Y_{n,m}$ to be as small as possible, so we keep the truncation close to the lowest possible level.

Our next step is to show $\mathbb{E}Y_{n,m}^2 \sim \log n$. For this we need upper and lower bounds. since $\mathbb{P}(|Y_{n,m}| > x) \leqslant \mathbb{P}(|\xi_1| > x)$ and is 0 for $x > c_n$, we have

$$\begin{aligned}\mathbb{E}Y_{n,m}^2 &\leqslant \int_0^{c_n} 2y\mathbb{P}(|\xi_1| > y)\, dy = 1 + \int_1^{c_n} \frac{2}{y}\, dy \\ &= 1 + 2\log c_n = 1 + \log n + 2\log\log\log n \sim \log n\end{aligned}$$

In the other direction, we observe $\mathbb{P}(|Y_{n,m}| > x) = \mathbb{P}(|\xi_1| > x) - \mathbb{P}(|\xi_1| > c_n)$ and the right-hand side is $\geqslant \left(1 - (\log\log n)^{-2}\right)\mathbb{P}(|\xi_1| > x)$ when $x \leqslant \sqrt{n}$ so

$$\mathbb{E}Y_{n,m}^2 \geqslant \left(1 - (\log\log n)^{-2}\right)\int_1^{\sqrt{n}} \frac{2}{y}\, dy \sim \log n$$

Since $\sum_{m=1}^n \mathrm{Var}(Y_m) \sim n\log n$, we apply CLT to

$$\bar{X}_{n,m} = \frac{Y_{n,m}}{\sqrt{n\log n}}, \leqslant 1 \leqslant m \leqslant n$$

Let $T_n = \bar{X}_{n,1} + \cdots + \bar{X}_{n,n}$, Then $T_n \Rightarrow N(0,1)$. Since the choice of c_n guarantees $\mathbb{P}(S_n \neq T_n) \to 0$, the same result holds for S_n.

We have shown in that the Lindeberg condition implies the condition (3.32). In turn, this implies the so-called condition of **asymptotic negligibility**, that is, the condition that for every $\epsilon > 0$,

$$\max_{1\leqslant m\leqslant n} \mathbb{P}(|X_{n,m}| \geqslant \epsilon) \to 0, \quad n \to \infty \tag{3.37}$$

It's easy to check that the triangular array with infinite variance in Example 3.29 satisfying this condition. Consequently, we may say that

> Lindeberg-Feller's theorem provide a condition of validity of the central limit theorem for sums of independent random variables under the condition of asymptotic negligibility

Limit theorems in which the condition of asymptotic negligibility is imposed on individual terms are usually called theorems with a *classical formulation*.

It is easy, however, to give examples of nondegenerate random variables for which neither the Lindeberg condition nor the asymptotic negligibility condition is satisfied, but nevertheless the central limit theorem is satisfied. Here is the simplest example.

Example 3.30. Let $\{\xi_k\}$ be a sequence of independent normally distributed random variables with $\mathbb{E}\xi_k = 0$,

$$\mathbb{E}\xi_1^2 = 1 \quad \mathbb{E}\xi_k^2 = 2^{k-2}, k \geqslant 2.$$

Define

$$X_{n,k} = \frac{\xi_k}{\sqrt{\sum_{k=1}^n \mathbb{E}\xi_k^2}} = \frac{\xi_k}{2^{\frac{n-1}{2}}},$$

Let $S_n = X_{n,1} + \cdots + X_{n,n}$. Obviously, S_n is normally distributed, with $\mathbb{E}S_n = 0$ and $\mathbb{E}S_n^2 = 1$. Hence $S_n = N(0,1)$ for all n.

However, it is easily verified that here the asymptotic negligibility condition is satisfied :

$$\max_{1\leqslant k\leqslant n} \mathbb{P}(|X_{n,k}| \geqslant \epsilon) = \max_{1\leqslant k\leqslant n} \mathbb{P}\left(|\xi_k| \geqslant \epsilon\, 2^{\frac{n-1}{2}}\right)$$

Since

$$\frac{\xi_n}{2^{\frac{n}{2}-1}} \sim N(0,1)$$

we have

$$\mathbb{P}\left(|\xi_n| \geqslant \epsilon\, 2^{\frac{n-1}{2}}\right) \geqslant \mathbb{P}\left(|N(0,1)| \geqslant \sqrt{2}\epsilon\right)$$

Thus we have that $\{X_{n,k}\}$ is not asymptotic negligibility.

Nonclassical conditions* We shall suppose that we are given a "triangle array" of random variables, i.e., for each $n \geqslant 1$ we have n independent random variables

$$X_{n1}, X_{n2}, \cdots, X_{nn}$$

with $\mathbb{E}X_{nk} = 0$, $\operatorname{Var} X_{nk} = \sigma_{nk}^2 > 0$, and $\sum_{k=1}^n \sigma_{nk}^2 \to 1$. Let $S_n = \sum_{k=1}^n X_{nk}$ and denote by $F_{n,k}$ the c.d.f. of $X_{n,k}$, $\Phi(x)$ the c.d.f. of $N(0,1)$ and

$$\Phi_{nk}(x) \coloneqq \Phi\left(\frac{x}{\sigma_{nk}}\right)$$

The theorem below provides a sufficient (and necessary) condition for the central limit theorem without assuming the "classical" condition of asymptotic negligibility. In this sense, condition (Λ), presented below, is an example of "nonclassical" conditions which reflect the title of this section.

Theorem. *To have $S_n \Rightarrow N(0, \sigma^2)$, it is sufficient (and necessary) that for every $\epsilon > 0$, the condition*

$$\sum_{k=1}^{n} \int_{|x|>\epsilon} |x| \, |F_{nk}(x) - \Phi_{nk}(x)| \, dx \to 0, \quad n \to \infty \tag{Λ}$$

is satisfied.

3.4 Poisson convergence

3.4.1 The basic limit theorem

Our first result is sometimes facetiously called the *"law of small numbers"* . These names derive from the fact that the Poisson appears as the limit of a sum of indicators of events that have small probabilities.

Theorem 3.27. *For each n let $X_{n,m}, 1 \leqslant m \leqslant n$ be independent random variables with $\mathbb{P}(X_{n,m} = 1) = p_{n,m}, \mathbb{P}(X_{n,m} = 0) = 1 - p_{n,m}$. Suppose*

(i) $\sum_{m=1}^{n} p_{n,m} \to \lambda \in (0, \infty)$, *and*

(ii) $\max_{1 \leqslant m \leqslant n} p_{n,m} \to 0$

Let $S_n = X_{n,1} + \cdots + X_{n,n}$, then $S_n \Rightarrow \text{Poisson}(\lambda)$.

Note that in the spirit of the Lindeberg-Feller theorem, no single term contributes very much to the sum. In contrast to that theorem, the contributions, when positive, are not small.

First proof. Let $\varphi_{n,m}(t) = \mathbb{E}e^{itX_{n,m}} = (1 - p_{n,m}) + p_{n,m}e^{it}$. Then it suffices to show that, for all t,

$$\varphi_n(t) = \mathbb{E}e^{itS_n} = \prod_{m=1}^{n} \left(1 + p_{n,m}\left(e^{it} - 1\right)\right) \to \exp(\lambda(e^{it} - 1)).$$

Taking logarithm,

$$\begin{aligned} \log \varphi_n(t) &= \sum_{m=1}^{n} \log\left[1 + p_{n,m}\left(e^{it} - 1\right)\right] \\ &= \sum_{m=1}^{n} p_{n,m}\left(e^{it} - 1\right) + \sum_{m=1}^{n} \theta_{n,m} p_{n,m}^2 |e^{it} - 1|^2 . \end{aligned}$$

Observe that the first term $\sum_{m=1}^{n} p_{n,m}\left(e^{it} - 1\right) \to \lambda\left(e^{it} - 1\right)$, and the second is domainted by

$$\sum_{m=1}^{n} p_{n,m}^2 \leqslant \max_{1 \leqslant m \leqslant n} p_{m,n} \sum_{m=1}^{n} p_{n,m} \to 0$$

Thus we complete the proof. □

REMARK. Note that S_n is the sum of n independent random variables, but S_n does not converge to the normal distribution. The problem is that the hypotheses of the Lindeberg-Feller theorem are not satisfied. Let $\xi_{n,m} = X_{n,m} - p_{n,m}$ for $1 \leqslant m \leqslant n$, then

$$\mathbb{E}\xi_{n,m} = 0 \quad \text{and} \quad \mathbb{E}\xi_{n,m}^2 = p_{n,m} - p_{n,m}^2$$

Thus $\sum_{m=1}^n \mathbb{E}\xi_{n,m}^2 \to \lambda$, but the Lindeberg's condition is false : for fiven $\epsilon > 0$,

$$\begin{aligned}\sum_{m=1}^n \mathbb{E}\xi_{n,m}^2 1_{\{|\xi_{n,m}|>\epsilon\}} &= \sum_{m=1}^n \mathbb{E}|X_{n,m} - p_{n,m}|^2 1_{|X_{n,m}-p_{n,m}|>\epsilon} \\ &= \sum_{m=1}^n (1-p_{n,m})^2 p_{n,m} \to \lambda\,.\end{aligned}$$

Theorem 3.27 generalizes trivially to give the following result.

Theorem 3.28. *Let $X_{n,m}, 1 \leqslant m \leqslant n$ be independent nonnegative integer valued random variables with $\mathbb{P}(X_{n,m} = 1) = p_{n,m}$, $\mathbb{P}(X_{n,m} \geqslant 2) = \epsilon_{n,m}$.*

(i) $\sum_{m=1}^n p_{n,m} \to \lambda \in (0,\infty)$,

(ii) $\max_{1\leqslant m\leqslant n} p_{n,m} \to 0$, *and*

(iii) $\sum_{m=1}^n \epsilon_{n,m} \to 0$

Let $S_n = X_{n,1} + \cdots + X_{n,n}$, then $S_n \Rightarrow$ Poisson(λ).

Proof. Let $Y'_{n,m} = 1$ if $X_{n,m} = 1$, and 0 otherwise. Let $T_n = Y_{n,1} + \cdots + Y_{n,n}$, (i) - (ii) and Theorem 3.27 imply $T_n \Rightarrow$ Poisson(λ), (iii) tells us $S_n - T_n \to 0$ in probability, then the result follows. □

We will now consider some concrete situations in which Theorem 3.27 can be applied. In each case we are considering a situation in which $p_{n,m} = c/n$, so we approximate the distribution of the sum by a Poisson with mean c.

Example 3.31. In a calculus class with 400 students, the number of students who have their birthday on the day of the final exam has approximately a Poisson distribution with mean $400/365 = 1.096$. This means that

the probability no one was born on that date is about $e^{-1.096} = 0.334$. Similar reasoning shows that the number of babies born on a given day or the number of people who arrive at a bank between 1: 15 and 1: 30 should have a Poisson distribution.

Example 3.32. Suppose we roll two dice 36 times. The probability of "double ones" (one on each die) is 1/36 so the number of times this occurs should have approximately a Poisson distribution with mean 1. Comparing the Poisson approximation with exact probabilities shows that the agreement is good even though the number of trials is small.

k	0	1	2	3
Poisson	0.3678	0.3678	0.1839	0.0613
exact	0.3627	0.3730	0.1865	0.0604

After we give the second proof of Theorem 3.27, we will discuss rates of convergence. Those results will show that for large n the largest discrepancy occurs for $k = 1$ and is about $(2en)^{-1}$ $(= 0.0051$ in this case).

Second proof, rates of convergence Our second proof of Theorem 3.27 requires a little work but provides information about the rate of convergence. The definition of total variation distance is given in Section 3.6.

The next two lemmas are the keys to our second proof.

Lemma 3.29. *Suppose μ_1, μ_2, ν_1, ν_2 are probability measures on $\mathbb{Z}$.*

(i) If $\mu_1 \times \mu_2$ denotes the product measure on $\mathbb{Z} \times \mathbb{Z}$, then

$$\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\| \leqslant \|\mu_1 - \nu_1 + \| \mu_2 - \nu_2\| .$$

(ii) If $\mu_1 * \mu_2$ denotes the convolution of μ_1 and μ_2, that is,

$$(\mu_1 * \mu_2)(x) = \sum_y \mu_1(x - y)\mu_2(y) \quad \text{for all } x \in \mathbb{Z} ,$$

then

$$\|\mu_1 * \mu_2 - \nu_1 * \nu_2\| \leqslant \|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\| .$$

Proof. To show (i), observe that

$$\begin{aligned}
2\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\| &= \sum_{x,y} |\mu_1(x)\mu_2(y) - \nu_1(x)\nu_2(y)| \\
&\leqslant \sum_{x,y} |\mu_1(x)\mu_2(y) - \nu_1(x)\mu_2(y)| + \sum_{x,y} |\nu_1(x)\mu_2(y) - \nu_1(x)\nu_2(y)| \\
&= \sum_y \mu_2(y) \sum_x |\mu_1(x) - \nu_1(x)| + \sum_x \nu_1(x) \sum_y |\mu_2(y) - \nu_2(y)| \\
&= 2\,\|\mu_1 - \nu_1\| + 2\,\|\mu_2 - \nu_2\|
\end{aligned}$$

which gives the desired result.

To show (ii), note that

$$\begin{aligned}
2\|\mu_1 * \mu_2 - \nu_1 * \nu_2\| &= \sum_x \left| \sum_y \mu_1(x-y)\mu_2(y) - \sum_y \nu_1(x-y)\nu_2(y) \right| \\
&\leqslant \sum_x \sum_y |\mu_1(x-y)\mu_2(y) - \nu_1(x-y)\nu_2(y)| \\
&= 2\,\|\mu_1 \times \mu_2 - \nu_1 \times \nu_2\|
\end{aligned}$$

which gives the desired result. □

Lemma 3.30. $\|\text{Bernoulli}(p) - \text{Poisson}(p)\| \leqslant p^2$.

Proof. Let μ be Bernoulli distribution with parameter p, ν be a Poisson distribution with parameter p. Then

$$\begin{aligned}
2\|\mu - \nu\| &= |\mu(0) - \nu(0)| + |\mu(1) - \nu(1)| + \sum_{n \geqslant 2} \nu(n) \\
&= \left|1 - p - e^{-p}\right| + \left|p - pe^{-p}\right| + 1 - e^{-p}(1+p)\,,
\end{aligned}$$

since $1 - x \leqslant e^{-x} \leqslant 1$ for $x \geqslant 0$, the above

$$\begin{aligned}
&= e^{-p} - 1 + p + p\,(1 - e^{-p}) + 1 - e^{-p} - pe^{-p} \\
&= 2p\,(1 - e^{-p}) \leqslant 2p^2
\end{aligned}$$

which gives the desired result. □

Second proof of Theorem 3.27. Let $\mu_{n,m}$ be the distribution of $X_{n,m}$. Let μ_n be the distribution of S_n. Let $\nu_{n,m}, \nu_n$, and ν be Poisson distributions with means $p_{n,m}$, $\lambda_n = \sum_{m=1}^n p_{n,m}$, and λ respectively. since $\mu_n = \mu_{n,1} * \cdots * \mu_{n,n}$ and $\nu_n = \nu_{n,1} * \cdots * \nu_{n,n}$, the three lemmas above imply

$$\|\mu_n - \nu_n\| \leqslant \sum_{m=1}^n \|\mu_{n,m} - \nu_{n,m}\| \leqslant \sum_{m=1}^n p_{n,m}^2$$

Using the definition of total variation distance now gives

$$\sup_A |\mu_n(A) - \nu_n(A)| \leqslant \sum_{m=1}^n p_{n,m}^2 \tag{3.38}$$

Assumptions (i) and (ii) imply that the right-hand side $\to 0$, since $\nu_n \Rightarrow \nu$ as $n \to \infty$ the result follows. □

REMARK. The proof above is due to Hodges and Le Cam. By different methods, C. Stein has proved

$$\sup_A |\mu_n(A) - \nu_n(A)| \leqslant \frac{1}{\lambda \vee 1} \sum_{m=1}^n p_{n,m}^2$$

Example 3.33 (Rates of convergence)**.** When $p_{n,m} = \frac{1}{n}$, (3.38) becomes

$$\sup_A |\mu_n(A) - \nu_n(A)| \leqslant \frac{1}{n}\,.$$

To assess the quality of this bound, we will compare the Poisson and binomial probabilities for k successes.

k	Poisson(1)	Binomial$(n, \frac{1}{n})$	
0	e^{-1}	$\left(1 - \frac{1}{n}\right)^n$	
1	e^{-1}	$n \cdot n^{-1} \left(1 - \frac{1}{n}\right)^{n-1}$	$= \left(1 - \frac{1}{n}\right)^{n-1} /2!$
2	$e^{-1}/2!$	$\binom{n}{2} n^{-2} \left(1 - \frac{1}{n}\right)^{n-3}$	$= \left(1 - \frac{1}{n}\right)^{n-1} /2!$
3	$e^{-1}/3!$	$\binom{n}{3} n^{-3} \left(1 - \frac{1}{n}\right)^{n-3}$	$= \left(1 - \frac{2}{n}\right) \left(1 - \frac{1}{n}\right)^{n-2} /3!$

since $(1 - x) \leqslant e^{-x}$, we have $\mu_n(0) - \nu_n(0) \leqslant 0$. Expanding

$$\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots$$

gives

$$(n-1)\log\left(1-\frac{1}{n}\right) = -\frac{n-1}{n} - \frac{n-1}{2n^2} - \cdots = -1 + \frac{1}{2n} + O\left(n^{-2}\right)$$

So

$$n\left(\left(1-\tfrac{1}{n}\right)^{n-1} - e^{-1}\right) = ne^{-1}\left(\exp\left\{1/2n + O\left(n^{-2}\right)\right\} - 1\right) \to e^{-1}/2$$

and it follows that

$$n\left(\mu_n(1) - \nu_n(1)\right) \to e^{-1}/2$$
$$n\left(\mu_n(2) - \nu_n(2)\right) \to e^{-1}/4$$

For $k \geqslant 3$, using $(1-2/n) \leqslant (1-1/n)^2$ and $(1-x) \leqslant e^{-x}$ shows $\mu_n(k) - \nu_n(k) \leqslant 0$ 80

$$\sup_{A \subset Z} |\mu_n(A) - \nu_n(A)| \approx 3/4en$$

There is a large literature on Poisson approximations for dependent events. Here we consider

3.4.2 Two examples with dependence

Example 3.34 (Matching)**.** Let π be a random permutation of $\{1, 2, \cdots, n\}$, let $X_{n,m} = 1$ if m is a fixed point (0 otherwise), and let $S_n = X_{n,1} + \cdots + X_{n,n}$ be the number of fixed points. We want to compute

$$\mathbb{P}\left(S_n = 0\right).$$

For a more exciting story consider men checking hats or wives swapping husbands. Let $A_{n,m} = \{X_{n,m} = 1\}$ The inclusion-exclusion formula implies

$$\begin{aligned}\mathbb{P}\left(\cup_{m=1}^n A_m\right) =& \sum_m \mathbb{P}\left(A_m\right) - \sum_{\ell<m} \mathbb{P}\left(A_\ell \cap A_m\right) \\ &+ \sum_{k<\ell<m} \mathbb{P}\left(A_k \cap A_\ell \cap A_m\right) - \cdots \\ =& n \cdot \frac{1}{n} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \cdots\end{aligned}$$

since the number of permutations with k specified fixed points is $(n-k)!$ Canceling some factorials gives

$$\mathbb{P}\left(S_n > 0\right) = \sum_{m=1}^n \frac{(-1)^{m-1}}{m!} \quad \text{so} \quad \mathbb{P}\left(S_n = 0\right) = \sum_{m=0}^n \frac{(-1)^m}{m!}$$

Recognizing the second sum as the first $n+1$ terms in the expansion of e^{-1} gives

$$\begin{aligned}\left|\mathbb{P}\left(S_n=0\right)-e^{-1}\right| &= \left|\sum_{m=n+1}^{\infty}\frac{(-1)^m}{m!}\right| \\ &\leqslant \frac{1}{(n+1)!}\left|\sum_{k=0}^{\infty}(n+2)^{-k}\right| = \frac{1}{(n+1)!}\cdot\left(1-\frac{1}{n+2}\right)^{-1}\end{aligned}$$

a much better rate of convergence than $1/n$. To compute the other probabilities, we observe that by considering the locations of the fixed points

$$\begin{aligned}\mathbb{P}\left(S_n=k\right) &= \binom{n}{k}\frac{1}{n(n-1)\cdots(n-k+1)}\mathbb{P}\left(S_{n-k}=0\right) \\ &= \frac{1}{k!}\mathbb{P}\left(S_{n-k}=0\right)\to e^{-1}\frac{1}{k!}\end{aligned}$$

Hence we have proved that

$$S_n \Rightarrow \text{Poisson}(1)\,.$$

Example 3.35 (Occupancy problem)**.** Suppose that $r=r(n)$ balls are placed at random into n boxes. It follows from the Poisson approximation to the binomial that if $n\to\infty$ and $\frac{r(n)}{n}\to c$, then the number of balls in a given box will approach Poisson(c). The last observation should explain why the fraction of empty boxes approached e^{-c} in Example 2.5.

Here we will show :

Claim : If $ne^{-r(n)/n}\to\lambda\in[0,\infty)$, the number of empty boxes approaches Poisson(λ).

To see where the answer comes from, notice that in the Poisson approximation the probability that a given box is empty is $e^{-r(n)/n}\approx\lambda/n$, so if the occupancy of the various boxes were independent, the resuld follow from Theorem 3.27. To prove the result, we begin by observing

$$\mathbb{P}\left(\text{ boxes } i_1,i_2,\cdots,i_k \text{ are empty }\right)=\left(1-\frac{k}{n}\right)^r$$

If we let $p_m(r,n) =$ the probability exactly m boxes are empty when $r = r(n)$ balls are put in n boxes, then

$$\mathbb{P}(\text{ no empty box }) = 1 - \mathbb{P}(\text{ at least one empty box }) .$$

So by inclusion-exclusion

$$p_0(r,n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^r .$$

By considering the locations of the empty boxes

$$p_m(r,n) = \binom{n}{m} \left(1 - \frac{m}{n}\right)^r p_0(r, n-m) .$$

To evaluate the limit of $p_m(r,n)$ we begin by showing that if $ne^{-r(n)/n} \to \lambda$ then by Stirling's formula,

$$\binom{n}{k} \left(1 - \frac{k}{n}\right)^r \to \frac{\lambda^k}{k!}$$

then

$$p_0(r,n) \to \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{k!} = e^{-\lambda}$$

and

$$p_m(r,n) \to e^{-\lambda} \frac{\lambda^m}{m!} .$$

Thus, the number of empty boxes approaches Poisson(λ).

Example 3.36 (Coupon collector's problem)**.** Let $X_1, X_2, \cdots$ be i.i.d. uniform on $\{1, 2, \cdots, n\}$ and

$$T_n = \inf \{m : \{X_1, \cdots X_m\} = \{1, 2, \cdots, n\}\} ,$$

Since $T_n \leqslant m$ if and only if m balls fill up all n boxes, it follows from Theorem 3.27 that

$$P\left(T_n - n \log n \leqslant nx\right) \to \exp\left(-e^{-x}\right)$$

To see this, let $r(n) = n \log n + nx$ then $ne^{-r(n)/n} \to e^{-x}$.

For a concrete instance of the previous result consider: What is the probability that in a village of $2190(=6\cdot 365)$ people all birthdays are represented? Do you think the answer is much different for $1825(=5\cdot 365)$ people?

Solution. Here $n=365$, so $365\log 365 = 2153$ and

$$\begin{aligned} P\left(T_{365}\leqslant 2190\right) &= P\left(\left(T_{365}-2153\right)/365 \leqslant 37/365\right) \\ &\approx \exp\left(-e^{-0.1014}\right) = \exp(-0.9036) = 0.4051 \\ P\left(T_{365}\leqslant 1825\right) &= P\left(\left(T_{365}-2153\right)/365 \leq -328/365\right) \\ &\approx \exp\left(-e^{0.8986}\right) = \exp(-2.4562) = 0.085 \end{aligned}$$

3.4.3 An introduction to Poisson process

Let $N(t)$ be the number of arrivals at a bank in the time interval $[0,t]$, denote $N(s,t]$ as $N(t)-N(s)$, is the number of arrivals at a bank in the time interval $(s,t]$. Suppose

(i) the numbers of arrivals in disjoint intervals are independent,

(ii) the distribution of $N(s,t]$ only depends on $t-s$,

(iii) $\mathbb{P}(N(h)=1)=\lambda h+o(h)$ and $\mathbb{P}(N(h)\geqslant 2)=o(h)$.

Cliam : $N(t)$ has a Poisson distribution with mean λt.

Proof. To see this, fix t, let

$$X_{n,m} = N\left((m-1)\frac{t}{n}, m\frac{t}{n}\right], \quad \text{for } 1\leqslant m\leqslant n$$

and apply Theorem 3.28, which gives the desired result. □

Thus, we introduce the Poisson process, as following

Definition 3.5. A family of random variables $\{N(t), t\geqslant 0\}$, satisfying

(i) $\{N(t)\}$ has independent increments, that is, for any $0=t_0<t_1<\ldots<t_n$,

$$N\left(t_k\right)-N\left(t_{k-1}\right), 1\leqslant k\leqslant n$$

are independent,

(ii) $\{N(t)\}$ has stationary increments, and $N(t)-N(s)$ is Poisson $(\lambda(t-s))$

is called a **Poisson process with rate (intensity)** λ, denoted by $\mathrm{PP}(\lambda)$.

In fact, we only care the Poisson process with almost surely right-continuous paths. We have another method to construct this Poisson.

Construct a Poisson process A simple way to construct a Poisson process of rate λ is to take a sequence $S_1, S_2, \cdots$ of independent exponential random variables of parameter λ, to set $J_0 = 0$, $J_n = S_1 + \cdots + S_n$, then set

$$N(t) = n \quad \text{if} \quad J_n \leqslant t < J_{n+1} \tag{3.39}$$

or equivalently

$$N(t) = \sum_{n=1}^{\infty} 1_{\{J_n \leqslant t\}} \quad \text{for all } t \geqslant 0\,. \tag{3.40}$$

The following diagram illustrates a typical path, and it's easy to check that $\{J_n\}$ is exactly the jump times of the right-continuous process $(N_t)_{t\geqslant 0}$.

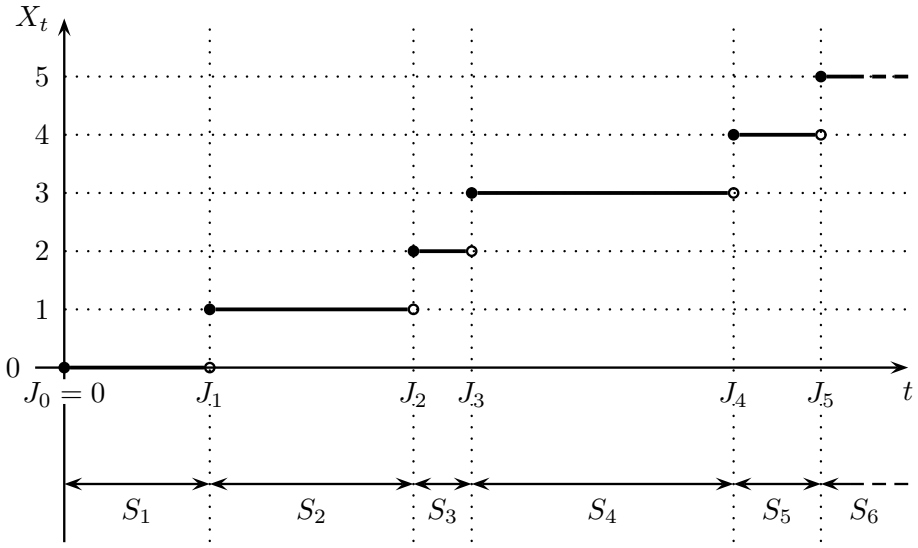

Figure 3.1: Construct a Poisson process

We will show that, this process is exactly Poisson process. Firstly, for each $t > 0$, $N(t)$ has a Poisson(λt) distribution.

Proof. For any $n \in \mathbb{N}$,

$$\mathbb{P}(N(t) = n) = \mathbb{P}(J_n \leqslant t < J_{n+1}) = \int_0^t f_{J_n}(s)\mathbb{P}\left(S_{n+1} > t - s\right) ds$$

Note that $J_n = S_1 + \cdots S_n$ has a gamma(n, λ) distribution, that is

$$f_{J_n}(s) = \lambda e^{-\lambda s} \cdot \frac{(\lambda s)^{n-1}}{(n-1)!} 1_{\{s>0\}} .$$

So we have

$$\begin{aligned}\mathbb{P}(N(t) = n) &= \int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} \cdot e^{-\lambda(t-s)} ds \\ &= \frac{\lambda^n}{(n-1)!} e^{-\lambda t} \int_0^t s^{n-1} ds = e^{-\lambda t} \frac{(\lambda t)^n}{n!} ,\end{aligned}$$

which proves the desired result. □

It's a little diffuse to check that the number of arrivals in disjoint intervals is independent, we omit it, a proof of the following theorem can be found in my notes : *Introduction to Stochastic processes*.

Claim : for any $s > 0$, $\{N(t+s) - N(s)\}$ is identically distributed with $\{N(t)\}$, and independent of $\{N(t) : 0 \leqslant t \leqslant s\}$.

Clearly, we have shown that we construct a Poisson process with right-continuous paths.

3.5 Limit Theorems in $\mathbb{R}^d$

We begin by considering weak convergence in a general metric space (S, ρ). In this setting we say $\{X_n\}$ weakly converges to X, write $X_n \Rightarrow X$, if and only if

$$Ef(X_n) \to Ef(X_\infty)$$

for all bounded continuous function f. As in Section 3.1, it will be useful to have several equivalent definitions of weak convergence.

Recall that f is said to be *Lipschitz continuous* on (S, ρ), if there is a constant C so that $|f(x) - f(y)| \leq C\rho(x, y)$.

Theorem 3.31. *The following statements are equivalent:*

(i) $Ef(X_n) \to Ef(X_\infty)$ *for all bounded continuous* f

(ii) $Ef(X_n) \to Ef(X_\infty)$ *for all bounded Lipschitz continuous* f

(iii) *For all closed sets* K, $\limsup P(X_n \in K) \leq P(X_\infty \in K)$

(iv) *For all open sets* G, $\liminf P(X_n \in G) \geq P(X_\infty \in G)$

(v) *For all sets* A *with* $P(X_\infty \in \partial A) = 0$, $\lim P(X_n \in A) = P(X_\infty \in A)$

Theorem 3.32 (Skorokhod). *μ_n and μ are p.m.'s on $(S, \mathcal{B})$ and $\mu_n \Rightarrow \mu$. We can find random variables X_n and X with distributions μ_n and μ, so that*

$$X_n \to X \quad a.s. \tag{3.41}$$

Theorem 3.33 (Continuous mapping theorem). *Let g be a measurable function on $(S, \mathcal{B})$, and denote by D_g all the discontinuty point of g. If $X_n \Rightarrow X$ and $\mathbb{P}(X \in D_g) = 0$, then*

$$g(X_n) \Rightarrow g(X).$$

If in addition g is bounded then $\mathbb{E}\, g(X_n) \to \mathbb{E}\, g(X_\infty)$.

3.6 Appendix : Total variance distance

Definition 3.6. $(\Omega, \mathcal{F})$ is a measurable space, μ, ν are two probability masures on it. The **total variance distance** between μ and ν is defined by

$$\|\mu - \nu\| \coloneqq \sup_{A \subset F} |\mu(A) - \nu(A)| \tag{3.42}$$

Clearly, total variance distance really defines a distance (or a metric) on all the probabbility mesures on $(\Omega, \mathcal{F})$.

Remark. There are some equivalent definitions.

(i) $\|\mu - \nu\| = \frac{1}{2}|\mu - \nu|(\Omega)$, where $|\mu - \nu| = (\mu - \nu)^+ + (\mu - \nu)^-$.

To see this, by the Hahn decomposition of signed measure $(\mu - \nu)$, we have

$$\Omega = \Omega^+ \cup \Omega^-$$

so that $(\mu - \nu)(A \cap \Omega^+) \geqslant 0$ and $(\mu - \nu)(A \cap \Omega^-) \leqslant 0$ for all $A \in \mathcal{F}$. Thus

$$\begin{aligned} |\mu - v|(\Omega) &= (\mu - \nu)^+(\Omega) + (\mu - v)^-(\Omega) \\ &= \mu\left(\Omega^+\right) - \nu\left(\Omega^+\right) - [\mu\left(\Omega^-\right) - \nu\left(\Omega^-\right)] \\ &= 2(\mu\left(\Omega^+\right) - \nu\left(\Omega^+\right)) \leqslant 2\|\mu - \nu\|. \end{aligned}$$

On the other hand,

$$\begin{aligned} \|\mu - \nu\| &= \sup_{A \subset F} |\mu(A) - \nu(A)| \\ &\leqslant \sup_{A \subset F} (\mu - \nu)^+(A) \vee (\mu - \nu)^-(A) \\ &\leqslant (\mu - \nu)^+(\Omega) \vee (\mu - \nu)^-(\Omega) = \frac{1}{2}|\mu - \nu|(\Omega) \end{aligned}$$

(ii) $\|\mu - \nu\| = \frac{1}{2}\sup\{|\int \phi \, \mathrm{d}(\mu - \nu)| : \phi \text{ is measurable and } |\phi| \leqslant 1\}$.

Obviously, $|\int \phi \, \mathrm{d}(\mu-\nu)| \leqslant |\mu-\nu|(\Omega)$ for all measurable ϕ and $|\phi| \leqslant 1$. On the other hand, note that

$$\begin{aligned} |\mu-\nu|(\Omega) &= \int (1_{\Omega^+} - 1_{\Omega^-}) \, \mathrm{d}(\mu-\nu) \\ &\leqslant \sup \left\{ \left| \int \phi \mathrm{d}(\mu-\nu) \right| : \phi \text{ is measurable and } |\phi| \leqslant 1 \right\} \end{aligned}$$

Thus (ii) follows from (i).

When the probability space is a countable set $(S, 2^S)$, then

$$\|\mu-\nu\| = \sup_{A \subset S} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_z |\mu(z) - \nu(z)| \tag{3.43}$$

and we have

Proposition 3.34. *μ_n converges to μ in total variance distance if and only if $\mu_n(x) \to \mu(x)$ for all $x \in S$, which is equivalnet to $\mu_n \Rightarrow \nu$.*

Proof. One direction is trivial. We cannot have $\|\mu_n - \mu\| \to 0$ unless $\mu_n(x) \to \mu(x)$ for each $x \in S$.

To prove the converse, note that if $\mu_n(x) \to \mu(x)$,

$$\sum_{x \in S} |\mu_n(x) - \mu(x)| \to 0$$

by the dominated convergence theorem. □

Exercise 36. Show that $\|\mu-\nu\| \leqslant \delta$ if and only if there are random variables X and Y with distributions μ and ν so that $P(X \neq Y) \leqslant \delta$.