MATH 170B LECTURE NOTE 2: ELEMENTARY LIMIT THEOREMS

HANBAEK LYU

1. OVERVIEW OF LIMIT THEOREMS

The primary subject in this note is the sequence of i.i.d. RVs and their partial sums. Namely, let $X_1, X_2, \cdots$ be an (infinite) sequence of i.i.d. RVs, and define their nth partial sum $S_n = X_1 + X_2 + \cdots + X_n$ for all $n \geq 1$. If we call X_i the ith step size or *increment*, then the sequence of RVs $(S_n)_{n\geq 1}$ is called a *random walk*, where we usually set $S_0 = 0$. Think of X_i as the gain or loss after betting once in a casino. Then S_n is the net gain of fortune after betting n times. Of course there are ups and downs in the short term, but what we want to analyze using probability theory is the long-term behavior of the random walk $(S_n)_{n\geq 1}$. Results of this type is called limit theorems.

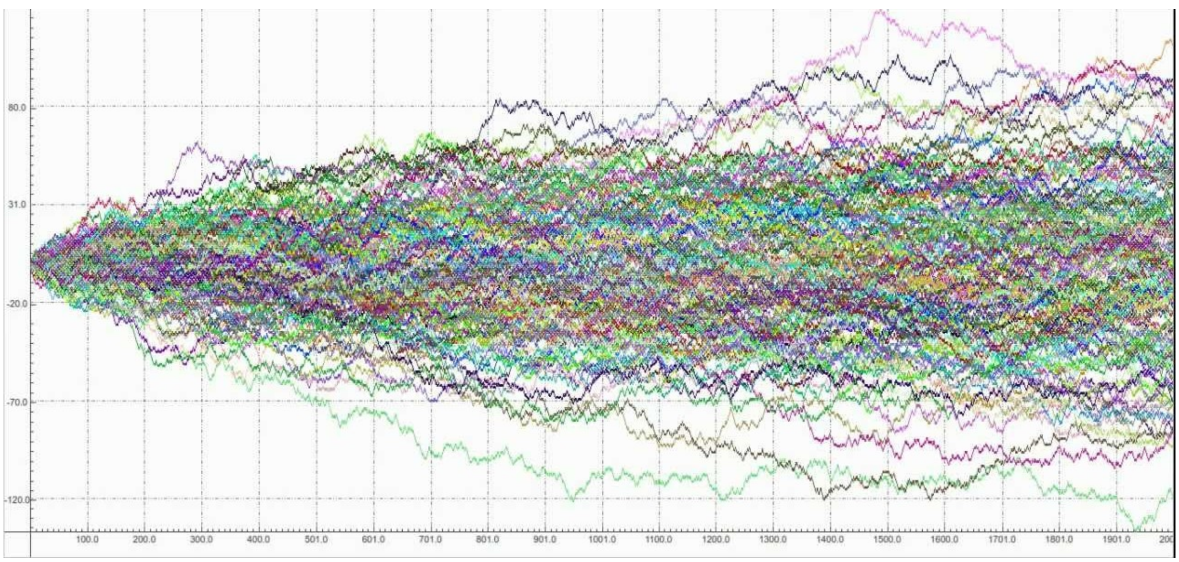

FIGURE 1. Simulation of simple random walks

Suppose each increment X_k has a finite mean μ. Then by linearity of expectation and independence of the increments, we have

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{\mathbb{E}[S_n]}{n} = \mu, \tag{1}$$

$$\text{Var}\left(\frac{S_n}{n}\right) = \frac{\text{Var}(S_n)}{n^2} = \frac{n\text{Var}(X_1)}{n^2} = \frac{\text{Var}(X_1)}{n}. \tag{2}$$

So the sample mean S_n/n has constant expectation and shrinking variance. Hence it makes sense to guess that it should behave as the constant μ, without taking the expectation. That is,

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu. \tag{3}$$

But this expression is shaky, since the left hand side is a limit of RVs while the right hand side is a constant. In what sense the random sample means converge to μ? This is the content of the *law of large numbers*, for which we will prove a weak and a strong versions.

The first limit theorem we will encounter is called the Weak Law of Large Numbers (WLLN), which is stated below:

Theorem 1.1 (WLLN)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with mean $\mu<\infty$ and let $S_n=\sum_{k=1}^{n} X_i$, $n\geq 1$ be a random walk. Then for any positive constant $\varepsilon>0$,*

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{S_n}{n}-\mu\right|>\varepsilon\right)=0. \tag{4}$$

In words, the probability that the sample mean S_n/n is *not* within ε distance from its expectation μ decays to zero as n tends to infinity. In this case, we say the sequence of RVs $(S_n/n)_{n\geq 1}$ converges to μ *in probability*.

The second version of law of large numbers is call the *strong law of large numbers* (SLLN), which is available if the increments have finite variance.

Theorem 1.2 (SLLN)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n=\sum_{k=1}^{n} X_i$, $n\geq 1$ be a random walk. Suppose $\mathbb{E}[X_1]=\mu<\infty$ and $\mathbb{E}[X_1^2]<\infty$. Then*

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{S_n}{n}=\mu\right)=1. \tag{5}$$

To make sense out of this, notice that the limit of sample mean $\lim_{n\to} S_n/n$ is itself a RV. Then SLLN says that this RV is well defined and its value is μ with probability 1. In this case, we say the sequence of RVs $(S_n/n)_{n\geq 1}$ converges to μ *with probability 1* or *almost surely*.

Perhaps one of the most celebrated theorems in probability theory is the *central limit theorem* (CLT), which tells about how the sample mean S_n/n "fluctuates" around its mean μ. From 2, if we denote $\sigma^2=\mathrm{Var}(X_1)<\infty$, we know that $\mathrm{Var}(S_n/n)=\sigma^2/n\to 0$ as $n\to\infty$. So the fluctuation decays as we add up more increments. To see the effect of fluctuation, we first center the sample mean by subtracting its expectation and "zoom in" by dividing by the standard deviation $\sigma/\sqrt{n}$. This is where the name 'central limit' comes from: it describes the limit of centered random walks.

Theorem 1.3 (CLT)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n=\sum_{k=1}^{n} X_i$, $n\geq 1$ be a random walk. Suppose $\mathbb{E}[X_1]=\mu<\infty$ and $\mathbb{E}[X_1^2]=\sigma^2<\infty$. Let $Z\sim N(0,1)$ be a standard normal RV and define*

$$Z_n=\frac{S_n-\mu n}{\sigma\sqrt{n}}=\frac{S_n/n-\mu}{\sigma/\sqrt{n}}. \tag{6}$$

Then for all $z\in\mathbb{R}$,

$$\lim_{n\to\infty}\mathbb{P}(Z_n\leq z)=\mathbb{P}(Z\leq z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{x^2}{2}}\,dx. \tag{7}$$

In words, the centered and rescaled RV Z_n is asymptotically distributed as a standard normal RV $Z\sim N(0,1)$. In this case, we say Z_n converges to Z as $n\to\infty$ *in distribution*. This is a remarkable result since as long as the increments X_k have finite mean and variance, it does not matter which distribution that they follow: the 'central limit' always looks like a standard normal distribution. Later in this section, we will prove this result by using the MGF of S_n and Taylor-expanding it up to the second order term.

In 170A, we will only study the Weak Law of Large Numbers and the Central Limit Theorem in a very special case when X_i's are Binomial RVs.

2. Bounding tail probabilities

In this subsection, we introduce two general inequalities called the Markov's and Chebyshef's inequalities. They are useful in bounding tail probabilities of the form $\mathbb{P}(X \geq x)$ using the expectation $\mathbb{E}[X]$ and variance $\mathrm{Var}(X)$, respectively. Their proofs are quite simple but they have lots of nice applications and implications.

Proposition 2.1 (Markov's inequality)**.** *Let $X \geq 0$ be a nonnegative RV with finite expectation. Then for any $a > 0$, we have*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{8}$$

Proof. Consider an auxiliary RV Y define as follows:

$$Y = \begin{cases} a & \text{if } X \geq a \\ 0 & \text{if } X < a. \end{cases} \tag{9}$$

Note that we always have $Y \leq X$. Hence we should have $\mathbb{E}[Y] \leq \mathbb{E}[X]$. But since $\mathbb{E}[Y] = a\mathbb{P}(X \geq a)$, we have

$$\lambda \mathbb{P}(X \geq a) \leq \mathbb{E}[X]. \tag{10}$$

Dividing both sides by $a > 0$ gives the assertion. □

Example 2.2. We will show that, for any RV Z, $\mathbb{E}[Z^2 = 0]$ implies $\mathbb{P}(Z = 0) = 1$. Indeed, Markov's inequality gives that for any $a > 0$,

$$\mathbb{P}(Z^2 \geq a) \leq \frac{\mathbb{E}[Z^2]}{a} = 0. \tag{11}$$

This means that $\mathbb{P}(Z^2 = 0) = 1$, so $\mathbb{P}(Z = 0) = 1$. ▲

Proposition 2.3 (Chebyshef's inequality)**.** *Let X be any RV with $\mathbb{E}[X] = \mu < \infty$ and $\mathrm{Var}(X) < \infty$. Then for any $a > 0$, we have*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}. \tag{12}$$

Proof. Applying Markov's inequality for the nonnegative RV $(X - \mu)^2$, we get

$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}((X - \mu)^2 \geq a^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}. \tag{13}$$

□

Example 2.4. Let $X \sim \mathrm{Exp}(\lambda)$. Since $\mathbb{E}[X] = 1/\lambda$, for any $a > 0$, the Markov's inequality gives

$$\mathbb{P}(X \geq a) \leq \frac{1}{a\lambda}, \tag{14}$$

while the true probability is

$$\mathbb{P}(X \geq a) = e^{-\lambda a}. \tag{15}$$

On the other hand, $\mathrm{Var}(X) = 1/\lambda^2$ so Chebyshef's inequality gives

$$\mathbb{P}(|X - 1/\lambda| \geq a) = \frac{1}{a^2\lambda^2}. \tag{16}$$

If $1/\lambda \leq a$, the true probability is

$$\begin{aligned} \mathbb{P}(|X - 1/\lambda| \geq a) &= \mathbb{P}(X \geq a + 1/\lambda) + \mathbb{P}(X \leq -a + 1/\lambda) && (17)\\ &= \mathbb{P}(X \geq a + 1/\lambda) = e^{-\lambda(a+1/\lambda)} = e^{-1-\lambda a}. && (18) \end{aligned}$$

As we can see, both Markov's and Chebyshef's inequalities give loose estimates, but the latter gives a slightly stronger bound. ▲

Example 2.5 (Chebyshef's inequality for bounded RVs)**.** Let X be a RV taking values from the interval $[a, b]$. Suppose we don't know anything else about X. Can we say anything useful about tail probability $\mathbb{P}(X \geq \lambda)$? If we were to use Markov's inequality, then certainly $a \leq \mathbb{E}[X] \leq b$ and in the worst case $\mathbb{E}[X] = b$. Hence we can at least conclude

$$\mathbb{P}(X \geq \lambda) \leq \frac{b}{\lambda}. \tag{19}$$

On the other hand, let's get a bound on $\mathrm{Var}(X)$ and use Chebyshef's inequality instead. We claim that

$$\mathrm{Var}(X) \leq \frac{(b-a)^4}{4}, \tag{20}$$

which would yield by Chebyshef's inequality that

$$\mathbb{P}(|X - \mathbb{E}[X]| \leq \lambda) \leq \frac{(b-a)^2}{4\lambda^2}. \tag{21}$$

Intuitively speaking, $\mathrm{Var}(X)$ is the largest when the value of X is as much spread out as possible at the two extreme values, a and b. Hence the largest variance will be achieved when X takes a and b with equal probabilities. In this case, $\mathbb{E}[X] = (a+b)/2$ so

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2 + b^2}{2} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{4}. \tag{22}$$

▲

Exercise 2.6. Let X be a RV taking values from the interval $[a, b]$.

(i) Use the usual 'completing squares' trick for a second moment to show that

$$0 \leq \mathbb{E}[(X-t)^2] = (t - \mathbb{E}[X])^2 + \mathrm{Var}(X) \quad \forall t \in \mathbb{R}. \tag{23}$$

(ii) Conclude that $\mathbb{E}[(X-t)^2]$ is minimized when $t = \mathbb{E}[X]$ and the minimum is $\mathrm{Var}(X)$.

(iii) By plugging in $t = (a+b)/2$ in (23), show that

$$\mathrm{Var}(X) = \mathbb{E}[(X-a)(X-b)] + \frac{(b-a)^2}{4} - \left(\mathbb{E}[X] - \frac{a+b}{2}\right)^2. \tag{24}$$

(iv) Show that $\mathbb{E}[(X-a)(X-b)] \leq 0$.

(v) Conclude that $\mathrm{Var}(X) \leq (b-a)^2/4$, where the equality holds if and only if X takes the extreme values a and b with equal probabilities.

Exercise 2.7 (Paley-Zigmund inequality)**.** Let X be a nonnegative RV with $\mathbb{E}[|X|] < \infty$. Fix a constant $\theta \geq 0$. We prove the Paley-Zigmund inequality, which gives a lower bound on the tail probabilities and also implies the so-called 'second moment method'.

(i) Write $X = X\mathbf{1}(X \geq \theta\mathbb{E}[X]) + X\mathbf{1}(X < \theta\mathbb{E}[X])$. Show that

$$\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}(X \leq \theta\mathbb{E}[X])] + \mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])] \tag{25}$$

$$\leq \theta\mathbb{E}[X] + \mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])]. \tag{26}$$

(ii) Use Cauchy-Schwartz inequality (Exc 2.11 in Lecture note 2) to show

$$(\mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])])^2 \leq \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}(X > \theta\mathbb{E}[X])^2] \tag{27}$$

$$= \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}(X > \theta\mathbb{E}[X])] \tag{28}$$

$$= \mathbb{E}[X^2]\mathbb{P}(X > \theta\mathbb{E}X). \tag{29}$$

(iii) From (i) and (ii), derive

$$\mathbb{E}[X] \leq \theta\mathbb{E}[X] + \sqrt{\mathbb{E}[X^2]\mathbb{P}(X > \theta\mathbb{E}X)}. \tag{30}$$

Conclude that

$$\mathbb{P}(X > \theta\mathbb{E}X) \geq \frac{(1-\theta)^2\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{31}$$

(iv) (Second moment method) From (iii), conclude that

$$\mathbb{P}(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{32}$$

3. The WLLN and convergence in probability

In this subsection, we prove the weak law of large numbers (Theorem 1.1) and study the notion of convergence in probability. Assuming finite variance for each increment, the weak law is an easy consequence of Chebyshef's inequality.

Theorem 3.1 (WLLN with second moment)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with finite mean $\mu < \infty$ and finite variance. Let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. Then for any positive constant $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0. \tag{33}$$

Proof. By Chebyshef's inequality, for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}, \tag{34}$$

where the last expression converges to 0 as $n \to \infty$. □

The proof of the full WLLN without the finite second moment assumption needs another technique called 'truncation'. We won't cover this technicality in this course and take Theorem 1.1 for granted.

The weak law of large numbers is the first time that we encounter the notion of 'convergence in probability'. We say a sequence of RVs converge to a constant in probability if the the probability of staying away from that constant goes to zero:

Definition 3.2. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let $\mu \in \mathbb{R}$ be a constant. We say X_n converges to μ *in probability* if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|X_n - \mu\right| > \varepsilon\right) = 0. \tag{35}$$

Before we proceed further, let us take a moment and think about the definition of convergence in probability. Recall that a sequence of real numbers $(x_n)_{n\geq 0}$ *converges* to x if for each 'error level' $\varepsilon > 0$, there exists a large integer $N(\varepsilon) > 0$ such that

$$|x_n - x| < \varepsilon \quad \forall n \geq N(\varepsilon). \tag{36}$$

If we would like to say that a sequence of RVs $(X_n)_{\geq 0}$ 'converges' to some real number x, how should we formulate this? Since X_n is an RV, $\{|X_n - x| < \varepsilon\}$ is an event. On the other hand, we can also view each x_n as an RV, even though it is a real number. Then we can rewrite (38) as

$$\mathbb{P}(|x_n - x| < \varepsilon) = 1 \quad \forall n \geq N(\varepsilon). \tag{37}$$

For general RVs, requiring $\mathbb{P}(|X_n - x| < \varepsilon) = 1$ for any large n might not be possible. But we can fix any desired level of 'confidence', $\delta > 0$, and require

$$\mathbb{P}(|x_n - x| < \varepsilon) \geq 1 - \delta \tag{38}$$

for sufficiently large n. This is precisely (35).

Example 3.3 (Empirical frequency)**.** Let A be an event of interest. We would like to estimate the unknown probability $p = \mathbb{P}(A)$ by observing a sequence of indpendent experiments. namely, let $(X_k)_{k\geq 0}$ be a sequence of i.i.d. RVs where $X_k = \mathbf{1}(A)$ is the indicator variable of the event A for each $k \geq 1$. Let $\hat{p}_n := (X_1 + \cdots + X_n)/n$. Since $\mathbb{E}[X_1] = \mathbb{P}(A) = p$, by WLLN we conclude that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\hat{p}_n - p| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty. \tag{39}$$

▲

Example 3.4 (Polling)**.** Let E_A be the event that a randomly select voter supports candidate A. Using a poll, we would like to estimate $p = \mathbb{P}(E_A)$, which can be understood as the proportion of supporters of candidate A. As before, we observe a sequence of i.i.d. indicator variables $X_k = \mathbf{1}(E_A)$. Let $\hat{p}_n := S_n/n$ be the empirical proportion of supporters of A out of n samples. We know by WLLN that $\hat{p}_n$ converges to p in probability. But if we want to guarantee a certain confidence level α for an error bound ε, how many samples should be take?

By Chebyshef's inequality, we get the following estimate:

$$\mathbb{P}\left(|\hat{p}_n - p| > \varepsilon\right) \leq \frac{\mathrm{Var}(\hat{p}_n)}{\varepsilon^2} = \frac{\mathrm{Var}[X_1]}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}. \tag{40}$$

Note that for the last inequality, we noticed that $X_1 \in [0,1]$ and used Exercise 2.6 (or you can use that for $Y \sim \text{Bernoulli}(p)$, $\mathrm{Var}(Y) = p(1-p) \leq 1/4$). Hence, for instance, if $\varepsilon = 0.01$ and $\alpha = 0.95$, then we would need to set n large enough so that

$$\mathbb{P}\left(|\hat{p}_n - p| > 0.01\right) \leq \frac{10000}{4n} \leq 0.05. \tag{41}$$

This yields $n \geq 50,000$. In other words, if we survey at least $n = 50,000$ independent voters, then the empirical frequency $\hat{p}_n$ is between $p - 0.01$ and $p + 0.01$ with probability at least 0.95. Still in other words, the true frequency p is between $\hat{p}_n - 0.01$ and $\hat{p}_n + 0.01$ with probability at least 0.95

if $n \geq 50,000$. We don't actually need this many samples. We will improve this result later using central limit theorem. ▲

Exercise 3.5 (Monte Carlo integration)**.** Let $(X_k)_{k\geq 1}$ be i.i.d. Uniform($[0,1]$) RVs and let $f : [0,1] \to \mathbb{R}$ be a continuous function. For each $n \geq 1$, let

$$I_n = \frac{1}{n}\left(f(X_1) + f(X_2) + \cdots + f(X_n)\right). \tag{42}$$

(i) Suppose $\int_0^1 |f(x)|\,dx < \infty$. Show that $I_n \to I := \int_0^1 f(x)\,dx$ in probability.
(ii) Further assume that $\int_0^1 |f(x)|^2\,dx < \infty$. Use Chebyshef's inequality to show that

$$\mathbb{P}\left(|I_n - I| \geq a/\sqrt{n}\right) \leq \frac{\mathrm{Var}(f(X_1))}{a^2} = \frac{1}{a^2}\left(\int_0^1 f(x)^2\,dx - I^2\right). \tag{43}$$

Exercise 3.6. Let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. Exp(λ) RVs. Define $Y_n = \min(X_1, X_2, \cdots, X_n)$.
(i) For each $\varepsilon > 0$, show that $\mathbb{P}(|Y_n - 0| > \varepsilon) = e^{-\lambda\varepsilon n}$.
(ii) Conclude that $Y_n \to 0$ in probability as $n \to \infty$.

Example 3.7. For each integer $n \geq 1$, define a RV X_n by

$$X_n = \begin{cases} n & \text{with prob. } 1/n \\ 1/n & \text{with prob. } 1 - 1/n. \end{cases} \tag{44}$$

Then $X_n \to 0$ in probability as $n \to \infty$. Indeed, for each $\varepsilon > 0$,

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = 1/n \tag{45}$$

for all $n > 1/\varepsilon$. Hence $\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = 0$. However, note that

$$\mathbb{E}[X_n] = 1 + n^{-1} - n^{-2} \to 1 \quad \text{as } n \to \infty. \tag{46}$$

This example indicates that convergence in probability only cares about probability of the event $\mathbb{P}(|X_n - \mathbb{E}[X_n]| > \varepsilon)$ but not the actual value of X_n when that 'bad' event occurs. ▲

Example 3.8 (Winning strategy)**.** Consider gambling in Vegas for a simple game: after each fair coin flip, one gains twice the bet if heads or loses if tails. There is a simple, always-winning strategy that is banned by Vegas: Double the bet every time you lose and stop after you win the first time.

To analyze this strategy, let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. Uniform($\{-1,1\}$) variables. Let τ be the first time that $X_n = 1$, the first time you win the game. Suppose you start betting \$1 at the first game, and let Y be your net gain after you exit the game. Since you are doubling the bet every time you lose and you gain twice the bet on your first win,

$$Y = -1 - 2^1 - 2^2 - \cdots - 2^{\tau-1} + 2^{\tau+1}. \tag{47}$$

Then

$$\mathbb{E}[Y \mid \tau = t] = -(1 + 2^1 + 2^2 + \cdots + 2^{t-1}) + 2^{t+1} = -(2^t - 1) + 2^{t+1} = 2^t + 1. \tag{48}$$

Hence by iterated expectation,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid \tau]] = \mathbb{E}[(2^\tau + 1)] = \sum_{t=1}^{\infty} (2^t + 1)\mathbb{P}(\tau = t) = \sum_{t=1}^{\infty} (2^t + 1)\frac{1}{2^{t-1}}\frac{1}{2} = \sum_{t=1}^{\infty} 1 + 2^{-t} = \infty. \tag{49}$$

So your expected gain is infinity! ▲

Example 3.9 (Coupon collector's problem)**.** Let $(X_t)_{\geq t}$ be a sequence of i.i.d. Uniform($\{1,2,\cdots,n\}$) variables. Think of the value of X_t as the label of the coupon you collect at tth trial. We are interested in how many times we need to reveal a new random coupon to collect a full set of n distinct coupons. That is, let

$$\tau^n = \min\{r \geq 1 \,|\, \#\{X_1, X_2, \cdots, X_r\} = n\}. \tag{50}$$

Because of the possible overlap, we expect n reveals should not get us the full set of n coupons. Indeed,

$$\mathbb{P}(\tau^n = n) = \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{1}{n} = \frac{n!}{n^n}. \tag{51}$$

Certainly this probability rapidly goes to zero as $n \to \infty$. So we need to reveal more than n coupons. But how many? The answer turns out to be $\tau^n \approx n \log n$. More precisely,

$$\frac{\tau^n}{n \log n} \to 1 \quad \text{as } n \to \infty \text{ in probability.} \tag{52}$$

A change of perspective might help us. Instead of waiting to collect all n coupons, let's progressively collect k distinct coupons for $k = 1$ to n. Namely, for each $1 \leq k \leq n$, define

$$\tau^k = \min\{r \geq 1 \,|\, \#\{X_1, X_2, \cdots, X_r\} = k\}. \tag{53}$$

So τ^k is the first time that we collect k distinct coupons.

Now consider what has to happen to collect $k+1$ distinct coupons from k distinct coupons? Here is an example. Say at time τ^2 we have coupons $\{1,3\}$. τ^3 is the first time that we pick up a new coupon from except 1 and 3. This happens with probability $(n-2)/n$ and since each draw is i.i.d.,

$$\tau^3 - \tau^2 \sim \text{Geom}\left(\frac{n-2}{n}\right). \tag{54}$$

A similar reasoning shows

$$\tau^{k+1} - \tau^k \sim \text{Geom}\left(\frac{n-k}{n}\right). \tag{55}$$

So starting from the first coupon, we wait a Geom($1/n$) time to get a new coupon, and wait a Geom($2/n$) time to get another new coupon, and so on. Note that these geometric waiting times are all independent. So we can decompose τ^n into a sum of independent geometric RVs:

$$\tau^n = \sum_{k=1}^{n-1} (\tau^{k+1} - \tau^k). \tag{56}$$

Then using the estimates in Exercise 3.10, it is straightforward to show that

$$\mathbb{E}[\tau^n] \approx n \log n, \quad \text{Var}(\tau^n) \leq n^2. \tag{57}$$

In Exercise 3.11, we will show (52) using Chebyshef's inequality. ▲

Exercise 3.10. In this exercise, we estimate some partial sums using integral comparison.

(i) For any integer $d \geq 1$, show that

$$\sum_{k=2}^{n} \frac{1}{k^d} \leq \int_1^n \frac{1}{x^d}\, dx \leq \sum_{k=1}^{n-1} \frac{1}{k^d} \tag{58}$$

by considering the upper and lower sum for the Riemann integral $\int_1^n x^{-d}\, dx$.

(ii) Show that

$$\log n \leq \sum_{k=1}^{n-1} \frac{1}{k} \leq 1 + \log(n-1). \tag{59}$$

(iii) Show that for all $d \geq 2$,

$$\sum_{k=1}^{n-1} \frac{1}{k^d} \leq \sum_{k=1}^{\infty} \frac{1}{k^d} \leq 1 + \int_1^{\infty} \frac{1}{x^d}\, dx \leq 2. \tag{60}$$

Exercise 3.11. For each $n \geq 1$, let $X_{1,n}, X_{2,n}, \cdots, X_{n,n}$ be a sequence of independent geometric RVs where $X_{k,n} \sim \text{Geom}((n-k)/n)$. Define $\tau^n = X_{1,n} + X_{2,n} + \cdots + X_{n,n}$.

(i) Show that $\mathbb{E}[\tau^n] = n\sum_{k=1}^{n-1} k^{-1}$. Using Exercise 3.10 (ii), deduce that

$$n \log n \leq \mathbb{E}[\tau^n] \leq n \log(n-1) + n. \tag{61}$$

(ii) Using $\text{Var}(\text{Geom}(p)) = (1-p)/p^2 \leq p^{-2}$ and Exercise 3.10 (iii), show that

$$\text{Var}(\tau^n) \leq n^2 \sum_{k=1}^{n-1} k^{-2} \leq 2n^2. \tag{62}$$

(iii) By Chebyshef's inequality, show that for each $\varepsilon > 0$,

$$\mathbb{P}\left(|\tau^n - \mathbb{E}[\tau^n]| > \varepsilon n \log n\right) \leq \frac{\text{Var}(\tau^n)}{\varepsilon^2 n^2 \log^2 n} \leq \frac{2}{\varepsilon^2 \log^2 n}. \tag{63}$$

Conclude that

$$\frac{\tau^n - \mathbb{E}[\tau^n]}{n \log n} \to 0 \quad \text{as } n \to \infty \text{ in probability.} \tag{64}$$

(iv) By using part (i), conclude that

$$\frac{\tau^n}{n \log n} \to 1 \quad \text{as } n \to \infty \text{ in probability.} \tag{65}$$

4. Central limit theorem

Let $(X_t)_{t\geq 0}$ be a sequence of i.i.d. RVs with finite mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ for $n \geq 1$. We have calculated the mean and variance of the sample mean S_n/n:

$$\mathbb{E}[S_n/n] = \mu, \quad \text{Var}(S_n/n) = \sigma^2/n. \tag{66}$$

Since $\text{Var}(S_n/n) \to 0$ as $n \to \infty$, we expect the sequence of RVs S_n/n to converge its mean μ in probability.

Central limit theorem is a limit theorem for the sample mean with different regime, namely, it describes the 'fluctuation' of the sample mean around its expectation, as $n \to \infty$. For this purpose, we need to standardize the sample mean so that the mean is zero and variance is unit. Namely, let

$$Z_n = \frac{S_n/n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \tag{67}$$

so that

$$\mathbb{E}[Z_n] = 0, \quad \text{Var}(Z_n) = 1. \tag{68}$$

Since the variance is kept at 1, we should not expect the sequence of RVs $(Z_n)_{n\geq 0}$ converge to some constant in probability, as in the law of large number situation. Instead, Z_n should converge to some other RV, if it ever converges in some sense. Central limit theorem states that Z_n converges to the standard normal RV $Z \sim N(0,1)$ 'in distribution'.

Let us state the central limit theorem (Theorem 1.3).

Theorem 4.1 (CLT)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. Suppose $\mathbb{E}[X_1] < \infty$ and $\mathbb{E}[X_1^2] = \sigma^2 < \infty$. Let $Z \sim N(0,1)$ be a standard normal RV and define*

$$Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}}. \tag{69}$$

Then Z_n converges to Z as $n \to \infty$ in distribution, namely,

$$\lim_{n\to} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z). \tag{70}$$

Proof. First notice that we can assume $\mathbb{E}[X_1] = 0$ without loss of generality (why?). Then $\sigma^2 = \mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \mathbb{E}[X_1^2]$. Our proof is based on computing moment generating function of Z_n, and we will show that this converges to the MGF of the standard normal. (This is why we learned moment generating function.)

Since the increments X_i's are i.i.d., we have

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}]\cdots\mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX_1}]^n. \tag{71}$$

Since we are assuming $\mathbb{E}[X_1] = 0$ and $\mathbb{E}(X_1^2) < \infty$, we have

$$\mathbb{E}[e^{tX_1}] = 1 + \frac{\sigma^2}{2}t^2 + O(t^3), \tag{72}$$

where $O(t^3)$ contains the rest of terms of order $t \geq 3$. Hence

$$\mathbb{E}[e^{tS_n}] = \left(1 + \frac{\sigma^2}{2}t^2 + O(t^3)\right)^n. \tag{73}$$

This yields

$$\mathbb{E}[e^{tZ_n}] = \mathbb{E}[e^{tS_n/(\sigma\sqrt{n})}] = \mathbb{E}[e^{(t/\sigma\sqrt{n})S_n}] \tag{74}$$

$$= \left(1 + \frac{t^2}{2n} + O(n^{-3/2})\right)^n. \tag{75}$$

Recall that

$$\lim_{n\to\infty} \left(1 + \frac{t^2}{2n}\right)^n = e^{t^2/2} = \mathbb{E}[e^{tZ}]. \tag{76}$$

Since the $O(n^{-3/2})$ term vanishes as $n \to \infty$, this shows that $\mathbb{E}[e^{tZ_n}] \to \mathbb{E}[e^{tZ}]$ as $n \to \infty$. Since MGFs determine distribution of RVs (see Theorem 4.14 (ii) in Note 4), it follows that the CDF of Z_n converges to that of Z. □

As a typical application of CLT, we can approximate Binomial(n,p) variables by normal RVs.

Exercise 4.2. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. Poisson(λ) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - n\lambda)/\sqrt{n\lambda}$. Show that as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim \text{Poisson}(n\lambda)$, then

$$\frac{Y_n - n\lambda}{\sqrt{n\lambda}} \Rightarrow Z \sim N(0,1). \tag{77}$$

(iii) From (ii) deduce that we have the following approximation

$$\mathbb{P}(Y_n \le x) \approx \mathbb{P}\left(Z \le \frac{x - n\lambda}{\sqrt{n\lambda}}\right), \tag{78}$$

which becomes more accurate as $n \to \infty$.

Exercise 4.3. Let $(X_n)_{n\ge 1}$ be a sequence of i.i.d. Bernoulli(p) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - np)/\sqrt{np(1-p)}$. Show that as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim \text{Binomial}(n,p)$, then

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \Rightarrow Z \sim N(0,1). \tag{79}$$

(iii) From (ii), deduce that have the following approximation

$$\mathbb{P}(Y_n \le x) \approx \mathbb{P}\left(Z \le \frac{x - np}{\sqrt{np(1-p)}}\right), \tag{80}$$

which becomes more accurate as $n \to \infty$.

Example 4.4 (Polling revisited)**.** Let $(X_n)_{n\ge 1}$ be a sequence of i.i.d. Bernoulli(p) RVs. Denote $\hat{p}_n = n^{-1}(X_1 + \cdots + X_n)$. In Exercise 3.4, we used Chebyshef's inequality to deduce that

$$\mathbb{P}(|\hat{p}_n - p| \le 0.01) \ge 0.95 \tag{81}$$

whenever $n \ge 50,000$. In this example, we will use CLT to improve this lower bound on n.

First, from Exercise 4.3, it is immediate to deduce the following convergence in distribution

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \Rightarrow Z \sim N(0,1). \tag{82}$$

Hence for any $\varepsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(|\hat{p}_n - p| \le \varepsilon\right) &= \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) && (83)\\
&\ge \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le 2\varepsilon\sqrt{n}\right) && (84)\\
&\approx \mathbb{P}\left(|Z| \le 2\varepsilon\sqrt{n}\right) = 2\mathbb{P}(0 \le Z \le 2\varepsilon\sqrt{n}), && (85)
\end{aligned}$$

where for the inequality we have used the fact that $p(1-p) \le 1/4$ for all $0 \le p \le 1$. The last expression is at least 0.95 if and only if

$$\mathbb{P}(0 \le Z \le 2\varepsilon\sqrt{n}) \ge 0.475. \tag{86}$$

From the table of standard normal distribution, we know that $\mathbb{P}(0 \le Z \le 1.96) = 0.475$. Hence (86) holds if and only if $2\varepsilon\sqrt{n} \ge 1.96$, or equivalently,

$$n \ge \left(\frac{0.98}{\varepsilon}\right)^2. \tag{87}$$

For instance, $\varepsilon = 0.01$ gives $n \ge 9604$. This is a drastic improvement from $n \ge 50,000$ via Chebyshef.

Exercise 4.5. Let $X_1, Y_1, \cdots, X_n, Y_n$ be i.i.d. Uniform($[0,1]$) RVs. Let

$$W_n = \frac{(X_1 + \cdots + X_n) - (Y_1 + \cdots + Y_n)}{n}. \tag{88}$$

Find a numerical approximation to the quantity

$$\mathbb{P}(|W_{20} - \mathbb{E}[W_{20}]| < 0.001). \tag{89}$$

5. The SLLN and almost sure convergence

Let $(X_n)_{n\ge1}$ be i.i.d. RVs with finite mean $\mathbb{E}[X_1] = \mu$ and let $S_n = X_1 + \cdots + X_n$ for all $n \ge 1$. The weak law of large numbers states that the sample mean S_n/n converges to μ in probability, that is,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0 \quad \forall \varepsilon > 0. \tag{90}$$

On the other hand, the Strong Law of Large Numbers (SLLN) tells us that a similar statement holds where the limit is inside the probability bracket. Namely,

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0 \quad \forall \varepsilon > 0. \tag{91}$$

If we view the limit on the left hand side as a RV, then (91) in fact states that this limit RV is 0 with probability 1:

$$\mathbb{P}\left(\lim_{n\to\infty} \left|\frac{S_n}{n} - \mu\right| = 0\right) = 1. \tag{92}$$

This is equivalent to the following familiar form of SLLN in Theorem 1.2:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1. \tag{93}$$

Definition 5.1. Let $(X_n)_{n\ge1}$ be a sequence of RVs and let a be a real number. We say that X_n converges to a *almost surely* (or *with probability* 1) if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = a\right) = 1. \tag{94}$$

Example 5.2. In this example, we will see that convergence in probability does not necessarily imply convergence with probability 1. Define a sequence of RVs $(T_n)_{n\ge1}$ as follows. Let $T_1 = 1$, and $T_2 \sim$ Uniform($\{2,3\}$), $T_3 \sim$ Uniform($\{4,5,6\}$), and so on. In general, $X_k \sim$ Uniform$\{(k-1)k/2, \cdots, k(k+1)/2\}$ for all $k \ge 2$. Let $X_n = \mathbf{1}$(some T_k takes value n). Think of

$$T_n = n\text{th arrival time of customers} \tag{95}$$

$$X_n = \mathbf{1}(\text{some customer arrives at time } n). \tag{96}$$

Then note that

$$\mathbb{P}(X_1 = 1) = 1, \tag{97}$$

$$\mathbb{P}(X_2 = 1) = \mathbb{P}(X_3 = 1) = 1/2, \tag{98}$$

$$\mathbb{P}(X_4 = 1) = \mathbb{P}(X_5 = 1) = \mathbb{P}(X_6 = 1) = 1/3, \tag{99}$$

and so on. Hence it is clear that $\lim_{n\to\infty} \mathbb{P}(X_n = 1) = 0$. Since X_n is an indicator variable, this yields that $\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = 0$ for all $\varepsilon > 0$, that is, X_n converges to 0 in probability. On the other hand, $X_n \to 0$ a.s. means $\mathbb{P}(\lim_{n\to\infty} X_n = 0)$, which implies that $X_n = 0$ for all but finitely many n's. However, $X_n = 1$ for infinitely many n's since customer always arrive after any large time N. Hence X_n cannot converge to 0 almost surely. ▲

Exercise 5.3. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let a be a real number. Suppose X_n converges to a with probability 1.

(i) Show that

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \leq \varepsilon\right) = 1 \quad \forall \varepsilon > 0. \tag{100}$$

(ii) Fix $\varepsilon > 0$. Let A_k be the event that $|X_n - a| \leq \varepsilon$ for all $n \geq k$. Show that $A_1 \subseteq A_2 \subseteq \cdots$ and

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \leq \varepsilon\right) \leq \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right). \tag{101}$$

(iii) Show that for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - a| \leq \varepsilon\right) \geq \lim_{n\to\infty} \mathbb{P}\left(A_n\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \geq \mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \leq \varepsilon\right) = 1. \tag{102}$$

Conclude that $X_n \to a$ in probability.

A typical tool for proving convergence with probability 1 is the following.

Exercise 5.4 (Borel-Cantelli lemma)**.** Let $(A_n)_{n\geq 1}$ be a sequence of events such that

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty. \tag{103}$$

We will show that

$$\mathbb{P}\left(A_n \text{ occurs only for finitely many } n\text{'s}\right) = 1. \tag{104}$$

(i) Let $N = \sum_{n=1}^{\infty} \mathbf{1}(A_n)$, which is the number of n's such that A_n occurs. Use Fubini's theorem to show that

$$\mathbb{E}[N] = \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbf{1}(A_n)\right] = \sum_{n=1}^{\infty} \mathbb{E}[\mathbf{1}(A_n)] = \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty. \tag{105}$$

(ii) Deduce that the RV N must not take ∞ with positive probability. Hence $\mathbb{P}(N < \infty) = 1$, as desired.

Exercise 5.5. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and fix $x \in \mathbb{R}$. We will show that $X_n \to x$ a.s. if the tail probabilities are 'summable'. (This is the typical application of the Borel-Cantelli lemma.)

(i) Fix $\varepsilon > 0$. Suppose $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - x| > \varepsilon) < \infty$. Use Borel-Cantelli lemma to deduce that $|X_n - x| > \varepsilon$ for only finitely many n's.

(ii) Conclude that, if $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - x| > \varepsilon) < \infty$ for all $\varepsilon > 0$, then $X_n \to x$ a.s.

Example 5.6. Let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. Exp(λ) RVs. Define $Y_n = \min(X_1, X_2, \cdots, X_n)$. Recall that in Exercise 3.6, we have shown that

$$\mathbb{P}(|Y_n - 0| > \varepsilon) = e^{-\lambda \varepsilon n}. \tag{106}$$

for all $\varepsilon > 0$ and that $Y_n \to 0$ in probability as $n \to \infty$. In fact, $Y_n \to 0$ with probability 1. To see this, we note that, for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|Y_n - 0| > \varepsilon) = \sum_{n=1}^{\infty} e^{-\lambda \varepsilon n} = \frac{e^{-\lambda \varepsilon}}{1 - e^{\lambda \varepsilon}} < \infty. \tag{107}$$

By Borel-Cantelli lemma (or Exercise 5.5), we conclude that $Y_n \to 0$ a.s. ▲

Example 5.7. Let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. Uniform([0, 1]) RVs.

(i) We show that $X_n^{1/n}$ converges to 1 almost surely, as $n \to \infty$. Fix any $\epsilon > 0$. Since $X_n \geq 1$,

$$\begin{aligned}\mathbb{P}(|(X_n)^{1/n} - 1| > \epsilon) &= \mathbb{P}\big((X_n)^{1/n} > (1+\epsilon) \text{ or } (X_n)^{1/n} < (1-\epsilon)\big) = \mathbb{P}((X_n)^{1/n} < (1-\epsilon)) \\ &= \mathbb{P}\big(X_n < (1-\epsilon)^n\big) = (1-\epsilon)^n,\end{aligned} \tag{108}$$

which goes to 0 as $n \to \infty$. Therefore, $X_n^{1/n}$ converges to 1 in probability. However, since $\sum_{n=1}^{\infty} (1-\epsilon)^n < \infty$, we have that $X_n^{1/n}$ also converges to 1 almost surely.

(ii) Define $U_n = \max\{X_1, X_2^2, X_3^3, \ldots, X_{n-1}^{n-1}, X_n^n\}$. We show that the sequence U_n converges in probability to 1. For an $\epsilon < 1$ fixed

$$\begin{aligned}\mathbb{P}(|U_n - 1| \geq \epsilon) = \mathbb{P}(U_n \leq 1 - \epsilon) &= \mathbb{P}(X_1 \leq 1-\epsilon, X_2^2 \leq 1-\epsilon, \ldots, X_n^n \leq 1-\epsilon) && (109)\\ &= \mathbb{P}(X_1 \leq 1-\epsilon)\mathbb{P}(X_2^2 \leq 1-\epsilon)\cdots\mathbb{P}(X_n^n \leq 1-\epsilon) && (110)\\ &= \mathbb{P}(X_1 \leq 1-\epsilon)\mathbb{P}(X_2 \leq (1-\epsilon)^{1/2})\cdots\mathbb{P}(X_n \leq (1-\epsilon)^{1/n}) && (111)\\ &= (1-\epsilon)\cdot(1-\epsilon)^{1/2}\cdots(1-\epsilon)^{1/n} = (1-\epsilon)^{1+1/2+\cdots+1/n} \to 0, && (112)\end{aligned}$$

since $1 + 1/2 + \cdots + 1/n \to \infty$, as $n \to \infty$.

(iii) Define $V_n = \max\{X_1, X_2^{2^2}, X_3^{3^2}, \ldots, X_{n-1}^{(n-1)^2}, X_n^{n^2}\}$. Does the sequence V_n converges in probability to 1? Similarly as in the previous part, for a fixed $\epsilon < 1$ we have

$$\mathbb{P}(|U_n - 1| \geq \epsilon) = (1-\epsilon)^{1+1/2^2+\cdots+1/n^2}, \tag{113}$$

which doesn't converge to zero (but to a positive number), since $1 + 1/2^2 + \cdots + 1/n^2$ is a convergent series. Therefore, V_n doesn't converge to zero in probability. Hence it also doesn't converge to 0 a.s.

▲

Now we prove the strong law of large numbers. The proof of full statement (Theorem 1.2) with finite second moment assumption has extra technicality, so here we prove the result under a stronger assumption of finite fourth moment.

Theorem 5.8 (SLLN with fourth moment)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. RVs such that $\mathbb{E}[X_n^4] < \infty$. Let $S_n = X_1 + \cdots + X_n$ for all $n \geq 1$. Then S_n/n converges to $\mathbb{E}[X_1]$ with probability 1.*

Proof. Our aim is to show that

$$\sum_{n=1}^{\infty} \mathbb{E}[(S_n/n)^4] < \infty. \tag{114}$$

Then by Borel-Cantelli lemma, $(S_n/n)^4$ converges to 0 with probability 1. Hence S_n/n converges to 0 with probability 1, as desired.

For a preparation, we first verify that we have finite first and second moments for X_1. It is easy to verify the inequality $|x| \leq 1 + x^4$ for all $x \in \mathbb{R}$, so we have

$$\mathbb{E}[|X_1|] \leq 1 + \mathbb{E}[X_1^4] < \infty. \tag{115}$$

Hence $\mathbb{E}[X_1]$ exists. By shifting, we may assume that $\mathbb{E}[X_1] = 0$. Similarly, it holds that $x^2 \leq c + x^4$ for all $x \in \mathbb{R}$ if $c > 0$ is large enough. Hence $\mathbb{E}[X_1^2] < \infty$.

Note that

$$\mathbb{E}\left[S_n^4\right] = \mathbb{E}\left[\left(\sum_{k=1}^{n} X_k\right)^4\right] = \mathbb{E}\left[\sum_{1\leq i,j,k,\ell\leq n} X_i X_j X_k X_\ell\right] = \sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}\left[X_i X_j X_k X_\ell\right]. \tag{116}$$

Note that by independence and the assumption that $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_i X_j X_k X_\ell] = 0$ if at least one of the four indices does not repeat. For instance,

$$\mathbb{E}[X_1 X_2^3] = \mathbb{E}[X_1]\mathbb{E}[X_2^3] = 0, \tag{117}$$

$$\mathbb{E}[X_1 X_2^2 X_3] = \mathbb{E}[X_1]\mathbb{E}[X_2^2]\mathbb{E}[X_3] = 0. \tag{118}$$

Hence by collecting terms based on number of overlaps, we have

$$\sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}\left[X_i X_j X_k X_\ell\right] = \sum_{i=1}^{n} \mathbb{E}[X_i^4] + \binom{4}{2} \sum_{1\leq i<j\leq n} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] \tag{119}$$

$$= n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2. \tag{120}$$

Thus for all $n \geq 1$,

$$\mathbb{E}\left[(S_n/n)^4\right] = \frac{n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2}{n^4} \leq \frac{n^2\mathbb{E}[X_1^4] + 3n^2\mathbb{E}[X_1^2]^2}{n^4} = \frac{\mathbb{E}[X_1^4] + 3\mathbb{E}[X_1^2]}{n^2}. \tag{121}$$

Summing over all n, this gives

$$\sum_{n=1}^{\infty} \mathbb{E}[(S_n/n)^4] \leq (\mathbb{E}[X_1^4] + 3\mathbb{E}[X_1^2]) \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty. \tag{122}$$

Hence by Borell-Cantelli lemma, we conclude that $(S_n/n)^4$ converges to 0 with probability 1. The same conclusion holds for S_n/n. This shows the assertion. □