PROBABILITY THEORY - PART 2
INDEPENDENT RANDOM VARIABLES

MANJUNATH KRISHNAPUR

Contents

1. Introduction

In this second part of the course, we shall study independent random variables. Much of what we do is devoted to the following single question: Given independent random variables with known distributions, what can you say about the distribution of the sum? In the process of finding answers, we shall weave through various topics. Here is a guide to the essential aspects that you might pay attention to.

Firstly, the results. We shall cover fundamental limit theorems of probability, such as the weak and strong law of large numbers, central limit theorems, poisson limit theorem, in addition to results on random series with independent summands. We shall also talk about the various modes of convergence of random variables.

The second important aspect will be the various techniques. These include the first and second moment methods, Borel-Cantelli lemmas, zero-one laws, inequalities of Chebyshev and Bernstein and Hoeffding, Kolmogorov's maximal inequality. In addition, we mention the outstandingly useful tool of characteristic functions as well as the less profound but very common and useful techniques of proofs such as truncation and approximation.

Thirdly, we shall try to introduce a few basic problems/constructs in probability that are of interest in themselves and that appear in many guises in all sorts of probability problems. These include the coupon collector problem, branching processes, Pólya's urn scheme and Brownian motion. Many more could have been included if there was more time[1].

2. Some basic tools in probability

We collect three basic tools in this section. Their usefulness cannot be overstated.

2.1. First and second moment methods.

In popular language, average value is often mistaken for typical value. This is not always correct, for example, in many populations, a typical person has much lower income than the average (because a few people have a large fraction of the total wealth). For a mathematical example, suppose $X = 10^6$ with probability 10^{-3} and $X = 0$ with probability $1 - 10^{-3}$. Then $\mathbf{E}[X] = 1000$ although with probability 0.999 its value is zero. Thus the typical value is close to zero.

Since it is often easier to calculate expectations and variances (for example, expectation of a sum is sum of expectations) than to calculate probabilities (example, tail probability of a sum of random variables), the following inequalities that bound certain probabilities in terms of moments may be expected to be somewhat useful. In fact, they are extremely useful as we shall shortly see!

Lemma 1. *Let $X \geq 0$ be a r.v.*

*(1) (**Markov's inequality or first moment method**). For any $t > 0$, we have $\mathbf{P}(X \geq t) \leq t^{-1}\mathbf{E}[X]$.*

[1] References: Dudley's book is an excellent source for the first aspect and some of the second but does not have much of the third. Durrett's book is excellent in all three, especially the third, and has way more material than we can touch upon in this course. Lots of other standard books in probability have various non-negative and non-positive features.

(2) (***Paley-Zygmund inequality or second moment method***). *For any non-negative r.v. X, and any $0 \leq \alpha \leq 1$, we have*

$$\mathbf{P}\left(X > \alpha \mathbf{E}[X]\right) \geq (1-\alpha)^2 \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]}.$$

In particular, $\mathbf{P}\left(X > 0\right) \geq \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]}$.

Proof. (1) For any $t > 0$, clearly $t\mathbf{1}_{X \geq t} \leq X$. Positivity of expectations gives the inequality.

(2) $\mathbf{E}[X]^2 = \mathbf{E}[X\mathbf{1}_{X>0}]^2 \leq \mathbf{E}[X^2]\mathbf{E}[\mathbf{1}_{X>0}] = \mathbf{E}[X^2]\mathbf{P}(X > 0)$. Hence the second inequality follows. The first one is similar. Let $\mu = \mathbf{E}[X]$. By Cauchy-Schwarz, we have $\mathbf{E}[X\mathbf{1}_{X>\alpha\mu}]^2 \leq \mathbf{E}[X^2]\mathbf{P}(X > \alpha\mu)$. Further, $\mu = \mathbf{E}[X\mathbf{1}_{X<\alpha\mu}] + \mathbf{E}[X\mathbf{1}_{X>\alpha\mu}] \leq \alpha\mu + \mathbf{E}[X\mathbf{1}_{X>\alpha\mu}]$, whence, $\mathbf{E}[X\mathbf{1}_{X>\alpha\mu}] \geq (1-\alpha)\mu$. Thus,

$$\mathbf{P}(X > \alpha\mu) \geq \frac{\mathbf{E}[X\mathbf{1}_{X>\alpha\mu}]^2}{\mathbf{E}[X^2]} \geq (1-\alpha)^2 \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]}. \qquad \blacksquare$$

Remark 2. *Applying these inequalities to other functions of X can give more information. For example, if X has finite variance, $\mathbf{P}(|X - \mathbf{E}[X]| \geq t) = \mathbf{P}(|X - \mathbf{E}[X]|^2 \geq t^2) \leq t^{-2}Var(X)$, which is called* Chebyshev's inequality*. Higher the moments that exist, better the asymptotic tail bounds that we get. For example, if $\mathbf{E}[e^{\lambda X}] < \infty$ for some $\lambda > 0$, we get exponential tail bounds by $\mathbf{P}(X > t) = \mathbf{P}(e^{\lambda X} > e^{\lambda t}) \leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$. Note that X is not assumed to be non-negative in these examples as Markov's inequality is applied to the non-negative random variables $(X - \mathbf{E}[X])^2$ and $e^{\lambda X}$.*

In the next section we shall give several applications of the first and second moment methods. Now we give two other results of ubiquitous use. Their use comes from the fact that most often probabilists are engaged in showing that an event has probability zero or one!

2.2. **Borel-Cantelli lemmas.** If A_n is a sequence of events in a common probability space, $\limsup A_n$ consists of all ω that belong to infinitely many of these events. Probabilists often write the phrase "A_n infinitely often" (or "A_n i.o" in short) to mean $\limsup A_n$.

Lemma 3 (Borel Cantelli lemmas). *Let A_n be events on a common probability space.*

(1) *If $\sum_n \mathbf{P}(A_n) < \infty$, then $\mathbf{P}(A_n$ infinitely often$) = 0$.*

(2) *If A_n are independent and $\sum_n \mathbf{P}(A_n) = \infty$, then $\mathbf{P}(A_n$ infinitely often$) = 1$.*

Proof. (1) For any N, $\mathbf{P}\left(\cup_{n=N}^{\infty} A_n\right) \leq \sum_{n=N}^{\infty} \mathbf{P}(A_n)$ which goes to zero as $N \to \infty$. Hence $\mathbf{P}(\limsup A_n) = 0$.

(2) For any $N < M$, $\mathbf{P}(\cup_{n=N}^{M} A_n) = 1 - \prod_{n=N}^{M} \mathbf{P}(A_n^c)$. Since $\sum_n \mathbf{P}(A_n) = \infty$, it follows that $\prod_{n=N}^{M}(1 - \mathbf{P}(A_n)) \leq \prod_{n=N}^{M} e^{-\mathbf{P}(A_n)} \to 0$, for any fixed N as $M \to \infty$. Therefore, $\mathbf{P}\left(\cup_{n=N}^{\infty} A_n\right) = 1$ for all N, implying that $\mathbf{P}(A_n$ i.o.$) = 1$. $\blacksquare$

We shall give another proof later, using the first and second moment methods. It will be seen then that pairwise independence is sufficient for the second Borel-Cantelli lemma!

2.3. **Kolmogorov's zero-one law.** If $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, the set of all events that have probability equal to 0 or to 1 form a sigma algebra. Zero-one laws are theorems that (in special situations) identify specific sub-sigma-algebras of this. Such σ-algebras (and events within them) are sometimes said to be *trivial*. An equivalent statement is that all random variables measurable with respect to such a sigma algebra are constants $a.s.$

Definition 4. Let $(\Omega, \mathcal{F})$ be a measurable space and let $\mathcal{F}_n$ be sub-sigma algebras of $\mathcal{F}$. Then the tail σ-algebra of the sequence $\mathcal{F}_n$ is defined to be $\mathcal{T} := \cap_n \sigma\left(\cup_{k\geq n}\mathcal{F}_k\right)$. For a sequence of random variables $X_1, X_2, \ldots$, the tail sigma algebra (also denoted $\mathcal{T}(X_1, X_2, \ldots)$) is the tail of the sequence $\sigma(X_n)$.

How to think of it? If A is in the tail of $(X_k)_{k\geq 1}$, then $A \in \sigma(X_n, X_{n+1}, \ldots)$ for any n. That is, the tail of the sequence is sufficient to tell you whether the event occurred or not. For example, A could be the event that infinitely many X_k are positive.

Theorem 5 (Kolmogorov's zero-one law). *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.*

(1) If $\mathcal{F}_n$ is a sequence of independent sub-sigma algebras of $\mathcal{F}$, then the tail σ-algebra is trivial.

(2) If X_n are independent random variables, and A is a tail event, then $\mathbf{P}(A) = 0$ or $\mathbf{P}(A) = 1$.

Proof. The second statement follows immediately from the first. To prove the first, define $\mathcal{T}_n := \sigma\left(\cup_{k>n}\mathcal{F}_k\right)$. Then, $\mathcal{F}_1, \ldots, \mathcal{F}_n, \mathcal{T}_n$ are independent. Since $\mathcal{T} \subseteq \mathcal{T}_n$, it follows that $\mathcal{F}_1, \ldots, \mathcal{F}_n, \mathcal{T}$ are independent. Since this is true for every n, we see that $\mathcal{T}, \mathcal{F}_1, \mathcal{F}_2, \ldots$ are independent. Hence, $\mathcal{T}$ and $\sigma\left(\cup_n \mathcal{F}_n\right)$ are independent. But $\mathcal{T} \subseteq \sigma\left(\cup_n \mathcal{F}_n\right)$, hence, $\mathcal{T}$ is independent of itself. This implies that for any $A \in \mathcal{T}$, we must have $\mathbf{P}(A)^2 = \mathbf{P}(A \cap A) = \mathbf{P}(A)$ which forces $\mathbf{P}(A)$ to be 0 or 1. ■

Independence is crucial (but observe that X_k need not be identically distributed). If $X_k = X_1$ for all k, then the tail sigma-algebra is the same as $\sigma(X_1)$ which is not trivial unless X_1 is constant $a.s.$ As a more non-trivial example, let ξ_k, $k \geq 1$ be i.i.d. $N(0.1, 1)$ and let $\eta \sim \text{Ber}_{\pm}(1/2)$. Set $X_k = \eta\xi_k$. Intuitively it is clear that a majority of ξ_ks are positive. Hence, by looking at $(X_n, X_{n+1}, \ldots)$ and checking whether positive or negatives are in majority, we ought to be able to guess η. In other words, the non-constant random variable η is in the tail of the sequence $(X_k)_{k\geq 1}$.

The following exercise shows how Kolmogorov's zero-one law may be used to get non-trivial conclusions. Another interesting application (but not relevant to the course) will be given in a later section.

Exercise 6. *Let X_i be independent random variables. Which of the following random variables must necessarily be constant almost surely?* $\limsup X_n$, $\liminf X_n$, $\limsup n^{-1}S_n$, $\liminf S_n$.

2.4. **A few other useful tools.** While it is difficult to match the utility of the tools that we have mentioned so far, there are several others that are quite useful in many settings. Here we just mention some of them so that the interested reader is at least aware of their existence. For lack of time to cover good examples, we shall largely omit these in this course.

- *Bernstein-Hoeffding inequalities* give powerful bounds for the probability for a sum of independent random variables to deviate from its mean. We shall cover these in Section 12.
- *The probabilistic method* is the name given to the trivial-sounding statement that $\mathbf{P}(A) > 0$ implies that A is non-empty! Its usefulness comes from actual situations where one wants to construct an object with given properties and the easiest way turns out to be to construct a random object and show that it has the desired properties (with positive probability).
- *Lovász's local lemma* states that if A_i are events, each having probability at most p and each A_i is independent of all except (at most) d of the other events, and if $p(d+1) \le \frac{1}{e}$, then $A_1^c \cap \ldots \cap A_n^c$ has positive probability. This is obvious if A_is are independent (then we only need the condition $\mathbf{P}(A_i) < 1$ for each i). Hence, intuitively one feels that the result should hold if the dependence is weak, and the local lemma is a precise formulation of such a result.
- There are many other *zero-one* laws that are quite useful. We mention two - *ergodicity* (which is covered in the problem set) and *Hewitt-Savage zero-one law* (which we omit altogether).

3. Applications of first and second moment methods

The first and second moment methods are immensely useful. This is somewhat surprising, given the very elementary nature of these inequalities, but the following applications illustrate the ease with which they give interesting results.

3.1. **Borel-Cantelli lemmas.** If X takes values in $\mathbb{R} \cup \{+\infty\}$ and $\mathbf{E}[X] < \infty$ then $X < \infty$ $a.s.$ (if you like you may see it as a consequence of Markov's inequality!). Apply this to $X = \sum_{k=1}^{\infty} \mathbf{1}_{A_k}$ which has $\mathbf{E}[X] = \sum_{k=1}^{\infty} \mathbf{P}(A_k)$ which is given to be finite. Therefore $X < \infty$ $a.s.$ which implies that for $a.e.$ ω, only finitely many $\mathbf{1}_{A_k}(\omega)$ are non-zero. This is the first Borel-Cantelli lemma.

The second one is more interesting. Fix $n < m$ and define $X = \sum_{k=n}^{m} \mathbf{1}_{A_k}$. Then $\mathbf{E}[X] = \sum_{k=n}^{m} \mathbf{P}(A_k)$. Also,

$$\mathbf{E}[X^2] = \mathbf{E}\left[\sum_{k=n}^{m}\sum_{\ell=n}^{m} \mathbf{1}_{A_k}\mathbf{1}_{A_\ell}\right] = \sum_{k=n}^{m} \mathbf{P}(A_k) + \sum_{k\neq\ell} \mathbf{P}(A_k)\mathbf{P}(A_\ell)$$

$$\le \left(\sum_{k=n}^{m} \mathbf{P}(A_k)\right)^2 + \sum_{k=n}^{m} \mathbf{P}(A_k).$$

Apply the second moment method to see that for any fixed n, as $m \to \infty$ (note that $X > 0$ is the same as $X \geq 1$),

$$\begin{aligned}\mathbf{P}(X \geq 1) &\geq \frac{\left(\sum_{k=n}^{m} \mathbf{P}(A_k)\right)^2}{\left(\sum_{k=n}^{m} \mathbf{P}(A_k)\right)^2 + \sum_{k=n}^{m} \mathbf{P}(A_k)} \\ &= \frac{1}{1 + \left(\sum_{k=n}^{m} \mathbf{P}(A_k)\right)^{-1}}\end{aligned}$$

which converges to 1 as $m \to \infty$, because of the assumption that $\sum \mathbf{P}(A_k) = \infty$. This shows that $\mathbf{P}(\cup_{k\geq n} A_k) = 1$ for any n and hence $\mathbf{P}(\limsup A_n) = 1$.

Note that this proof used independence only to claim that $\mathbf{P}(A_k \cap A_\ell) = \mathbf{P}(A_k)\mathbf{P}(A_\ell)$. Therefore, not only did we get a new proof, but we have shown that the second Borel-Cantelli lemma holds for *pairwise independent* events too!

3.2. **Coupon collector problem.** A bookshelf has (a large number) n books numbered $1, 2, \ldots, n$. Every night, before going to bed, you pick one of the books at random to read. The book is replaced in the shelf in the morning. How many days pass before you have picked up each of the books at least once?

Theorem 7. *Let T_n denote the number of days till each book is picked at least once. Then T_n is concentrated around $n \log n$ in a window of size n by which we mean that for any sequence of numbers $\theta_n \to \infty$, we have*

$$\mathbf{P}(|T_n - n\log n| < n\theta_n) \to 1.$$

The proof will proceed by computing the expected value of T_n and then showing that T_n is typically near its expected value.

A very useful elementary inequality: In the following proof and many other places, we shall have occasion to make use of the elementary estimate

$$1 - x \leq e^{-x} \quad \text{for all } x, \qquad 1 - x \geq e^{-x-x^2} \quad \text{for } |x| < \frac{1}{2}. \tag{1}$$

To see the first inequality, observe that $e^{-x} - (1-x)$ is equal to 0 for $x = 0$, has positive derivative for $x > 0$ and negative derivative for $x < 0$. To prove the second inequality, recall the power series expansion $\log(1-x) = -x - x^2/2 - x^3/3 - \ldots$ which is valid for $|x| < 1$. Hence, if $|x| < \frac{1}{2}$, then

$$\begin{aligned}\log(1-x) &\geq -x - x^2 + \frac{1}{2}x^2 - \frac{1}{2}\sum_{k=3}^{\infty} |x|^k \\ &\geq -x - x^2\end{aligned}$$

since $\sum_{k=3}^{\infty} |x|^3 \leq x^2 \sum_{k=3}^{\infty} 2^{-k} \leq \frac{1}{2}x^2$.

Proof of Theorem 7. Fix an integer $t \geq 1$ and let $X_{t,k}$ be the indicator that the k^{th} book is not picked up on the first t days. Then, $\mathbf{P}(T_n > t) = \mathbf{P}(S_{t,n} \geq 1)$ where $S_{t,n} = X_{t,1} + \ldots + X_{t,n}$ is the number of books not yet picked in the first t days. As $\mathbf{E}[X_{t,k}] = (1-1/n)^t$ and $\mathbf{E}[X_{t,k}X_{t,\ell}] = (1-2/n)^t$ for $k \neq \ell$, we also compute that thefirst two moments of $S_{t,n}$ and use (1) to get

$$ne^{-\frac{t}{n}-\frac{t}{n^2}} \leq \mathbf{E}[S_{t,n}] = n\left(1-\frac{1}{n}\right)^t \leq ne^{-\frac{t}{n}}. \tag{2}$$

and

$$\mathbf{E}[S_{t,n}^2] = n\left(1-\frac{1}{n}\right)^t + n(n-1)\left(1-\frac{2}{n}\right)^t \leq ne^{-\frac{t}{n}} + n(n-1)e^{-\frac{2t}{n}}. \tag{3}$$

The left inequality on the first line is valid only for $n \geq 2$ which we assume.

Now set $t = n\log n + n\theta_n$ and apply Markov's inequality to get

$$\mathbf{P}(T_n > n\log n + n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \leq \mathbf{E}[S_{t,n}] \leq ne^{-\frac{n\log n + n\theta_n}{n}} \leq e^{-\theta_n} = o(1). \tag{4}$$

On the other hand, taking $t = n\log n - n\theta_n$ (where we take $\theta_n < \log n$, of course!), we now apply the second moment method. For any $n \geq 2$, by using (3) we get $\mathbf{E}[S_{t,n}^2] \leq e^{\theta_n} + e^{2\theta_n}$. The first inequality in (2) gives $\mathbf{E}[S_{t,n}] \geq e^{\theta_n - \frac{\log n - \theta_n}{n}}$. Thus,

$$\mathbf{P}(T_n > n\log n - n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \geq \frac{\mathbf{E}[S_{t,n}]^2}{\mathbf{E}[S_{t,n}^2]} \geq \frac{e^{2\theta_n - 2\frac{\log n - \theta_n}{n}}}{e^{\theta_n} + e^{2\theta_n}} = 1 - o(1) \tag{5}$$

as $n \to \infty$. From (4) and (5), we get the sharp bounds

$$\mathbf{P}\left(|T_n - n\log(n)| > n\theta_n\right) \to 0 \text{ for any } \theta_n \to \infty. \qquad \blacksquare$$

Here is an alternate approach to the same problem. It brings out some other features well. But we shall use elementary conditioning and appeal to some intuitive sense of probability.

Alternate proof of Theorem 7. Let $\tau_1 = 1$ and for $k \geq 2$, let τ_k be the number of draws after $k-1$ distinct coupons have been seen till the next new coupon appears. Then, $T_n = \tau_1 + \ldots + \tau_n$.

We make two observations about τ_ks. Firstly, they are independent random variables. This is intuitively clear and we invite the reader to try writing out a proof from definitions. Secondly, the distribution of τ_k is Geo($\frac{n-k+1}{n}$). This is so since, after having seen $(k-1)$ coupons, in every draw, there is a chance of $(n-k+1)/n$ to see a new (unseen) coupon.

If $\xi \sim$ Geo(p) (this means $\mathbf{P}(\xi = k) = p(1-p)^{k-1}$ for $k \geq 1$), then $\mathbf{E}[\xi] = \frac{1}{p}$ and $\text{Var}(\xi) = \frac{1-p}{p^2}$, by direct calculations. Therefore,

$$\mathbf{E}[T_n] = \sum_{k=1}^{n} \frac{n}{n-k+1} = n\log n + O(n),$$

$$\begin{aligned}\text{Var}(T_n) &= n\sum_{k=1}^{n} \frac{k-1}{(n-k+1)^2} = n\sum_{j=1}^{n} \frac{n-j}{j^2} \\ &\leq Cn^2\end{aligned}$$

with $C = \sum_{j=1}^{\infty} \frac{1}{j^2}$. Thus, if $\theta_n \uparrow \infty$, then fix N such that $|\mathbf{E}[T_n] - n\log n| \le \frac{1}{2} n\theta_n$ for $n \ge N$. Then,

$$\begin{aligned}\mathbf{P}\left\{|T_n - n\log n| \ge n\theta_n\right\} &\le \mathbf{P}\left\{|T_n - \mathbf{E}[T_n]| \ge \frac{1}{2}n\theta_n\right\} \\ &\le \frac{\text{Var}(T_n)}{\frac{1}{4}n^2\theta_n^2} \\ &\le \frac{4C}{\theta_n^2}\end{aligned}$$

which goes to zero as $n \to \infty$, proving the theorem. ■

3.3. **Branching processes:** Consider a Galton-Watson branching process with offsprings that are i.i.d ξ. We quickly recall the definition informally. The process starts with one individual in the $0th$ generation who has ξ_1 offsprings and these comprise the first generation. Each of the offsprings (if any) have new offsprings, the number of offsprings being independent and identical copies of ξ. The process continues as long as there are any individuals left[2].

Let Z_n be the number of offsprings in the n^{th} generation. Take $Z_0 = 1$.

Theorem 8. *Let $m = \mathbf{E}[\xi]$ be the mean of the offspring distribution.*

(1) If $m < 1$, then w.p.1, the branching process dies out. That is $\mathbf{P}(Z_n = 0 \text{ for all large } n) = 1$.

(2) If $m > 1$, then the process survives with positive probability, i.e., $\mathbf{P}(Z_n \ge 1 \text{ for all } n) > 0$.

Proof. In the proof, we compute $\mathbf{E}[Z_n]$ and $\text{Var}(Z_n)$ using elementary conditional probability concepts. By conditioning on what happens in the $(n-1)^{\text{st}}$ generation, we write Z_n as a sum of Z_{n-1} independent copies of ξ. From this, one can compute that $\mathbf{E}[Z_n|Z_{n-1}] = mZ_{n-1}$ and if we assume that ξ has variance σ^2 we also get $\text{Var}(Z_n|Z_{n-1}) = Z_{n-1}\sigma^2$. Therefore, $\mathbf{E}[Z_n] = \mathbf{E}[\mathbf{E}[Z_n|Z_{n-1}]] = m\mathbf{E}[Z_{n-1}]$ from which we get $\mathbf{E}[Z_n] = m^n$. Similarly, from the formula $\text{Var}(Z_n) = \mathbf{E}[\text{Var}(Z_n|Z_{n-1})] + \text{Var}(\mathbf{E}[Z_n|Z_{n-1}])$ we can compute that

$$\begin{aligned}\text{Var}(Z_n) &= m^{n-1}\sigma^2 + m^2\text{Var}(Z_{n-1}) \\ &= \left(m^{n-1} + m^n + \ldots + m^{2n-1}\right)\sigma^2 \qquad \text{(by repeating the argument)} \\ &= \sigma^2 m^{n-1}\frac{m^{n+1}-1}{m-1}.\end{aligned}$$

[2] For those who are not satisfied with the informal description, here is a precise definition: Let $V = \bigcup_{k=1}^{\infty} \mathbb{N}_+^k$ be the collection of all finite tuples of positive integers. For $k \ge 2$, say that $(v_1, \ldots, v_k) \in \mathbb{N}_+^k$ is a child of $(v_1, \ldots, v_{k-1}) \in \mathbb{N}_+^{k-1}$. This defines a graph G with vertex set V and edges given by connecting vertices to their children. Let G_1 be the connected component of G containing the vertex (1). Note that G_1 is a tree where each vertex has infinitely many children. Given any $\eta : V \to \mathbb{N}$ (equivalently, $\eta \in \mathbb{N}^V$), define T_η as the subgraph of G_1 consisting of all vertices $(v_1, \ldots, v_k)$ for which $v_j \le \eta((v_1, \ldots, v_{j-1}))$ for $2 \le j \le k$. Also define $Z_{k-1}(\eta) = \#\{(v_1, \ldots, v_k) \in T\}$ for $k \ge 2$ and let $Z_0 = 1$. Lastly, given a probability measure μ on $\mathbb{N}$, consider the product measure $\mu^{\otimes V}$ on $\mathbb{N}^V$. Under this measure, the random variables $\eta(u)$, $u \in V$ are i.i.d. and denote the offspring random variables. The random variable Z_k denotes the number of individuals in the kth generation. The random tree T_η is called the Galton-Watson tree.

(1) By Markov's inequality, $\mathbf{P}(Z_n > 0) \le \mathbf{E}[Z_n] = m^n \to 0$. Since the events $\{Z_n > 0\}$ are decreasing, it follows that $\mathbf{P}(\text{extinction}) = 1$.

(2) If $m = \mathbf{E}[\xi] > 1$, then as before $\mathbf{E}[Z_n] = m^n$ which increases exponentially. But that is not enough to guarantee survival. Assuming that ξ has finite variance σ^2, apply the second moment method to write

$$\mathbf{P}(Z_n > 0) \ge \frac{\mathbf{E}[Z_n]^2}{\text{Var}(Z_n) + \mathbf{E}[Z_n]^2} \ge \frac{1}{1 + \frac{\sigma^2}{m-1}}$$

which is a positive number (independent of n). Again, since $\{Z_n > 0\}$ are decreasing events, we get $\mathbf{P}(\text{non-extinction}) > 0$.

The assumption of finite variance of ξ can be removed as follows. Since $\mathbf{E}[\xi] = m > 1$, we can find A large so that setting $\eta = \min\{\xi, A\}$, we still have $\mathbf{E}[\eta] > 1$. Clearly, η has finite variance. Therefore, the branching process with η offspring distribution survives with positive probability. Then, the original branching process must also survive with positive probability! (A coupling argument is the best way to deduce the last statement: Run the original branching process and kill every child after the first A. If inspite of the violence the population survives, then ...) ■

Remark 9. *The fundamental result of branching processes also asserts the a.s extinction for the critical case $m = 1$. We omit this for now as it does not follow directly from the first and second moment methods.*

3.4. **How many prime divisors does a number typically have?** For a natural number k, let $\nu(k)$ be the number of (distinct) prime divisors of n. What is the typical size of $\nu(n)$ as compared to n? We have to add the word typical, because if p is a prime number then $\nu(p) = 1$ whereas $\nu(2 \times 3 \times \ldots \times p) = p$. Thus there are arbitrarily large numbers with $\nu = 1$ and also numbers for which ν is as large as we wish. To give meaning to "typical", we draw a number at random and look at its ν-value. As there is no natural way to pick one number at random, the usual way of making precise what we mean by a "typical number" is as follows.

Formulation: Fix $n \ge 1$ and let $[n] := \{1, 2, \ldots, n\}$. Let μ_n be the uniform probability measure on $[n]$, i.e., $\mu_n\{k\} = 1/n$ for all $k \in [n]$. Then, the function $\nu : [n] \to \mathbb{R}$ can be considered a random variable, and we can ask about the behaviour of these random variables. Below, we write $\mathbf{E}_n$ to denote expectation w.r.t μ_n.

Theorem 10 (Hardy, Ramanujan). *With the above setting, for any $\delta > 0$, as $n \to \infty$ we have*

$$\mu_n \left\{ k \in [n] : \left| \frac{\nu(k)}{\log\log n} - 1 \right| > \delta \right\} \to 0. \tag{6}$$

Proof. (**Turan**). Fix n and for any prime p define $X_p : [n] \to \mathbb{R}$ by $X_p(k) = \mathbf{1}_{p|k}$. Then, $\nu(k) = \sum\limits_{p \le k} X_p(k)$. We define $\psi(k) := \sum\limits_{p \le \sqrt[4]{k}} X_p(k)$. Then, $\psi(k) \le \nu(k) \le \psi(k) + 4$ since there can be at most four primes larger than $\sqrt[4]{k}$ that divide k. From this, it is clearly enough to show (6) for ψ in place of ν (why?).

We shall need the first two moments of ψ under μ_n. For this we first note that $\mathbf{E}_n[X_p] = \frac{\lfloor \frac{n}{p} \rfloor}{n}$ and $\mathbf{E}_n[X_pX_q] = \frac{\lfloor \frac{n}{pq} \rfloor}{n}$. Observe that $\frac{1}{p} - \frac{1}{n} \le \frac{\lfloor \frac{n}{p} \rfloor}{n} \le \frac{1}{p}$ and $\frac{1}{pq} - \frac{1}{n} \le \frac{\lfloor \frac{n}{pq} \rfloor}{n} \le \frac{1}{pq}$.

By linearity $\mathbf{E}_n[\psi] = \sum\limits_{p \le \sqrt[4]{n}} \mathbf{E}[X_p] = \sum\limits_{p \le \sqrt[4]{n}} \frac{1}{p} + O(n^{-\frac{3}{4}})$. Similarly

$$\begin{aligned}
\mathrm{Var}_n[\psi] &= \sum_{p \le \sqrt[4]{n}} \mathrm{Var}[X_p] + \sum_{p \ne q \le \sqrt[4]{n}} \mathrm{Cov}(X_p, X_q) \\
&= \sum_{p \le \sqrt[4]{n}} \left(\frac{1}{p} - \frac{1}{p^2} + O(n^{-1}) \right) + \sum_{p \ne q \le \sqrt[4]{n}} O(n^{-1}) \\
&= \sum_{p \le \sqrt[4]{n}} \frac{1}{p} - \sum_{p \le \sqrt[4]{n}} \frac{1}{p^2} + O(n^{-\frac{1}{2}}).
\end{aligned}$$

We make use of the following two facts. Here, $a_n \sim b_n$ means that $a_n/b_n \to 1$.

$$\sum_{p \le \sqrt[4]{n}} \frac{1}{p} \sim \log\log n \qquad \sum_{p=1}^{\infty} \frac{1}{p^2} < \infty.$$

The second one is obvious, while the first one is not hard, (see exercise 11 below)). Thus, we get $\mathbf{E}_n[\psi] = \log\log n + O(n^{-\frac{3}{4}})$ and $\mathrm{Var}_n[\psi] = \log\log n + O(1)$. Thus, by Chebyshev's inequality,

$$\mu_n \left\{ k \in [n] : \left| \frac{\psi(k) - \mathbf{E}_n[\psi]}{\log\log n} \right| > \delta \right\} \le \frac{\mathrm{Var}_n(\psi)}{\delta^2 (\log\log n)^2} = O\left(\frac{1}{\log\log n} \right).$$

From the asymptotics $\mathbf{E}_n[\psi] = \log\log n + O(n^{-\frac{3}{4}})$ we also get (for n large enough)

$$\mu_n \left\{ k \in [n] : \left| \frac{\psi(k)}{\log\log n} - 1 \right| > \delta \right\} \le \frac{\mathrm{Var}_n(\psi)}{\delta^2 (\log\log n)^2} = O\left(\frac{1}{\log\log n} \right). \blacksquare$$

Exercise 11. $\sum\limits_{p \le \sqrt[4]{n}} \frac{1}{p} \sim \log\log n$. *[**Note:** This is not trivial although not too hard. Consult some Number theory book.].*

4. Applications of Borel-Cantelli lemmas and Kolmogorov's zero-one law

We already mentioned a few direct consequences of Kolmogorov's zero-one law, such as the constancy of $\limsup_{n \to \infty} \frac{S_n}{n}$. Let us give a couple more.

4.1. **Random power series.** Let X_n be i.i.d. Exp(1). Consider the random power series $\sum_{n=0}^{\infty} X_n(\omega)z^n$. For fixed ω, we know that the radius of convergence is $R(\omega) = (\limsup |X_n(\omega)|^{1/n})^{-1}$. Since this is a tail random variable, by Kolmogorov's zero-one law, it must be constant. In other words, there is a deterministic r_0 such that $R(\omega) = r_0$ $a.s.$

But what is the radius of convergence? It cannot be determined by the zero-one law. We may use Borel-Cantelli lemma to determine it. Observe that $\mathbf{P}(|X_n|^{\frac{1}{n}} > t) = e^{-t^n}$ for any $t > 0$. If $t = 1 + \epsilon$ with $\epsilon > 0$, this decays very fast and is summable. Hence, $|X_n|^{\frac{1}{n}} \leq 1 + \epsilon$ $a.s.$. and hence $R \leq 1 + \epsilon$ $a.s.$ Take intersection over rational ϵ to get $R \leq 1$ $a.s.$. For the other direction, if $t < 1$, then $e^{-t^n} \to 1$ and hence $\sum_n e^{-t^n} = \infty$. Since X_n are independent, so are the events $\{|X_n|^{\frac{1}{n}} > t\}$. By the second Borel-Cantelli lemma, it follows that with probability 1, there are infinitely many n such that $|X_n|^{\frac{1}{n}} \geq 1 - \epsilon$. Again, take intersection over rational ϵ to conclude that $R \geq 1$ $a.s.$ This proves that the radius of convergence is equal to 1 almost surely.

In a homework problem, you are asked to show the same for a large class of distributions and also to find the radius of convergence for more general random series of the form $\sum_{n=0}^{\infty} c_n X_n z^n$.

4.2. **Percolation on a lattice.** This application is really an excuse to introduce a beautiful object of probability. Consider the lattice $\mathbb{Z}^2$, points of which we call vertices. By an edge of this lattice we mean a pair of adjacent vertices $\{(x, y), (p, q)\}$ where $x = p, |y - q| = 1$ or $y = q, |x - p| = 1$. Let E denote the set of all edges. X_e, $e \in E$ be i.i.d Ber(p) random variables indexed by E. Consider the subset of all edges e for which $X_e = 1$. This gives a random subgraph of $\mathbb{Z}^2$ called the *bond percolation graph at level* p. We denote the subgraph by G_ω for ω in the probability space.

Question: What is the probability that in the percolation subgraph, there is an infinite connected component?

Let $A = \{\omega : G_\omega \text{ has an infinite connected component}\}$. If there is an infinite component, changing X_e for finitely many e cannot destroy it. Conversely, if there was no infinite cluster to start with, changing X_e for finitely many e cannot create one. In other words, A is a tail event for the collection X_e, $e \in E$! Hence, by Kolmogorov's 0-1 law[3], $\mathbf{P}_p(A)$ is equal to 0 or 1. Is it 0 or is it 1?

In a pathbreaking work of Harry Kesten, it was proved in 1980s that $\mathbf{P}_p(A) = 0$ if $p \leq \frac{1}{2}$ and $\mathbf{P}_p(A) = 1$ if $p > \frac{1}{2}$. The same problem can be considered on $G = \mathbb{Z}^3$, keeping each edge with probability p and deleting it with probability $1 - p$, independently of all other edges. It is again known (and not too difficult to show) that there is some number $p_c \in (0, 1)$ such that $\mathbf{P}_p(A) = 0$ if $p < p_c$ and $\mathbf{P}_p(A) = 1$ if $p > p_c$. The value of p_c is not known, and more importantly, it is not known whether $\mathbf{P}_{p_c}(A)$ is 0 or 1! This is a typical situation - Kolmogorov's law may tell us that the probability of an event is 0 or 1, but deciding between these two possibilities can be very difficult!

[3] You may be slightly worried that the zero-one law was stated for a sequence but we have an array here. Simply take a bijection $f : \mathbb{N} \to \mathbb{Z}^2$ and define $Y_n = X_{f(n)}$ and observe that the event that we want is in the tail of the sequence $(Y_n)_{n\in\mathbb{N}}$. This shows that we could have stated Kolmogorov's zero one law for a countable collection $\mathcal{F}_i$, $i \in I$, of independent sigma algebras. The tail sigma algebra should then be defined as $\bigcap\limits_{F\subseteq I, |F|<\infty} \sigma(\bigcup\limits_{i\in I\setminus F} \mathcal{F}_i)$

4.3. **Random walk.** Let X_i be i.i.d. $\text{Ber}_{\pm}(1/2)$ and let $S_n = X_1 + \ldots + X_n$ for $n \geq 1$ and $S_0 = 0$ ($S = (S_n)$ is called *simple, symmetric random walk on integers*). Let A be the event that the random walk returns to the origin infinitely often, i.e., $A = \{\omega : S_n(\omega) = 0 \text{ infinitely often}\}$.

Then A is not a tail event. Indeed, suppose $X_k(\omega) = (-1)^k$ for $k \geq 2$. Then, if $X_1(\omega) = -1$, the event A occurs (i.e., $A \ni \omega$) while if $X_1(\omega) = +1$, then A does not occur (i.e., $A \not\ni \omega$). This proves that $A \notin \sigma(X_2, X_3, \ldots)$ and hence, it is not a tail event.

Alternately, you may write $A = \limsup A_n$ where $A_n = \{\omega : S_n(\omega) = 0\}$ and try to use Borel-Cantelli lemmas. It can be shown with some effort that $\mathbf{P}(A_{2n}) \asymp \frac{1}{\sqrt{n}}$ and hence $\sum_n \mathbf{P}(A_n) = \infty$. However, the events A_n are not independent (even pairwise), and hence we cannot apply the second Borel-Cantelli to conclude that $\mathbf{P}(A) = 1$.

Nevertheless, the last statement that $\mathbf{P}(A) = 1$ is true. It is a theorem of Pólya that the random walk returns to the origin in one and two dimensions but not necessarily in three and higher dimensions! If you like a challenge, use the first or second moment methods to show it in the one-dimensional case under consideration (Hint: Let R_n be the number of returns in the first n steps and try to compute/estimate its first two moments).

5. A SHORT PREVIEW OF LAWS OF LARGE NUMBERS AND OTHER THINGS TO COME

If a fair coin is tossed 100 times, we expect that the number of times it turns up heads is close to 50. What do we mean by that, for after all the number of heads could be any number between 0 and 100? What we mean of course, is that the number of heads is unlikely to be far from 50. The weak law of large numbers expresses precisely this.

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite variance. Let $\mu = \mathbf{E}[X_1]$ and $\sigma^2 = \text{Var}(X_1)$. Let $S_n = X_1 + \ldots + X_n$ and $\bar{X}_n = \frac{1}{n} S_n$ be the sample mean of the first n observations.

By linearity of expectations, $\mathbf{E}[\bar{X}_n] = \mu$ and by the independence of X_is, we have $\text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i) = n\sigma^2$. In particular, the standard deviation of S_n is $\sigma\sqrt{n}$, showing that if $\mu = 0$, then the sum of n i.i.d. variables is of order $\sqrt{n}$. This is one of the most important facts in probability and statistics (sometimes called square root law).

Returning to the sample mean, using Chebyshev's inequality, for any $\delta > 0$, we have

$$\mathbf{P}\{|\bar{X}_n - \mu| \geq \delta\} \leq \frac{1}{\delta^2} \mathbf{E}[(\bar{X}_n - \mu)^2] = \frac{\sigma^2}{\delta^2 n}.$$

In particular, $\mathbf{P}\{|\bar{X}_n - \mu| \geq \delta\} \to 0$ as $n \to \infty$. In words, the distribution of the random variable $\bar{X}_n$ puts most of its mass close to the point μ (if n is large). This is one way of making precise the idea that sample mean is close to the population mean. Thus we have proved the following theorem, under the extra assumption that $\text{Var}(X_1)$ is finite.

Theorem 12 (Kolmogorov's weak law of large numbers). *Let $X_1, X_2 \ldots$ be i.i.d random variables. If $\mathbf{E}[|X_1|] < \infty$, then for any $\delta > 0$, as $n \to \infty$, we have $\mathbf{P}\{|\bar{X}_n - \mu| \geq \delta\} \to 0$.*

The full statement (without finite variance assumption) will be proved later. One idea might be to use Markov's inequality for $|\bar{X}_n - \mu|$ (instead of squaring). That gives $\mathbf{P}\{|\bar{X}_n - \mu| \geq \delta\} \leq \frac{1}{\delta}\mathbf{E}[|\bar{X}_n - \mu|]$. If we use triangle inequality we get

$$\mathbf{E}[|\bar{X}_n - \mu|] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[|X_i - \mu|] = \mathbf{E}[|X_1 - \mu|].$$

This would have been useful only if the bound were to go to zero as $n \to \infty$. In summary, whatever be the merits of mean absolute deviation as a measure of dispersion, it is not comparable to standard deviation as a theoretical tool.

We shall also prove the strong law of large numbers which we state now.

Theorem 13 (**Kolmogorov's SLLN**). *Let X_n be i.i.d with $\mathbf{E}[|X_1|] < \infty$. Then, $\mathbf{P}\left\{\lim_{n\to\infty} \bar{X}_n = \mu\right\} = 1$.*

What is the difference between the weak and strong law? Let $A_{n,\delta} = \{|\bar{X}_n - \mu| \geq \delta\}$. Then,

- Weak law is equivalent to the statement that $\mathbf{P}\{A_{n,\delta}\} \to 0$ for every $\delta > 0$.
- Strong law is equivalent to the statement that $\mathbf{P}\{A_{n,\delta} \text{ i.o }\} = 0$ for every $\delta > 0$. This is because, $\lim_{n\to\infty} \bar{X}_n(\omega) = \mu$ if and only if $\omega \in \bigcap_k \bigcup_m \bigcap_{n\geq m} (A_{n,\frac{1}{k}})^c$.

It is easy to figure from this that the strong law implies the weak law but not conversely. We do not elaborate on this now, as we shall later study various modes of convergence of random variables and their relative strengths.

How can we prove the strong law? Fix $\delta > 0$. One way to show that $\mathbf{P}\{A_{n,\delta} \text{ i.o }\} = 0$ is to show that $\sum_n \mathbf{P}\{A_{n,\delta}\} < \infty$ and invoke the first Borel-Cantelli lemma. But even under finite variance assumption, the bound we have on $\mathbf{P}\{A_{n,\delta}\}$ if $\sigma^2/n\delta^2$, which is not summable. We need a better bound.

Proof of SLLN under fourth moment assumption. Assume that $\mathbf{E}[X_1^4] < \infty$. Without loss of generality, assume that $\mu = 0$ (else replace X_i by $X_i - \mu$).

Since $\mathbf{E}[\bar{X}_n] = 0$ too, we get $\mathbf{P}\{A_{n,\delta}\} \leq \delta^{-4}\mathbf{E}[|\bar{X}_n|^4]$ by Markov's inequality. Further,

$$\mathbf{E}[|\bar{X}_n - \mu|^4] = \frac{1}{n}\mathbf{E}\left[\left(\sum_{i=1}^{n} X_i\right)^4\right] = \frac{1}{n^4}\sum_{i,j,k,\ell\leq n}\mathbf{E}[X_iX_jX_kX_\ell]$$
$$= \frac{1}{n^4}\left(n\mathbf{E}[X_1^4] + 3n(n-1)\mathbf{E}[X_1^2]^2\right)$$

since $\mathbf{E}[X_iX_jX_kX_\ell] = 0$ if any one of the indices $i, j, k\ell$ is distnct from the rest (if i is distinct from j, k, ℓ then it factors out as $\mathbf{E}[X_i] = 0$). Thus, we have proved that $\mathbf{P}\{A_{n,\delta}\} \leq C\delta^{-4}n^{-2}$ where $C = 3\mathbf{E}[X_1^2]^2 + \mathbf{E}[X_1^4]$.

Thus, $\sum_n \mathbf{P}\{A_{n,\delta}\} < \infty$ and we see that $\mathbf{P}\{A_{n,\delta} \text{ i.o }\} = 0$ for any $\delta > 0$. ■

6. Modes of convergence

Before going to the proofs of laws of large numbers under optimal conditions, we try to understand the different senses in which random variables can converge to other random variables. Let us recall all the modes of convergence we have introduced so far.

Definition 14. Let X_n, X be real-valued random variables on a common probability space.

- $X_n \overset{a.s.}{\to} X$ (X_n converges to X almost surely) if $\mathbf{P}\left\{\omega : \lim X_n(\omega) = X(\omega)\right\} = 1$.
- $X_n \overset{P}{\to} X$ (X_n converges to X in probability) if $\mathbf{P}\{|X_n - X| > \delta\} \to 0$ as $n \to \infty$ for any $\delta > 0$.
- $X_n \overset{L^p}{\to} X$ (X_n converges to X in L^p) if $\|X_n - X\|_p \to 0$ (i.e., $\mathbf{E}[|X_n - X|^p] \to 0$. This makes sense for any $0 < p \leq \infty$ although $\|\cdots\|_p$ is a norm only for $p \geq 1$. Usually it is understood that $\mathbf{E}[|X_n|^p]$ and $\mathbf{E}[|X|^p]$ are finite, although the definition makes sense without that.
- $X_n \overset{d}{\to} X$ (X_n converges to X in distribution) if the distribution of $\mu_{X_n} \overset{d}{\to} \mu_X$ where μ_X is the distribution of X. This definition (but not the others) makes sense even if the random variables X_n, X are all defined on different probability spaces.

Now, we study the inter-relationships between these modes of convergence.

6.1. **Almost sure and in probability.** Are they really different? Usually looking at Bernoulli random variables elucidates the matter.

Example 15. *Suppose A_n are events in a probability space. Then one can see that*

(1) $\mathbf{1}_{A_n} \overset{P}{\to} 0 \iff \lim\limits_{n\to\infty} \mathbf{P}(A_n) = 0,$

(2) $\mathbf{1}_{A_n} \overset{a.s.}{\to} 0 \iff \mathbf{P}(\limsup A_n) = 0.$

By Fatou's lemma, $\mathbf{P}(\limsup A_n) \geq \limsup \mathbf{P}(A_n)$, and hence we see that a.s convergence of $\mathbf{1}_{A_n}$ to zero implies convergence in probability. The converse is clearly false. For instance, if A_n are independent events with $\mathbf{P}(A_n) = n^{-1}$, then $\mathbf{P}(A_n)$ goes to zero but, by the second Borel-Cantelli lemma $\mathbf{P}(\limsup A_n) = 1$. This example has all the ingredients for the following two implications.

Lemma 16. *Suppose X_n, X are random variables on the same probability space. Then,*

(1) If $X_n \overset{a.s.}{\to} X$, then $X_n \overset{P}{\to} X$.

(2) If $X_n \overset{P}{\to} X$ "fast enough" so that $\sum_n \mathbf{P}(|X_n - X| > \delta) < \infty$ for every $\delta > 0$, then $X_n \overset{a.s.}{\to} X$.

Proof. Note that analogous to the example, in general

(1) $X_n \overset{P}{\to} X \iff \forall \delta > 0,\ \lim\limits_{n\to\infty} \mathbf{P}(|X_n - X| > \delta) = 0,$

(2) $X_n \overset{a.s.}{\to} X \iff \forall \delta > 0,\ \mathbf{P}(\limsup\{|X_n - X| > \delta\}) = 0.$

Thus, applying Fatou's lemma we see that a.s convergence implies convergence in probability. For the second part, observe that by the first Borel Cantelli lemma, if $\sum_n \mathbf{P}(|X_n - X| > \delta) < \infty$, then $\mathbf{P}(|X_n - X| > \delta \text{ i.o}) = 0$ and hence $\limsup |X_n - X| \le \delta \; a.s.$ Apply this to all rational δ and take countable intersection to get $\limsup |X_n - X| = 0$. Thus we get a.s. convergence. ■

The second statement is useful for the following reason. Almost sure convergence $X_n \overset{a.s.}{\to} 0$ is a statement about the joint distribution of the entire sequence $(X_1, X_2, \dots)$ while convergence in probability $X_n \overset{P}{\to} 0$ is a statement about the marginal distributions of X_ns. As such, convergence in probability is often easier to check. If it is fast enough, we also get almost sure convergence for free, without having to worry about the joint distribution of X_ns.

Note that the converse is not true in the second statement. On the probability space $([0,1], \mathcal{B}, \lambda)$, let $X_n = \mathbf{1}_{[0,1/n]}$. Then $X_n \overset{a.s.}{\to} 0$ but $\mathbf{P}(|X_n| \ge \delta)$ is not summable for any $\delta > 0$. Almost sure convergence implies convergence in probability, but no rate of convergence is assured.

Exercise 17. *(1) If $X_n \overset{P}{\to} X$, show that $X_{n_k} \overset{a.s.}{\to} X$ for some subsequence.*

(2) Show that $X_n \overset{P}{\to} X$ if and only if every subsequence of $\{X_n\}$ has a further subsequence that converges a.s.

(3) If $X_n \overset{P}{\to} X$ and $Y_n \overset{P}{\to} Y$ (all r.v.s on the same probability space), show that $aX_n + bY_n \overset{P}{\to} aX + bY$ and $X_nY_n \overset{P}{\to} XY$.

6.2. **In distribution and in probability.** We say that $X_n \overset{d}{\to} X$ if the distributions of X_n converges to the distribution of X. This is a matter of language, but note that X_n and X need not be on the same probability space for this to make sense. In comparing it to convergence in probability, however, we must take them to be defined on a common probability space.

Lemma 18. *Suppose X_n, X are random variables on the same probability space. Then,*

(1) If $X_n \overset{P}{\to} X$, then $X_n \overset{d}{\to} X$.

(2) If $X_n \overset{d}{\to} X$ and X is a constant a.s., then $X_n \overset{P}{\to} X$.

Proof. (1) Suppose $X_n \overset{P}{\to} X$. Since for any $\delta > 0$

$$\mathbf{P}(X_n \le t) \le \mathbf{P}(X \le t + \delta) + \mathbf{P}(X - X_n > \delta)$$
$$\text{and} \quad \mathbf{P}(X \le t - \delta) \le \mathbf{P}(X_n \le t) + \mathbf{P}(X_n - X > \delta),$$

we see that $\limsup \mathbf{P}(X_n \le t) \le \mathbf{P}(X \le t + \delta)$ and $\liminf \mathbf{P}(X_n \le t) \ge \mathbf{P}(X \le t - \delta)$ for any $\delta > 0$. Let t be a continuity point of the distribution function of$X$ and let $\delta \downarrow 0$. We immediately get $\lim_{n\to\infty} \mathbf{P}(X_n \le t) = \mathbf{P}(X \le t)$. Thus, $X_n \overset{d}{\to} X$.

(2) If $X = b$ $a.s.$ (b is a constant), then the cdf of X is $F_X(t) = \mathbf{1}_{t \ge b}$. Hence, $\mathbf{P}(X_n \le b - \delta) \to 0$ and $\mathbf{P}(X_n \le b + \delta) \to 1$ for any $\delta > 0$ as $b \pm \delta$ are continuity points of F_X. Therefore $\mathbf{P}(|X_n - b| > \delta) \le (1 - F_{X_n}(b+\delta)) + F_{X_n}(b - \delta)$ converges to 0 as $n \to \infty$. Thus, $X_n \xrightarrow{P} b$. ■

If $X_n = 1 - U$ and $X = U$, then $X_n \xrightarrow{d} X$ but of course X_n does not converge to X in probability! Thus the condition of X being constant is essential in the second statement. In fact, if X is any non-degenerate random variable, we can find X_n that converge to X in distribution but not in probability. For this, fix $T : [0,1] \to \mathbb{R}$ such that $T(U) \overset{d}{=} X$. Then define $X_n = T(1 - U)$. For all n the random variable X_n has the same distribution as X and hence $X_n \xrightarrow{d} X$. But X_n does not converge in probability to X (unless X is degenerate).

Exercise 19. *(1) Suppose that X_n is independent of Y_n for each n (no assumptions about independence across n). If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, then $(X_n, Y_n) \xrightarrow{d} (U, V)$ where $U \overset{d}{=} X$, $V \overset{d}{=} Y$ and U, V are independent. Further, $aX_n + bY_n \xrightarrow{d} aU + bV$.*

(2) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{d} Y$ (all on the same probability space), then show that $X_nY_n \xrightarrow{d} XY$.

6.3. **In probability and in L^p.** How do convergence in L^p and convergence in probability compare? Suppose $X_n \xrightarrow{L^p} X$ (actually we don't need $p \ge 1$ here, but only $p > 0$ and $\mathbf{E}[|X_n - X|^p] \to 0$). Then, for any $\delta > 0$, by Markov's inequality

$$\mathbf{P}(|X_n - X| > \delta) \le \delta^{-p}\mathbf{E}[|X_n - X|^p] \to 0$$

and thus $X_n \xrightarrow{P} X$. The converse is not true. In fact, even almost sure convergence does not imply convergence in L^p, as the following example shows.

Example 20. *On $([0,1], \mathcal{B}, \lambda)$, define $X_n = 2^n \mathbf{1}_{[0,1/n]}$. Then, $X_n \xrightarrow{a.s.} 0$ but $\mathbf{E}[X_n^p] = n^{-1}2^{np}$ for all n, and hence X_n does not go to zero in L^p (for any $p > 0$).*

As always, the fruitful question is to ask for additional conditions to convergence in probability that would ensure convergence in L^p. Let us stick to $p = 1$. Is there a reason to expect a (weaker) converse? Indeed, suppose $X_n \xrightarrow{P} X$. Then write $\mathbf{E}[|X_n - X|] = \int_0^\infty \mathbf{P}(|X_n - X| > t)dt$. For each t the integrand goes to zero. Will the integral go to zero? Surely, if $|X_n| \le 10$ a.s. for all n, (then the same holds for $|X|$) the integral reduces to the interval $[0, 20]$ and then by DCT (since the integrand is bounded by 1 which is integrable over the interval [0,20]), we get $\mathbf{E}[|X_n - X|] \to 0$.

As example 20 shows, the converse cannot be true in full generality. What goes wrong in that example is that with a small probability X_n can take a very very large value and hence the expected value stays away from zero. This observation makes the next definition more palatable. We put the new concept in a separate section to give it the due respect that it deserves.

7. Uniform integrability

Definition 21. A family $\{X_i\}_{i\in I}$ of random variables is said to be *uniformly integrable* if given any $\epsilon > 0$, there exists A large enough so that $\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>A}] < \epsilon$ for all $i \in I$.

Example 22. *A finite set of integrable random variables is uniformly integrable. More interestingly, an L^p-bounded family with $p > 1$ is u.i. For, if $\mathbf{E}[|X_i|^p] \le M$ for all $i \in I$ for some $M > 0$, then*

$$\mathbf{E}[|X_i|\,\mathbf{1}_{|X_i|>t}] \le \mathbf{E}\left[\left(\frac{|X_i|}{t}\right)^{p-1}|X_i|\,\mathbf{1}_{|X_i|>t}\right] \le \frac{1}{t^{p-1}}M$$

which goes to zero as $t \to \infty$. Thus, given $\epsilon > 0$, one can choose t so that $\sup_{i\in I}\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>t}] < \epsilon$.

This fails for $p = 1$, i.e., an L^1-bounded family of random variables need not be uniformly integrable. To see this, modify Example 20 by defining $X_n = n\mathbf{1}_{[0,\frac{1}{n}]}$.

However, a uniformly integrable family must be bounded in L^1. To see this find $A > 0$ so that $\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>A}] < 1$ for all i. Then, for any $i \in I$, we get $\mathbf{E}[|X_i|] = \mathbf{E}[|X_i|\mathbf{1}_{|X_i|<A}] + \mathbf{E}[|X_i|\mathbf{1}_{|X_i|\ge A}] \le A + 1$. Convince yourself that for any $p > 1$, there exist uniformly integrable families that are not bounded in L^p.

Exercise 23. *If $\{X_i\}_{i\in I}$ and $\{Y_j\}_{j\in J}$ are both u.i, then $\{X_i + Y_j\}_{(i,j)\in I\times J}$ is u.i. What about the family of products, $\{X_iY_j\}_{(i,j)\in I\times J}$?*

Lemma 24. *Suppose X_n, X are integrable random variables on the same probability space. Then, the following are equivalent.*

1. *$X_n \xrightarrow{L^1} X$.*
2. *$X_n \xrightarrow{P} X$ and $\{X_n\}$ is u.i.*

Proof. If $Y_n = X_n - X$, then $X_n \xrightarrow{L^1} X$ iff $Y_n \xrightarrow{L^1} 0$, while $X_n \xrightarrow{P} X$ iff $Y_n \xrightarrow{P} 0$ and by the first part of exercise 23, $\{X_n\}$ is u.i if and only if $\{Y_n\}$ is. Hence we may work with Y_n instead (i.e., we may assume that the limiting r.v. is 0 a.s).

First suppose $Y_n \xrightarrow{L^1} 0$. We already showed that $Y_n \xrightarrow{P} 0$. If $\{Y_n\}$ were not uniformly integrable, then there exists $\delta > 0$ such that for any positive integer k, there is some n_k such that $\mathbf{E}[|Y_{n_k}|\mathbf{1}_{|Y_{n_k}|\ge k}] > \delta$. This in turn implies that $\mathbf{E}[|Y_{n_k}|] > \delta$. But this contradicts $Y_n \xrightarrow{L^1} 0$.

Next suppose $Y_n \xrightarrow{P} 0$ and that $\{Y_n\}$ is u.i. Then, fix $\epsilon > 0$ and find $A > 0$ so that $\mathbf{E}[|Y_k|\mathbf{1}_{|Y_k|>A}] \le \epsilon$ for all k. Then,

$$\begin{aligned}\mathbf{E}[|Y_k|] &\le \mathbf{E}[|Y_k|\mathbf{1}_{|Y_k|\le A}] + \mathbf{E}[|Y_k|\mathbf{1}_{|Y_k|>A}] \\ &\le \int_0^A \mathbf{P}(|Y_k| > t)dt \;+\; \epsilon.\end{aligned}$$

Since $Y_n \xrightarrow{P} 0$ we see that $\mathbf{P}(|Y_k| > t) \to 0$ for all $t < A$. Further, $\mathbf{P}(|Y_k| > t) \leq 1$ for all k and 1 is integrable on $[0, A]$. Hence, by DCT the first term goes to 0 as $k \to \infty$. Thus $\limsup \mathbf{E}[|Y_k|] \leq \epsilon$ for any ϵ and it follows that $Y_k \xrightarrow{L^1} 0$. ■

Corollary 25. *Suppose X_n, X are integrable random variables and $X_n \xrightarrow{a.s.} X$. Then, $X_n \xrightarrow{L^1} X$ if and only if $\{X_n\}$ is uniformly integrable.*

To deduce convergence in mean from a.s convergence, we have so far always invoked DCT. As shown by Lemma 24 and corollary 25, uniform integrability is the sharp condition, so it must be weaker than the assumption in DCT. Indeed, if $\{X_n\}$ are dominated by an integrable Y, then whatever "A" works for Y in the u.i condition will work for the whole family $\{X_n\}$. Thus a dominated family is u.i., while the converse is false.

Remark: Like tightness of measures, uniform integrability is also related to a compactness question. On the space $L^1(\mu)$, apart from the usual topology coming from the norm, there is another one called *weak topology* (where $f_n \to f$ if and only if $\int f_n g d\mu \to \int f g d\mu$ for all $g \in L^\infty(\mu)$). The *Dunford-Pettis theorem* asserts that pre-compact subsets of $L^1(\mu)$ in this weak topology are precisely uniformly integrable subsets of $L^1(\mu)$! A similar question can be asked in L^p for $p > 1$ where weak topology means that $f_n \to f$ if and only if $\int f_n g d\mu \to \int f g d\mu$ for all $g \in L^q(\mu)$ where $q^{-1} + p^{-1} = 1$. Another part of Dunford-Pettis theorem asserts that pre-compact subsets of $L^p(\mu)$ in this weak topology are precisely those that are bounded in the $L^p(\mu)$ norm.

8. Weak law of large numbers

We have already seen the weak law of large numbers under the extra assumption of finite variance. Now we prove the weak law under assuming only the finiteness of the first moment.

Theorem 26 (Kolmogorov's weak law of large numbers). *Let $X_1, X_2 \ldots$ be i.i.d random variables. If $\mathbf{E}[|X_1|] < \infty$, then for any $\delta > 0$, as $n \to \infty$, we have*

$$\mathbf{P}\left(\Big| \frac{X_1 + \ldots + X_n}{n} - \mathbf{E}[X_1] \Big| > \delta \right) \to 0.$$

Let us introduce some terminology. If Y_n, Y are random variables on a probability space and $\mathbf{P}\{|Y_n - Y| \geq \delta\} \to 0$ as $n \to \infty$ for every $\delta > 0$, then we say that Y_n converges to Y *in probability* and write $Y_n \xrightarrow{P} Y$. In this language, the conclusion of the weak law of large numbers is that $\frac{1}{n} S_n \xrightarrow{P} \mathbf{E}[X_1]$ (the limit random variable happens to be constant).

Proof. Without loss of generality assume that $\mathbf{E}[X_1] = 0$. Fix n and write $X_k = Y_k + Z_k$, where $Y_k := X_k \mathbf{1}_{|X_k| \leq A_n}$ and $Z_k := X_k \mathbf{1}_{|X_k| > A_n}$ for some A_n to be chosen later. Then, Y_i are i.i.d, with some mean $\mu_n := \mathbf{E}[Y_1] = -\mathbf{E}[Z_1]$ that depends on A_n and goes to zero as $A_n \to \infty$. Fix $\delta > 0$ and choose n_0 large enough so that $|\mu_n| < \delta$ for $n \geq n_0$.

As $|Y_1| \leq A_n$, we get $\text{Var}(Y_1) \leq \mathbf{E}[Y_1^2] \leq A_n \mathbf{E}[|X_1|]$. By the Chebyshev bound that we used in the first step,

$$\mathbf{P}\left\{ \Big| \frac{S_n^Y}{n} - \mu_n \Big| > \delta \right\} \leq \frac{\text{Var}(Y_1)}{n\delta^2} \leq \frac{1}{n\delta^2}\mathbf{E}[Y_1^2] = \frac{1}{n\delta^2}\mathbf{E}[X_1^2 \mathbf{1}_{|X_1| \geq A_n}]. \tag{7}$$

If $n \geq n_0$ then $|\mu_n| < \delta$ and hence if $|\frac{1}{n}S_n^Z + \mu_n| \geq \delta$, then at least one of $Z_1, \ldots, Z_n$ must be non-zero.

$$\mathbf{P}\left\{ \Big| \frac{S_n^Z}{n} + \mu_n \Big| > \delta \right\} \;\leq\; n\mathbf{P}(Z_1 \neq 0) \;=\; n\mathbf{P}(|X_1| > A_n).$$

Thus, writing $X_k = (Y_k - \mu_n) + (Z_k + \mu_n)$, we see that

$$\begin{aligned}\mathbf{P}\left\{ \Big| \frac{S_n}{n} \Big| > 2\delta \right\} &\leq \mathbf{P}\left\{ \Big| \frac{S_n^Y}{n} - \mu_n \Big| > \delta \right\} + \mathbf{P}\left\{ \Big| \frac{S_n^Z}{n} + \mu_n \Big| > \delta \right\} \\ &\leq \frac{1}{n\delta^2}\mathbf{E}[X_1^2 \mathbf{1}_{|X_1| \geq A_n}] + n\mathbf{P}(|X_1| > A_n) \end{aligned} \tag{8}$$

We use the bound $X_1^2 \mathbf{1}_{|X_1| \leq A_n} \leq A_n |X_1|$ and Markov's inequality to get

$$\mathbf{P}\left\{ \Big| \frac{S_n}{n} \Big| > 2\delta \right\} \leq \frac{A_n \mathbf{E}[|X_1|]}{n\delta^2} + \frac{n}{A_n}\mathbf{E}[|X_1| \; \mathbf{1}_{|X_1| > A_n}].$$

Fix an arbitrary $\epsilon > 0$ and take $A_n = \alpha n$ with $\alpha := \epsilon\delta^2 \mathbf{E}[|X_1|]^{-1}$. The first term clearly becomes less than ϵ. The second term is bounded by $\alpha^{-1}\mathbf{E}[|X_1| \; \mathbf{1}_{|X_1| > \alpha n}]$, which goes to zero as $n \to \infty$ (for any fixed choise of $\alpha > 0$). Thus, we see that $\limsup \mathbf{P}\{|n^{-1}S_n| \geq \delta\} \leq \epsilon$. As this is valid for every $\epsilon > 0$, it follows that $\mathbf{P}\{|n^{-1}S_n| \geq \delta\} \to 0$ as $n \to \infty$. ■

Some remarks about the weak law.

(1) Did we require independence in the proof? If you notice, it was used in only one place, to say that $\text{Var}(S_n^Y) = n\text{Var}(Y_1)$ for which it suffices if Y_i were uncorrelated. In particular, if we assume that X_i *pairwise independent*, identically distributed and have finite mean, then the weak law of large numbers holds as stated.

(2) A simple example that violates law of large numbers is the Cauchy distribution with density $\frac{1}{\pi(1+t^2)}$. Observe that $\mathbf{E}[|X|^p] < \infty$ for all $p < 1$ but not $p = 1$. It is a fact (we shall probably see this later, you may try proving it yourself!) that $\frac{1}{n}S_n$ has exactly the same distribution as X_1. There is no chance of convergence in probability then!

(3) If X_k are i.i.d. random variables (possibly with $\mathbf{E}[|X_1|] = \infty$), let us say that weak law of large numbers is valid if there exist (non-random) numbers a_n such that $\frac{1}{n}S_n - a_n \xrightarrow{P} 0$. When X_i have finite mean, this holds with $a_n = \mathbf{E}[X]$.

It turns out that a necessary and sufficient condition for the existence of such a_n is that $t\mathbf{P}\{|X_1| \geq t\} \to 0$ as $t \to \infty$ (in which case, the weak law holds with $a_n = \mathbf{E}[X\mathbf{1}_{|X| \leq n}]$).

Cauchy distribution is an example that violates this condition. Find a distribution which satisfies the condition but does not have finite expectation.

9. Strong law of large numbers

Theorem 27 (Kolmogorov's SLLN). *Let X_n be i.i.d with $\mathbf{E}[|X_1|] < \infty$. Then, $\frac{S_n}{n} \overset{a.s.}{\to} \mathbf{E}[X_1]$ as $n \to \infty$.*

As we discussed earlier, under the assumption that $\mathbf{E}[X_1^4]$ is finite, $\mathbf{P}\left(|n^{-1}S_n| > \delta\right) = O(n^{-2})$ which is summable, and hence SLLN holds. But proving it under just first moment assumption is nontrivial. We present it now[4].

Proof. **Step 1:** It suffices to prove the theorem for integrable non-negative random variable, because we may write $X = X_+ - X_-$ and it is true that $S_n = S_n^+ - S_n^-$ where $S_n^+ = X_1^+ + \ldots + X_n^+$ and $S_n^- = X_1^- + \ldots + X_n^-$. Henceforth, we assume that $X_n \geq 0$ and $\mu = \mathbf{E}[X_1] < \infty$ (Caution: Don't also assume zero mean in addition to non-negativity!). One consequence of non-negativity is that

$$\frac{S_{N_1}}{N_2} \leq \frac{S_n}{n} \leq \frac{S_{N_2}}{N_1} \text{ if } N_1 \leq n \leq N_2. \tag{9}$$

Step 2: The second step is to prove the following claim. To understand the big picture of the proof, you may jump to the third step where the strong law is deduced using this claim, and then return to the proof of the claim.

Claim 28. *Fix any $\lambda > 1$ and define $n_k := \lfloor \lambda^k \rfloor$. Then, $\frac{S_{n_k}}{n_k} \overset{a.s.}{\to} \mathbf{E}[X_1]$ as $k \to \infty$.*

Proof of the claim: Fix j and for $1 \leq k \leq n_j$ write $X_k = Y_k + Z_k$ where $Y_k = X_k \mathbf{1}_{X_k \leq n_j}$ and $Z_k = X_k \mathbf{1}_{X_k > n_j}$ (why we chose the truncation at n_j is not clear at this point). Then, let J_δ be large enough so that for $j \geq J_\delta$, we have $\mathbf{E}[Z_1] \leq \delta$. Let $S_{n_j}^Y = \sum_{k=1}^{n_j} Y_k$ and $S_{n_j}^Z = \sum_{k=1}^{n_j} Z_k$. Since $S_{n_j} = S_{n_j}^Y + S_{n_j}^Z$ and $\mathbf{E}[X_1] = \mathbf{E}[Y_1] + \mathbf{E}[Z_1]$, we get

$$\begin{aligned}
\mathbf{P}\left\{ \left| \frac{S_{n_j}}{n_j} - \mathbf{E}[X_1] \right| > 2\delta \right\} &\leq \mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| + \left| \frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1] \right| > 2\delta \right\} \\
&\leq \mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} + \mathbf{P}\left\{ \left| \frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1] \right| > \delta \right\} \\
&\leq \mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} + \mathbf{P}\left\{ \frac{S_{n_j}^Z}{n_j} \neq 0 \right\}. \qquad (10)
\end{aligned}$$

[4] The proof given here is due to Etemadi. The presentation is adapted from a blog article of Terence Tao.

We shall show that both terms in (10) are summable over j. The first term can be bounded by Chebyshev's inequality

$$\mathbf{P}\left\{\Big|\frac{S^Y_{n_j}}{n_j}-\mathbf{E}[Y_1]\Big|>\delta\right\}\le\frac{1}{\delta^2 n_j}\mathbf{E}[Y_1^2]=\frac{1}{\delta^2 n_j}\mathbf{E}[X_1^2\mathbf{1}_{X_1\le n_j}]. \tag{11}$$

while the second term is bounded by the union bound

$$\mathbf{P}\left\{\frac{S^Z_{n_j}}{n_j}\ne 0\right\}\le n_j\mathbf{P}(X_1>n_j). \tag{12}$$

The right hand sides of (11) and (12) are both summable. To see this, observe that for any positive x, there is a unique k such that $n_k<x\le n_{k+1}$, and then

$$(a)\quad\sum_{j=1}^{\infty}\frac{1}{n_j}x^2\mathbf{1}_{x\le n_j}\le x^2\sum_{j=k+1}^{\infty}\frac{1}{\lambda^j}\le C_\lambda x,\qquad (b)\quad\sum_{j=1}^{\infty}n_j\mathbf{1}_{x>n_j}\le\sum_{j=1}^{k}\lambda^j\le C_\lambda x.$$

Here, we may take $C_\lambda=\frac{\lambda}{\lambda-1}$, but what matters is that it is some constant depending on λ (but not on x). We have glossed over the difference between $\lfloor\lambda^j\rfloor$ and λ^j but you may check that it does not matter (perhaps by replacing C_λ with a larger value). Setting $x=X_1$ in the above inequalities (a) and (b) and taking expectations, we get

$$\sum_{j=1}^{\infty}\frac{1}{n_j}\mathbf{E}[X_1^2\mathbf{1}_{X_1\le n_j}]\le C_\lambda\mathbf{E}[X_1].\qquad\sum_{j=1}^{\infty}n_j\mathbf{P}(X_1>n_j)\le C_\lambda\mathbf{E}[X_1].$$

As $\mathbf{E}[X_1]<\infty$, the probabilities on the left hand side of (11) and (12) are summable in j, and hence it also follows that $\mathbf{P}\left\{\Big|\frac{S_{n_j}}{n_j}-\mathbf{E}[X_1]\Big|>2\delta\right\}$ is summable. This happens for every $\delta>0$ and hence Lemma 16 implies that $\frac{S_{n_j}}{n_j}\overset{a.s.}{\to}\mathbf{E}[X_1]$ a.s. This proves the claim.

Step 3: Fix $\lambda>1$. Then, for any n, find k such that $\lambda^k<n\le\lambda^{k+1}$, and then, from (9) we get

$$\frac{1}{\lambda}\mathbf{E}[X_1]\le\liminf_{n\to\infty}\frac{S_n}{n}\le\limsup_{n\to\infty}\frac{S_n}{n}\le\lambda\mathbf{E}[X_1],\ \text{almost surely}.$$

Take intersection of the above event over all $\lambda=1+\frac{1}{m}$, $m\ge 1$ to get $\lim\limits_{n\to\infty}\frac{S_n}{n}=\mathbf{E}[X_1]\ a.s.$ ■

10. Applications of law of large numbers

We give three applications, two "practical" and one theoretical.

10.1. Wierstrass' approximation theorem.

Theorem 29. *The set of polynomials is dense in the space of continuous functions (with the sup-norm metric) on an interval of the line.*

Proof (Bernstein). Let $f \in C[0,1]$. For any $n \geq 1$, we define the *Bernstein polynomials* $Q_{f,n}(p) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}p^k(1-p)^{n-k}$. We show that $\|Q_{f,n} - f\| \to 0$ as $n \to \infty$, which is clearly enough. To achieve this, we observe that $Q_{f,n}(p) = \mathbf{E}[f(n^{-1}S_n)]$, where S_n has $\text{Bin}(n,p)$ distribution. Law of large numbers enters, because Binomial may be thought of as a sum of i.i.d Bernoullis.

For $p \in [0,1]$, consider $X_1, X_2, \ldots$ i.i.d Ber(p) random variables. For any $p \in [0,1]$, we have

$$\begin{aligned}
\left|\mathbf{E}_p\left[f\left(\frac{S_n}{n}\right)\right] - f(p)\right| &\leq \mathbf{E}_p\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\right] \\
&= \mathbf{E}_p\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\mathbf{1}_{|\frac{S_n}{n}-p|\leq\delta}\right] + \mathbf{E}_p\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\mathbf{1}_{|\frac{S_n}{n}-p|>\delta}\right] \\
&\leq \omega_f(\delta) + 2\|f\|\mathbf{P}_p\left\{\left|\frac{S_n}{n} - p\right| > \delta\right\} \qquad (13)
\end{aligned}$$

where $\|f\|$ is the sup-norm of f and $\omega_f(\delta) := \sup\{|f(x) - f(y)| : |x-y| < \delta\}$ is the modulus of continuity of f. Observe that $\text{Var}_p(X_1) = p(1-p)$ to write

$$\mathbf{P}_p\left\{\left|\frac{S_n}{n} - p\right| > \delta\right\} \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4\delta^2 n}.$$

Plugging this into (13) and recalling that $Q_{f,n}(p) = \mathbf{E}_p\left[f\left(\frac{S_n}{n}\right)\right]$, we get

$$\sup_{p\in[0,1]}\left|Q_{f,n}(p) - f(p)\right| \leq \omega_f(\delta) + \frac{\|f\|}{2\delta^2 n}$$

Since f is uniformly continuous (which is the same as saying that $\omega_f(\delta) \downarrow 0$ as $\delta \downarrow 0$), given any $\epsilon > 0$, we can take $\delta > 0$ small enough that $\omega_f(\delta) < \epsilon$. With that choice of δ, we can choose n large enough so that the second term becomes smaller than ϵ. With this choice of δ and n, we get $\|Q_{f,n} - f\| < 2\epsilon$. ■

Remark 30. *It is possible to write the proof without invoking WLLN. In fact, we did not use WLLN, but the Chebyshev bound. The main point is that the Bin(n,p) probability measure puts almost all its mass between $np(1-\delta)$ and $np(1+\delta)$ (in fact, in a window of width $\sqrt{n}$ around np). Nevertheless, WLLN makes it transparent why this is so.*

10.2. **Monte Carlo method for evaluating integrals.** Consider a continuous function $f : [a,b] \to \mathbb{R}$ whose integral we would like to compute. Quite often, the form of the function may be sufficiently complicated that we cannot analytically compute it, but is explicit enough that we can numerically evaluate (on a computer) $f(x)$ for any specified x. Here is how one can evaluate the integral by use of random numbers.

Suppose $X_1, X_2, \ldots$ are i.i.d uniform($[a,b]$). Then, $Y_k := f(X_k)$ are also i.i.d with $\mathbf{E}[Y_1] = \int_a^b f(x)dx$. Therefore, by WLLN,

$$\mathbf{P}\left(\Big| \frac{1}{n}\sum_{k=1}^{n} f(X_k) \; - \; \int_a^b f(x)dx \, \Big| > \delta \right) \to 0.$$

Hence if we can sample uniform random numbers from $[a,b]$, then we can evaluate $\frac{1}{n}\sum_{k=1}^{n} f(X_k)$, and present it as an approximate value of the desired integral!

In numerical analysis one uses the same idea, but with deterministic points. The advantage of random samples is that it works irrespective of the niceness of the function. The accuracy is not great, as the standard deviation of $\frac{1}{n}\sum_{k=1}^{n} f(X_k)$ is $Cn^{-1/2}$, so to decrease the error by half, one needs to sample four times as many points.

Exercise 31. *Since $\pi = \int_0^1 \frac{4}{1+x^2}dx$, by sampling uniform random numbers X_k and evaluating $\frac{1}{n}\sum_{k=1}^{n} \frac{4}{1+X_k^2}$ we can estimate the value of π! Carry this out on the computer to see how many samples you need to get the right value to three decimal places.*

10.3. **Accuracy in sample surveys.** Quite often we read about sample surveys or polls, such as "do you support the war in Iraq?". The poll may be conducted across continents, and one is sometimes dismayed to see that the pollsters asked a 1000 people in France and about 1800 people in India (a much much larger population). Should the sample sizes have been proportional to the size of the population?

Behind the survey is the simple hypothesis that each person is a Bernoulli random variable (1='yes', 0='no'), and that there is a probability p_i (or p_f) for an Indian (or a French person) to have the opinion yes. Are different peoples' opinions independent? Definitely not, but let us make that hypothesis. Then, if we sample n people, we estimate p by $\bar{X_n}$ where X_i are i.i.d Ber(p). The accuracy of the estimate is measured by its mean-squared deviation $\sqrt{\text{Var}(\bar{X_n})} = \sqrt{p(1-p)}n^{-\frac{1}{2}}$. Note that this does not depend on the population size, which means that the estimate is about as accurate in India as in France, with the same sample size! This is all correct, provided that the sample size is much smaller than the total population. Even if not satisfied with the assumption of independence, you must concede that the vague feeling of unease about relative sample sizes has no basis in fact...

11. The law of iterated logarithm

If $a_n \uparrow \infty$ is a deterministic sequence, then Kolmogorov's zero-one law implies that $\limsup \frac{S_n}{a_n}$ is constant $a.s.$ This motivates the following natural question.

Question: Let X_i be i.i.d Ber$_{\pm}(1/2)$ random variables. Find a_n so that $\limsup \frac{S_n}{a_n} = 1$ a.s.

The question is about the growth rate of sums of random independent ±1s. We know that $n^{-1}S_n \overset{a.s.}{\to} 0$ by the SLLN, hence, $a_n = n$ is "too much". What about n^α for some $\alpha < 1$? Applying Hoeffding's inequality (to be proved in the next section), we see that $\mathbf{P}(n^{-\alpha}S_n > t) \leq \exp\{-\frac{1}{2}t^2 n^{2\alpha-1}\}$. If $\alpha > \frac{1}{2}$, this is a summable sequence for any $t > 0$, and therefore $\mathbf{P}(n^{-\alpha}S_n > t \text{ i.o.}) = 0$. That is $\limsup n^{-\alpha}S_n \overset{a.s.}{\to} 0$ for $\alpha > \frac{1}{2}$. Alternately (instead of using Hoeffding's inequality), you may do the same using moments. For a positive integer p, we have

$$\begin{aligned}\mathbf{P}\{|S_n| \geq n^\alpha \delta\} &\leq \frac{1}{\delta^{2p}n^{2\alpha p}}\mathbf{E}[(X_1 + \ldots + X_n)^{2p}] \\ &\leq \frac{C_p}{\delta^{2p}n^{2\alpha p - p}}\end{aligned}$$

where we used the fact that $\mathbf{E}[S_n^{2p}] \leq C_p n^p$ (exercise!). If $\alpha = \frac{1}{2} + \epsilon$, then taking p large enough we can make this summable and hence $\frac{1}{n^\alpha}S_n \overset{a.s.}{\to} 0$.

What about $\alpha = \frac{1}{2}$? One can show that $\limsup n^{-\frac{1}{2}}S_n = +\infty$ a.s, which means that $\sqrt{n}$ is too slow compared to S_n. So the right answer is larger than $\sqrt{n}$ but smaller than $n^{\frac{1}{2}+\epsilon}$ for any $\epsilon > 0$. The sharp answer, due to Khinchine is one of the great results of probability theory. Khinchine proved it for Bernoulli random variables AND it was extended to general distributions with finite variance by Hartman and Wintner.

Result 32 (Law of iterated logarithm). Let X_i be i.i.d with zero mean and finite variance σ^2. Then,

$$\limsup_{n\to\infty} \frac{S_n}{\sigma\sqrt{2n\log\log n}} = +1 \text{ a.s.}$$

In fact the set of all limit points of the sequence $\left\{\frac{S_n}{\sigma\sqrt{2n\log\log n}}\right\}$ is almost surely equal to the interval $[-1, 1]$.

We skip the proof of LIL, because it is a bit involved, and there are cleaner ways to deduce it using Brownian motion (in this or a later course).

Exercise 33. *Let X_i be i.i.d random variables taking values ± 1 with equal probability. Show that*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \leq 1, \quad a.s.$$

12. HOEFFDING'S INEQUALITY

If X_n are i.i.d with finite mean, then we know that the probability for S_n/n to be more than δ away from its mean, goes to zero. How fast? Assuming finite variance, we saw that this probability decays at least as fast as n^{-1}. If we assume higher moments, we can get better bounds, but always polynomial decay in n. Here we assume that X_n are bounded a.s, and show that the decay is like a Gaussian.

Lemma 34. ***(Hoeffding's inequality).*** *Let $X_1, \ldots, X_n$ be independent, and assume that $|X_k| \le d_k$ w.p.1. For simplicity assume that $\mathbf{E}[X_k] = 0$. Then, for any $n \ge 1$ and any $t > 0$,*

$$\mathbf{P}\left(|S_n| \ge t\right) \le 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^n d_i^2}\right\}.$$

Remark 35. *The boundedness assumption on X_ks is essential. That $\mathbf{E}[X_k] = 0$ is for convenience. If we remove that assumption, note that $Y_k = X_k - \mathbf{E}[X_k]$ satisfy the assumptions of the theorem, except that we can only say that $|Y_k| \le 2d_k$ (because $|X_k| \le d_k$ implies that $|\mathbf{E}[X_k]| \le d_k$ and hence $|X_k - \mathbf{E}[X_k]| \le 2d_k$). Thus, applying the result to Y_ks, we get*

$$\mathbf{P}\left(|S_n - \mathbf{E}[S_n]| \ge t\right) \le 2\exp\left\{-\frac{t^2}{8\sum_{i=1}^n d_i^2}\right\}.$$

Proof. Without loss of generality, take $\mathbf{E}[X_k] = 0$. Now, if $|X| \le d$ w.p.1, and $\mathbf{E}[X] = 0$, for any $\lambda > 0$ use the convexity of exponential on $[-\lambda d, \lambda d]$ (note that λX lies inside this interval and hence a convex combination of $-\lambda d$ and λd), we get

$$e^{\lambda X} \le \frac{1}{2}\left(\left(1 + \frac{X}{d}\right)e^{\lambda d} + \left(1 - \frac{X}{d}\right)e^{-\lambda d}\right).$$

Therefore, taking expectations we get $\mathbf{E}[\exp\{\lambda X\}] \le \cosh(\lambda d)$. Take $X = X_k$, $d = d_k$ and multiply the resulting inequalities and use independence to get $\mathbf{E}[\exp\{\lambda S_n\}] \le \prod_{k=1}^n \cosh(\lambda d_k)$. Apply the elementary inequality $\cosh(x) \le \exp(x^2/2)$ to get

$$\mathbf{E}[\exp\{\lambda S_n\}] \le \exp\left\{\frac{1}{2}\lambda^2 \sum_{k=1}^n d_k^2\right\}.$$

From Markov's inequality we thus get $\mathbf{P}(S_n > t) \le e^{-\lambda t}\mathbf{E}[e^{\lambda S_n}] \le \exp\left\{-\lambda t + \frac{1}{2}\lambda^2 \sum_{k=1}^n d_k^2\right\}$. Optimizing this over λ gives the choice $\lambda = \frac{t}{\sum_{k=1}^n d_k^2}$ and the inequality

$$\mathbf{P}\left(S_n \ge t\right) \le \exp\left\{-\frac{t^2}{2\sum_{i=1}^n d_i^2}\right\}.$$

Working with $-X_k$ gives a similar inequality for $\mathbf{P}(-S_n > t)$ and adding the two we get the statement in the lemma. ■

The power of Hoeffding's inequality is that it is not an asymptotic statement but valid for every finite n and finite t. Here are two consequences. Let X_i be i.i.d bounded random variables with $\mathbf{P}(|X_1| \le d) = 1$.

(1) (**Large deviation regime**) Take $t = nu$ to get

$$\mathbf{P}\left(\left|\frac{1}{n}S_n - \mathbf{E}[X_1]\right| \ge u\right) = \mathbf{P}\left(|S_n - \mathbf{E}[S_n]| \ge nu\right) \le 2\exp\left\{-\frac{u^2}{8d^2}n\right\}.$$

This shows that for bounded random variables, the probability for the sample sum S_n to deviate by an order n amount from its mean decays exponentially in n. This is called the *large deviation regime* because the order of the deviation is the same as the typical order of the quantity we are measuring.

(2) (**Moderate deviation regime**) Take $t = u\sqrt{n}$ to get

$$\mathbf{P}\left(|S_n - \mathbf{E}[S_n]| \geq u\sqrt{n}\right) \leq 2\exp\left\{-\frac{u^2}{8d^2}\right\}.$$

This shows that S_n is within a window of size $\sqrt{n}$ centered at $\mathbf{E}[S_n]$. In this case the probability is not decaying with n, but the window we are looking at is of a smaller order namely, $\sqrt{n}$, as compared to S_n itself, which is of order n. Therefore this is known as *moderate deviation regime*. The inequality also shows that the tail probability of $(S_n - \mathbf{E}[S_n])/\sqrt{n}$ is bounded by that of a Gaussian with variance d. More generally, if we take $t = un^{\alpha}$ with $\alpha \in [1/2, 1)$, we get $\mathbf{P}\left(|S_n - \mathbf{E}[S_n]| \geq un^{\alpha}\right) \leq 2e^{-\frac{u^2}{8d^2}n^{2\alpha-1}}$.

As Hoeffding's inequality is very general, and holds for all finite n and t, it is not surprising that it is not asymptotically sharp. For example, CLT will show us that $(S_n - \mathbf{E}[S_n])/\sqrt{n} \xrightarrow{d} N(0, \sigma^2)$ where $\sigma^2 = \text{Var}(X_1)$. Since $\sigma^2 < d$, and the $N(0, \sigma^2)$ has tails like $e^{-u^2/2\sigma^2}$, the constant in the exponent given by Hoeffding's is not sharp in the moderate regime. In the large deviation regime, there is well studied theory. A basic result there says that $\mathbf{P}(|S_n - \mathbf{E}[S_n]| > nu) \approx e^{-nI(u)}$, where the function $I(u)$ can be written in terms of the moment generating function of X_1. It turns out that if $|X_i| \leq d$, then $I(u)$ is larger than $u^2/8d^2$ which is what Hoeffding's inequality gave us. Again, Hoeffding's is not sharp in the large deviation regime.

13. Random series with independent terms

In law of large numbers, we considered a sum of n terms scaled by n. A natural question is to ask about convergence of infinite series with terms that are independent random variables. Of course $\sum X_n$ will not converge if X_i are i.i.d (unless $X_i = 0$ a.s!). Consider an example.

Example 36. *Let a_n be i.i.d with finite mean. Important examples are $a_n \sim N(0,1)$ or $a_n = \pm 1$ with equal probability. Then, define $f(z) = \sum_n a_n z^n$. What is the radius of convergence of this series? From the formula for radius of convergence $R = \left(\limsup_{n\to\infty} |a_n|^{\frac{1}{n}}\right)^{-1}$, it is easy to find that the radius of convergence is exactly 1 (a.s.) [**Exercise**]. Thus we get a random analytic function on the unit disk.*

Now we want to consider a general series with independent terms. For this to happen, the individual terms must become smaller and smaller. The following result shows that if that happens in an appropriate sense, then the series converges a.s.

Theorem 37 (Khinchine). *Let X_n be independent random variables with finite second moment. Assume that $\mathbf{E}[X_n] = 0$ for all n and that $\sum_n$ Var$(X_n) < \infty$. Then $\sum X_n$ converges, a.s.*

Proof. A series converges if and only if it satisfies Cauchy criterion. To check the latter, consider N and consider

$$\text{(14)} \qquad \mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } n \geq N\right) = \lim_{m\to\infty} \mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } N \leq n \leq N+m\right).$$

Thus, for fixed N, m we must estimate the probability of the event $\delta < \max_{1\leq k\leq m} |S_{N+k} - S_N|$. For a fixed k we can use Chebyshev's to get $\mathbf{P}(\delta < |S_{N+k} - S_N|) \leq \delta^{-2}\text{Var}(X_N + X_{N+1} + \ldots + X_{N+m})$. However, we don't have a technique for controlling the maximum of $|S_{N+k} - S_N|$ over $k = 1, 2, \ldots, m$. This needs a new idea, provided by Kolmogorov's maximal inequality below.

Invoking 40, we get

$$\mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } N \leq n \leq N+m\right) \leq \delta^{-2} \sum_{k=N}^{N+m} \text{Var}(X_k) \leq \delta^{-2} \sum_{k=N}^{\infty} \text{Var}(X_k).$$

The right hand side goes to zero as $N \to \infty$. Thus, from (14), we conclude that for any $\delta > 0$,

$$\lim_{N\to\infty} \mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } n \geq N\right) = 0.$$

This implies that $\limsup S_n - \liminf S_n \leq \delta$ a.s. Take intersection over $\delta = 1/k$, $k = 1, 2\ldots$ to get that S_n converges a.s. ■

Remark 38. *What to do if the assumptions are not exactly satisfied? First, suppose that $\sum_n$ Var(X_n) is finite but $\mathbf{E}[X_n]$ may not be zero. Then, we can write $\sum X_n = \sum(X_n - \mathbf{E}[X_n]) + \sum \mathbf{E}[X_n]$. The first series on the right satisfies the assumptions of Theorem 37 and hence converges a.s. Therefore, $\sum X_n$ will then converge a.s if and only if the deterministic series $\sum_n \mathbf{E}[X_n]$ converges.*

Next, suppose we drop the finite variance condition too. Now X_n are arbitrary independent random variables. We reduce to the previous case by truncation. Suppose we could find some $A > 0$ such that $\mathbf{P}(|X_n| > A)$ is summable. Then set $Y_n = X_n \mathbf{1}_{|X_n|\leq A}$. By Borel-Cantelli, almost surely, $X_n = Y_n$ for all but finitely many n and hence $\sum X_n$ converges if and only if $\sum Y_n$ converges. Note that Y_n has finite variance. If $\sum_n \mathbf{E}[Y_n]$ converges and $\sum_n$ Var$(Y_n) < \infty$, then it follows from the argument in the previous paragraph and Theorem 37 that $\sum Y_n$ converges a.s. Thus we have proved

Lemma 39 (Kolmogorov's three series theorem - part 1). *Suppose X_n are independent random variables. Suppose for some $A > 0$, the following hold with $Y_n := X_n \mathbf{1}_{|X_n|\leq A}$.*

$$(a)\ \sum_n \mathbf{P}(|X_n| > A) < \infty. \qquad (b)\ \sum_n \mathbf{E}[Y_n] \textit{ converges.} \qquad (c)\ \sum_n \textit{Var}(Y_n) < \infty.$$

Then, $\sum_n X_n$ converges, almost surely.

Kolmogorov showed that if $\sum_n X_n$ converges a.s., then for any *$A > 0$, the three series (a), (b) and (c) must converge. Together with the above stated result, this forms a very satisfactory answer as the question of convergence of a random series (with independent entries) is reduced to that of checking the convergence of three non-random series! We skip the proof of this converse implication.*

14. KOLMOGOROV'S MAXIMAL INEQUALITY

It remains to prove the inequality invoked earlier about the maximum of partial sums of X_is. Note that the maximum of n random variables can be much larger than any individual one. For example, if Y_n are independent Exponential(1), then $\mathbf{P}(Y_k > t) = e^{-t}$, whereas $\mathbf{P}(\max_{k\le n} Y_k > t) = 1-(1-e^{-t})^n$ which is much larger. However, when we consider partial sums $S_1, S_2, \ldots, S_n$, the variables are not independent and a miracle occurs.

Lemma 40 (Kolmogorov's maximal inequality). *Let X_n be independent random variables with finite variance and $\mathbf{E}[X_n] = 0$ for all n. Then, $\mathbf{P}\left(\max_{k\le n} |S_k| > t\right) \le t^{-2}\sum_{k=1}^n Var(X_k)$.*

Proof. The second inequality follows from the first by considering X_ks and their negatives. Hence it suffices to prove the first inequality.

Fix n and let $\tau = \inf\{k \le n : |S_k| > t\}$ where it is understood that $\tau = n$ if $|S_k| \le t$ for all $k \le n$. Then, by Chebyshev's inequality,

$$\mathbf{P}(\max_{k\le n} |S_k| > t) = \mathbf{P}(|S_\tau| > t) \le t^{-2}\mathbf{E}[S_\tau^2]. \tag{15}$$

We control the second moment of S_τ by that of S_n as follows.

$$\begin{aligned}\mathbf{E}[S_n^2] &= \mathbf{E}\left[(S_\tau + (S_n - S_\tau))^2\right] \\ &= \mathbf{E}[S_\tau^2] + \mathbf{E}\left[(S_n - S_\tau)^2\right] + 2\mathbf{E}[S_\tau(S_n - S_\tau)] \\ &\ge \mathbf{E}[S_\tau^2] + 2\mathbf{E}[S_\tau(S_n - S_\tau)]. \qquad (16)\end{aligned}$$

We evaluate the second term by splitting according to the value of τ. Note that $S_n - S_\tau = 0$ when $\tau = n$. Hence,

$$\begin{aligned}\mathbf{E}[S_\tau(S_n - S_\tau)] &= \sum_{k=1}^{n-1} \mathbf{E}[\mathbf{1}_{\tau=k} S_k (S_n - S_k)] \\ &= \sum_{k=1}^{n-1} \mathbf{E}\left[\mathbf{1}_{\tau=k} S_k\right] \mathbf{E}[S_n - S_k] \quad \text{(because of independence)} \\ &= 0 \quad (\text{because } \mathbf{E}[S_n - S_k] = 0).\end{aligned}$$

In the second line we used the fact that $S_k \mathbf{1}_{\tau=k}$ depends on $X_1, \ldots, X_k$ only, while $S_n - S_k$ depends only on $X_{k+1}, \ldots, X_n$. From (16), this implies that $\mathbf{E}[S_n^2] \ge \mathbf{E}[S_\tau^2]$. Plug this into (15) to get $\mathbf{P}(\max_{k\le n} S_k > t) \le t^{-2}\mathbf{E}[S_n^2]$. ■

15. Central limit theorem - statement, heuristics and discussion

If X_i are i.i.d with zero mean and finite variance σ^2, then we know that $\mathbf{E}[S_n^2] = n\sigma^2$, which can roughly be interpreted as saying that $S_n \approx \sqrt{n}$ (That the sum of n random zero-mean quantities grows like $\sqrt{n}$ rather than n is sometimes called the *fundamental law of statistics*). The central limit theorem makes this precise, and shows that on the order of $\sqrt{n}$, the fluctuations (or randomness) of S_n are independent of the original distribution of X_1! We give the precise statement and some heuristics as to why such a result may be expected.

Theorem 41. *Let X_n be i.i.d with mean μ and finite variance σ^2.Then, $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges in distribution to $N(0,1)$.*

Informally, letting χ denote a standard Normal variable, we may write $S_n \approx n\mu + \sigma\sqrt{n}\chi$. This means, the distribution of S_n is hardly dependent on the distribution of X_1 that we started with, except for the two parameters - mean and variance. This is a statement about a remarkable symmetry, where replacing one distribution by another makes no difference to the distribution of the sum.

Heuristics:: Why should one expect such a statement to be true? Without loss of generality, let us take $\mu = 0$ and $\sigma^2 = 1$. First point to note is that the standard deviation of $S_n/\sqrt{n}$ is 1, which gives hope that in the limit we may get a non-degenerate distribution. Indeed, if the variance were going to zero, then we could only expect the limiting distribution to have zero variance and thus be degenerate. Further, since the variance is bounded above, it follows that the distributions of $S_n/\sqrt{n}$ form a tight family. Therefore, there must be subsequences that have distributional limits.

Let us make a leap of faith and assume that the entire sequence $S_n/\sqrt{n}$ converges in distribution to some Y. If so, what can be the distribution of Y? Observe that $(2n)^{-\frac{1}{2}}S_{2n} \stackrel{d}{\to} Y$ and further,

$$\frac{X_1 + X_3 + \ldots + X_{2n-1}}{\sqrt{n}} \stackrel{d}{\to} Y, \qquad \frac{X_2 + X_4 + \ldots + X_{2n}}{\sqrt{n}} \stackrel{d}{\to} Y.$$

But $(X_1, X_3, \ldots)$ is independent of $(X_2, X_4, \ldots)$. Therefore, by an earlier exercise, we also get

$$\left(\frac{X_1 + X_3 + \ldots + X_{2n-1}}{\sqrt{n}}\ ,\ \frac{X_2 + X_4 + \ldots + X_{2n}}{\sqrt{n}}\right) \stackrel{d}{\to} (Y_1, Y_2)$$

where Y_1, Y_2 are i.i.d copies of Y. But then, by yet another exercise, we get

$$\frac{S_{2n}}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\left(\frac{X_1 + X_3 + \ldots + X_{2n-1}}{\sqrt{n}} + \frac{X_2 + X_4 + \ldots + X_{2n}}{\sqrt{n}}\right) \stackrel{d}{\to} \frac{Y_1 + Y_2}{\sqrt{2}}$$

Thus we must have $Y_1 + Y_2 \stackrel{d}{=} \sqrt{2}Y$. If $Y_1 \sim N(0, \sigma^2)$, then certainly it is true that $Y_1 + Y_2 \stackrel{d}{=} \sqrt{2}Y$. We claim that $N(0, \sigma^2)$ are the only distributions that have this property. If so, then it gives a strong heuristic that the central limit theorem is true.

To show that $N(0,\sigma^2)$ is the only distribution that satisfies $Y_1+Y_2 \overset{d}{=} \sqrt{2}Y$ (where Y_1, Y_2, Y are i.i.d. $N(0,\sigma^2)$) is not trivial. The cleanest way is to use characteristic functions. If $\psi(t)$ denotes the characteristic function of Y, then

$$\psi(t) = \mathbf{E}\left[e^{itY}\right] = \mathbf{E}\left[e^{itY/\sqrt{2}}\right]^2 = \psi\left(\frac{t}{\sqrt{2}}\right)^2.$$

From this, by standard methods, one can deduce that $\psi(t) = e^{-at^2}$ for some $a > 0$ (**exercise**, but note that characteristic functions are always continuous). By uniqueness of characteristic functions, $Y \sim N(0, 2a)$. Since we expect $\mathbf{E}[Y^2] = 1$, we must get $N(0,1)$.

Apart from these heuristics, it is possible to prove the CLT for certain special distributions. One is of course the Demoivre-Laplace limit theorem (CLT for Bernoullis), which is well known and we omit it here. We just recall that sums of independent Bernoullis have binomial distribution, with explicit formula for the probability mass function and whose asymptotics can be calculated using Stirling's formula.

Instead, let us consider the less familiar case of exponential distribution. If X_i are i.i.d Exp(1) so that $\mathbf{E}[X_1] = 1$ and $\text{Var}(X_1) = 1$. Then $S_n \sim \text{Gamma}(n,1)$ and hence $\frac{S_n - n}{\sqrt{n}}$ has density

$$\begin{aligned} f_n(x) &= \frac{1}{\Gamma(n)} e^{-n-x\sqrt{n}} (n + x\sqrt{n})^{n-1}\sqrt{n} \\ &= \frac{e^{-n}n^{n-\frac{1}{2}}}{\Gamma(n)} e^{-x\sqrt{n}} \left(1 + \frac{x}{\sqrt{n}}\right)^{n-1} \\ &\to \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \end{aligned}$$

by elementary calculations (use Stirling's approximation for $\Gamma(n)$ and for terms involving x write the exponent as $-x\sqrt{n} + \log(1 + x/\sqrt{n})$ and use the Taylor expansion of logarithm). By an earlier exercise convergence of densities implies convergence in distribution and thus we get CLT for sums of exponential random variables.

Exercise 42. *Prove the CLT for $X_1 \sim$ Ber(p). Note that this also implies CLT for $X_1 \sim$ Bin(k,p).*

Lastly, we show how the CLT can be derived under strong assumptions by the method of moments. As justifying all the steps here would take time, let us simply present it as a heuristic for CLT for Bernoulli random variables. Let X_i be i.i.d. Ber$_{\pm}(1/2)$.

As usual $S_n = X_1 + \ldots + X_n$. Let us compute the moments of $S_n/\sqrt{n}$. By the symmetry of the distribution, all odd moments are zero. Consider an even moment

$$\mathbf{E}[(S_n/\sqrt{n})^{2p}] = \frac{1}{n^p} \sum_{1 \le i_1, \ldots, i_{2p} \le n} \mathbf{E}[X_{i_1} X_{i_2} \ldots X_{i_{2p}}].$$

If any X_k occurs an odd number of times in the expectation, then the term is zero.

16. Strategies of proof of central limit theorem

In the next few sections, we shall prove CLT as stated in Theorem 41 as well as a more general CLT for triangular arrays to be stated in Theorem 49. We shall in fact give two proofs, one via the replacement strategy of Lindeberg and another via characteristic functions. Both proofs teach useful techniques in probability. To separate the key ideas from technical details that are less essential, we shall first prove a weaker version of Theorem 41 (assuming that X_1 has finite third moment) by both approaches. Then we prove the more general Theorem 49 (which implies Theorem 41 anyway) by adding minor technical details to both approaches.

What are these two strategies? The starting point is the following fact that we have seen before.

Lemma 43. *$Y_n \overset{d}{\to} Y$ if and only if $\mathbf{E}[f(Y_n)] \to \mathbf{E}[f(Y)]$ for all $f \in C_b(\mathbb{R})$. Here $C_b(\mathbb{R})$ is the space of bounded continuous functions on $\mathbb{R}$.*

The implication that we shall use is one way, and let us recall how that is proved.

Proof of one implication. Suppose $\mathbf{E}[f(Y_n)] \to \mathbf{E}[f(Y)]$ for all $f \in C_b(\mathbb{R})$. Fix t, a continuity point of F_Y, and for each $k \ge 1$ define a function $f_k \in C_b(\mathbb{R})$ such that $0 \le f_k \le 1$, $f_k(x) = 1$ for $x \le t$ and $f_k(x) = 0$ for $x \ge t + \frac{1}{k}$. For example, we may take f_k to be linear in $[t, t+\frac{1}{k}]$.

As $f_k \in C_b(\mathbb{R})$, we get $\mathbf{E}[f_k(Y_n)] \to \mathbf{E}[f_k(Y)]$ as $n \to \infty$. But $F_Y(t) \le \mathbf{E}[f_k(Y)] \le F_Y(t+\frac{1}{k})$ and $F_{Y_n}(t) \le \mathbf{E}[f_k(Y_n)] \le F_{Y_n}(t+\frac{1}{k})$. Hence, $\limsup_{n\to\infty} F_{Y_n}(t) \le F_Y(t+\frac{1}{k})$. This being true for every k, we let $k \to \infty$ and get $\limsup_{n\to\infty} F_{Y_n}(t) \le F_Y(t)$. Similarly, use the function $g_k(x) := f_k(x+\frac{1}{k})$ to get

$$\liminf_{n\to\infty} F_{Y_n}(t) \ge \lim_{n\to\infty} \mathbf{E}[g_k(Y_n)] = \mathbf{E}[g_k(Y)] \ge F_Y(t - \frac{1}{k}).$$

Again, letting $k \to \infty$ and using continuity of F_Y at t we get $\liminf_{n\to\infty} F_{Y_n}(t) \ge F_Y(t)$. Thus, $Y_n \overset{d}{\to} Y$. ■

Continuous functions are more easy to work with than indicators of intervals, hence the usefulness of the above lemma. However, it is even more convenient that we can restrict to smaller subclasses of the space of continuous functions. We state two results to that effect.

Lemma 44. *Suppose $\mathbf{E}[f(Y_n)] \to \mathbf{E}[f(Y)]$ for all $f \in C_b^{(3)}(\mathbb{R})$, the space of all functions whose first three derivatives exist, are continuous and bounded. Then, $Y_n \overset{d}{\to} Y$.*

Proof. Repeat the proof given for Lemma 43 but take f_k to be a smooth function such that $0 \le f_k \le 1$, $f_k(x) = 1$ for $x \le t$ and $f_k(x) = 0$ for $x \ge t + \frac{1}{k}$. ■

Here is the further reduction, which unlike the first, is not so obvious! It is proved in the appendix, and goes by the name *Lévy's continuity theorem*.

Lemma 45 (Lévy's continuity theorem)**.** *Suppose $\mathbf{E}[e^{i\lambda Y_n}] \to \mathbf{E}[e^{i\lambda Y}]$ for all $\lambda \in \mathbb{R}$. Then, $Y_n \overset{d}{\to} Y$.*

In this lemma, we only check convergence of expectations for the very special class of functions $e_\lambda(y) := e^{i\lambda y}$, for $\lambda \in \mathbb{R}$. Note that by the expectation of a complex valued random variable $U + iV$ with U, V real-valued, we simply mean $\mathbf{E}[U] + i\mathbf{E}[V]$. The function $\varphi_Y : \mathbb{R} \to \mathbb{C}$ defined by $\varphi_Y(\lambda) = \mathbf{E}[e^{i\lambda Y}]$ is called the *characteristic function* of Y. It is a very useful tool in probability and analysis, and a brief introduction including the proof of the above lemma is give in the appendix 21.

17. Central limit theorem - two proofs assuming third moments

We give two proofs of the following slightly weaker version of CLT.

Theorem 46. *Let X_n be i.i.d with finite third moment, and having zero mean and unit variance. Then, $\frac{S_n}{\sqrt{n}}$ converges in distribution to $N(0,1)$.*

Once the ideas are clear, we prove a much more general version later, which will also subsume Theorem 41.

17.1. Proof via characteristic functions.

We shall need the following facts.

Exercise 47. *Let z_n be complex numbers such that $nz_n \to z$. Then, $(1+z_n)^n \to e^z$.*

We need a second fact that is proved in the appendix 21. It is quite easy to prove it incorrectly, but less so to prove it correctly!

Exercise 48. *Let $X \sim N(0,1)$. Then, $\mathbf{E}[e^{itX}] = e^{-\frac{1}{2}t^2}$.*

Proof of Theorem 46. By Lévy's continuity theorem (Lemma 45), it suffices to show that the characteristic functions of $n^{-\frac{1}{2}}S_n$ converge to the characteristic function of $N(0,1)$. The characteristic function of $S_n/\sqrt{n}$ is $\psi_n(t) := \mathbf{E}\left[e^{itS_n/\sqrt{n}}\right]$. Writing $S_n = X_1 + \ldots + X_n$ and using independence,

$$\begin{aligned}\psi_n(t) &= \mathbf{E}\left[\prod_{k=1}^n e^{itX_k/\sqrt{n}}\right] \\ &= \prod_{k=1}^n \mathbf{E}\left[e^{itX_k/\sqrt{n}}\right] \\ &= \psi\left(\frac{t}{\sqrt{n}}\right)^n\end{aligned}$$

where ψ denotes the characteristic function of X_1.

Use Taylor expansion to third order for the function $x \to e^{itx}$ to write,

$$e^{itx} = 1 + itx - \frac{1}{2}t^2x^2 - \frac{i}{6}t^3 e^{itx^*}x^3 \qquad \text{for some } x^* \in [0,x] \text{ or } [x,0].$$

Apply this with X_1 in place of x and $tn^{-1/2}$ in place of t. Then take expectations and recall that $\mathbf{E}[X_1]=0$ and $\mathbf{E}[X_1^2]=1$ to get

$$\psi\left(\frac{t}{\sqrt{n}}\right)=1-\frac{t^2}{2n}+R_n(t), \qquad \text{where } R_n(t)=-\frac{i}{6n^{\frac{3}{2}}}t^3\mathbf{E}\left[e^{itX_1^*}X_1^3\right].$$

Clearly, $|R_n(t)|\le C_t n^{-3/2}$ for a constant C_t (that depends on t but not n). Hence $nR_n(t)\to 0$ and by Exercise 47 we conclude that for each fixed $t\in\mathbb{R}$,

$$\psi_n(t)=\left(1-\frac{t^2}{2n}+R_n(t)\right)^n\to e^{-\frac{t^2}{2}}$$

which is the characteristic function of $N(0,1)$. ■

17.2. **Proof using Lindeberg's replacement idea.** Here the idea is more probabilistic. First we observe that the central limit theorem is trivial for $(Y_1+\ldots+Y_n)/\sqrt{n}$, if Y_i are independent $N(0,1)$ random variables. The key idea of Lindeberg is to go from $X_1+\ldots+X_n$ to $Y_1+\ldots+Y_n$ in steps, replacing each X_i by Y_i, one at a time, and arguing that the distribution does not change much!

Proof. We assume, without loss of generality, that X_i and Y_i are defined on the same probability space, are all independent, X_i have the given distribution (with zero mean and unit variance) and Y_i have $N(0,1)$ distribution.

Fix $f\in C_b^{(3)}(\mathbb{R})$ and let $\sqrt{n}U_k=\sum_{j=1}^{k-1}X_j+\sum_{j=k+1}^{n}Y_j$ and $\sqrt{n}V_k=\sum_{j=1}^{k}X_j+\sum_{j=k+1}^{n}Y_j$ for $0\le k\le n$ and empty sums are regarded as zero. Then, $V_0=S_n^Y/\sqrt{n}$ and $V_n=S_n^X/\sqrt{n}$. Also, $S_n^Y/\sqrt{n}$ has the same distribution as Y_1. Thus,

$$\begin{aligned}\mathbf{E}\left[f\left(\frac{1}{\sqrt{n}}S_n^X\right)\right]-\mathbf{E}[f(Y_1)]&=\sum_{k=1}^{n}\mathbf{E}\left[f\left(V_k\right)-f\left(V_{k-1}\right)\right]\\&=\sum_{k=1}^{n}\mathbf{E}\left[f\left(V_k\right)-f\left(U_k\right)\right]-\sum_{k=1}^{n}\mathbf{E}\left[f\left(V_{k-1}\right)-f\left(U_k\right)\right].\end{aligned}$$

By Taylor expansion, we see that

$$\begin{aligned}f(V_k)-f(U_k)&=f'(U_k)\frac{X_k}{\sqrt{n}}+f''(U_k)\frac{X_k^2}{2n}+f'''(U_k^*)\frac{X_k^3}{6n^{\frac{3}{2}}},\\f(V_{k-1})-f(U_k)&=f'(U_k)\frac{Y_k}{\sqrt{n}}+f''(U_k)\frac{Y_k^2}{2n}+f'''(U_k^{**})\frac{Y_k^3}{6n^{\frac{3}{2}}}.\end{aligned}$$

Take expectations and subtract. A key observation is that U_k is independent of X_k, Y_k. Therefore, $\mathbf{E}[f'(U_k)X_k^p]=\mathbf{E}[f'(U_k)]\mathbf{E}[X_k^p]$ etc. Consequently, using equality of the first two moments of X_k, Y_k, we get

$$\mathbf{E}[f(V_k)-f(V_{k-1})]=\frac{1}{6n^{\frac{3}{2}}}\left\{\mathbf{E}[f'''(U_k^*)X_k^3]+\mathbf{E}[f'''(U_k^{**})Y_k^3]\right\}.$$

Now, U_k^* and U_k^{**} are not independent of X_k, Y_k, hence we cannot factor the expectations. We put absolute values and use the bound on derivatives of f to get

$$\Big|\mathbf{E}[f(V_k)] - \mathbf{E}[f(V_{k-1})]\Big| \leq \frac{1}{n^{\frac{3}{2}}} C_f \left\{ \mathbf{E}[|X_1|^3] + \mathbf{E}[|Y_1|^3] \right\}.$$

Add up over k from 1 to n to get

$$\left|\mathbf{E}\left[f\left(\frac{1}{\sqrt{n}} S_n^X\right)\right] - \mathbf{E}[f(Y_1)]\right| \leq \frac{1}{n^{\frac{1}{2}}} C_f \left\{ \mathbf{E}[|X_1|^3] + \mathbf{E}[|Y_1|^3] \right\}$$

which goes to zero as $n \to \infty$. Thus, $\mathbf{E}[f(S_n/\sqrt{n})] \to \mathbf{E}[f(Y_1)]$ for any $f \in C_b^{(3)}(\mathbb{R})$ and consequently, by Lemma 44 we see that $\frac{1}{\sqrt{n}} S_n \xrightarrow{d} N(0,1)$. ■

18. Central limit theorem for triangular arrays

The CLT does not really require the third moment assumption, and we can modify the above proof to eliminate that requirement. Instead, we shall prove an even more general theorem, where we don't have one infinite sequence, but the random variables that we add to get S_n depend on n themselves. Further, observe that we assume independence but not identical distributions in each row of the triangular array.

Theorem 49 (Lindeberg-Feller CLT). *Suppose $X_{n,k}$, $k \leq n$, $n \geq 1$, are random variables. We assume that*

(1) For each n, the random variables $X_{n,1}, \ldots, X_{n,n}$ are defined on the same probability space, are independent, and have finite variances.

(2) $\mathbf{E}[X_{n,k}] = 0$ and $\sum_{k=1}^n \mathbf{E}[X_{n,k}^2] \to \sigma^2$, as $n \to \infty$.

(3) For any $\delta > 0$, we have $\sum_{k=1}^n \mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}| > \delta}] \to 0$ as $n \to \infty$.

Then, $X_{n,1} + \ldots + X_{n,n} \xrightarrow{d} N(0, \sigma^2)$ as $n \to \infty$.

First we show how this theorem implies the standard central limit theorem under second moment assumptions.

Proof of Theorem 41 from Theorem 49. Let $X_{n,k} = n^{-\frac{1}{2}} X_k$ for $k = 1, 2, \ldots, n$. Then, $\mathbf{E}[X_{n,k}] = 0$ while $\sum_{k=1}^n \mathbf{E}[X_{n,k}^2] = \frac{1}{n} \sum_{k=1}^n \mathbf{E}[X_1^2] = \sigma^2$, for each n. Further, $\sum_{k=1}^n \mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}| > \delta}] = \mathbf{E}[X_1^2 \mathbf{1}_{|X_1| > \delta\sqrt{n}}]$ which goes to zero as $n \to \infty$ by DCT, since $\mathbf{E}[X_1^2] < \infty$. Hence the conditions of Lindeberg Feller theorem are satisfied and we conclude that $\frac{S_n}{\sqrt{n}}$ converges in distribution to $N(0,1)$. ■

But apart from the standard CLT, many other situations of interest are covered by the Lindeberg-Feller CLT. We consider some examples.

Example 50. *Let $X_k \sim Ber(p_k)$ be independent random variables with $0 < p_k < 1$. Is S_n asymptotically normal? By this we mean, does $(S_n - \mathbf{E}[S_n])/\sqrt{Var(S_n)}$ converge in distribution to $N(0,1)$? Obviously the standard CLT does not apply.*

To fit it in the framework of Theorem 49, define $X_{n,k} = \frac{X_k - p_k}{\tau_n}$ where $\tau_n^2 = \sum_{k=1}^n p_k(1-p_k)$ is the variance of S_n. The first assumption in Theorem 49 is obviously satisfied. Further, $X_{n,k}$ has mean zero and variance $p_k(1-p_k)/\tau_n^2$ which sum up to 1 (when summed over $1 \le k \le n$). As for the crucial third assumption, observe that $\mathbf{1}_{|X_{n,k}|>\delta} = \mathbf{1}_{|X_k - p_k|>\delta\tau_n}$. If $\tau_n \uparrow \infty$ as $n \to \infty$, then the indicator becomes zero (since $|X_k - p_k| \le 1$). This shows that whenever $\tau_n \to \infty$, asymptotic normality holds for S_n.

If τ_n does not go to infinity, there is no way CLT can hold. We leave it for the reader to think about, just pointing out that in this case, X_1 has a huge influence on $(S_n - \mathbf{E}[S_n])/\tau_n$. Changing X_1 from 0 to 1 or vice versa will induce a big change in the value of $(S_n - \mathbf{E}[S_n])/\tau_n$ from which one can argue that the latter cannot be asymptotically normal.

The above analysis works for any uniformly bounded sequence of random variables. Here is a generalization to more general, independent but not identically distributed random variables.

Exercise 51. *Suppose X_k are independent random variables and $\mathbf{E}[|X_k|^{2+\delta}] \le M$ for some $\delta > 0$ and $M < \infty$. If $Var(S_n) \to \infty$, show that S_n is asymptotically normal.*

Here is another situation covered by the Lindeberg-Feller CLT but not by the standard CLT.

Example 52. *If X_n are i.i.d (mean zero and unit variance) random variable, what can we say about the asymptotics of $T_n := X_1 + 2X_2 + \ldots + nX_n$? Clearly $\mathbf{E}[T_n] = 0$ and $\mathbf{E}[T_n^2] = \sum_{k=1}^n k^2 \sim \frac{n^3}{3}$. Thus, if we expect any convergence to Gaussian, then it must be that $n^{-\frac{3}{2}}T_n \xrightarrow{d} N(0,1/3)$.*

To prove that this is indeed so, write $n^{-\frac{3}{2}}T_n = \sum_{k=1}^n X_{n,k}$, where $X_{n,k} = n^{-\frac{3}{2}}kX_k$. Let us check the crucial third condition of Theorem 49.

$$\begin{aligned}\mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] &= n^{-3}k^2\mathbf{E}[X_k^2\mathbf{1}_{|X_k|>\delta k^{-1}n^{3/2}}] \\ &\le n^{-1}\mathbf{E}[X^2\mathbf{1}_{|X|>\delta\sqrt{n}}] \qquad (\textit{since } k \le n)\end{aligned}$$

which when added over k gives $\mathbf{E}[X^2\mathbf{1}_{|X|>\delta\sqrt{n}}]$. Since $\mathbf{E}[X^2] < \infty$, this goes to zero as $n \to \infty$, for any $\delta > 0$.

Exercise 53. *Let $0 < a_1 < a_2 < \ldots$ be fixed numbers and let X_k be i.i.d. random variables with zero mean and unit variance. Find simple sufficient conditions on a_k to ensure asymptotic normality of $T_n := \sum_{k=1}^n a_k X_k$.*

19. Two proofs of the Lindeberg-Feller CLT

Now we prove the Lindeberg-Feller CLT by both approaches. It makes sense to compare with the earlier proofs and see where some modifications are required.

19.1. **Proof via characteristic functions.** As in the earlier proof, we need a fact comparing a product to an exponential.

Exercise 54. *If $z_k, w_k \in \mathbb{C}$ and $|z_k|, |w_k| \le \theta$ for all k, then* $\left| \prod\limits_{k=1}^{n} z_k - \prod\limits_{k=1}^{n} w_k \right| \le \theta^{n-1} \sum\limits_{k=1}^{n} |z_k - w_k|$.

Proof of Theorem 49. The characteristic function of $S_n = X_{n,1} + \ldots + X_{n,n}$ is given by $\psi_n(t) = \prod\limits_{k=1}^{n} \mathbf{E}\left[e^{itX_{n,k}}\right]$. Again, we shall use the Taylor expansion of e^{itx}, but we shall need both the second and first order expansions.

$$e^{itx} = \begin{cases} 1 + itx - \frac{1}{2}t^2x^2 - \frac{i}{6}t^3 e^{itx^*} x^3 & \text{for some } x^* \in [0,x] \text{ or } [x,0]. \\ 1 + itx - \frac{1}{2}t^2 e^{itx^+} x^2 & \text{for some } x^+ \in [0,x] \text{ or } [x,0]. \end{cases}$$

Fix $\delta > 0$ and use the first equation for $|x| \le \delta$ and the second one for $|x| > \delta$ to write

$$e^{itx} = 1 + itx - \frac{1}{2}t^2x^2 + \frac{\mathbf{1}_{|x|>\delta}}{2} t^2 x^2 (1 - e^{itx^+}) - \frac{i\mathbf{1}_{|x|\le\delta}}{6} t^3 x^3 e^{itx^*}.$$

Apply this with $x = X_{n,k}$, take expectations and write $\sigma_{n,k}^2 := \mathbf{E}[X_{n,k}^2]$ to get

$$\mathbf{E}[e^{itX_{n,k}}] = 1 - \frac{1}{2}\sigma_{n,k}^2 t^2 + R_{n,k}(t)$$

where, $R_{n,k}(t) := \frac{t^2}{2}\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|>\delta} X_{n,k}^2 \left(1 - e^{itX_{n,k}^+}\right)\right] - \frac{it^3}{6}\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|\le\delta} X_{n,k}^3 e^{itX_{n,k}^*}\right]$. We can bound $R_{n,k}(t)$ from above by using $|X_{n,k}|^3 \mathbf{1}_{|X_{n,k}|\le\delta} \le \delta X_{n,k}^2$ and $|1 - e^{itx}| \le 2$, to get

$$|R_{n,k}(t)| \le t^2 \mathbf{E}\left[\mathbf{1}_{|X_{n,k}|>\delta} X_{n,k}^2\right] + \frac{|t|^3\delta}{6}\mathbf{E}\left[X_{n,k}^2\right]. \tag{17}$$

We want to apply Exercise 54 to $z_k = \mathbf{E}\left[e^{itX_{n,k}}\right]$ and $w_k = 1 - \frac{1}{2}\sigma_{n,k}^2 t^2$. Clearly $|z_k| \le 1$ by properties of c.f. If we prove that $\max\limits_{k\le n} \sigma_{n,k}^2 \to 0$, then it will follow that $|w_k| \le 1$ and hence with $\theta = 1$ in Exercise 54, we get

$$\begin{aligned} \limsup_{n\to\infty} \left| \prod_{k=1}^{n} \mathbf{E}\left[e^{itX_{n,k}}\right] - \prod_{k=1}^{n}\left(1 - \frac{1}{2}\sigma_{n,k}^2 t^2\right) \right| &\le \limsup_{n\to\infty} \sum_{k=1}^{n} |R_{n,k}(t)| \\ &\le \frac{1}{6}|t|^3\sigma^2\delta \quad \text{(by 17)} \end{aligned}$$

To see that $\max\limits_{k\le n} \sigma_{n,k}^2 \to 0$, fix any $\delta > 0$ note that $\sigma_{n,k}^2 \le \delta^2 + \mathbf{E}\left[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}\right]$ from which we get

$$\max_{k\le n} \sigma_{n,k}^2 \le \delta^2 + \sum_{k=1}^{n} \mathbf{E}\left[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}\right] \to \delta^2.$$

As δ is arbitrary, it follows that $\max\limits_{k\le n}\sigma_{n,k}^2 \to 0$ as $n\to\infty$. As $\delta>0$ is arbitrary, we get

$$\lim_{n\to\infty}\prod_{k=1}^{n}\mathbf{E}\left[e^{itX_{n,k}}\right]=\lim_{n\to\infty}\prod_{k=1}^{n}\left(1-\frac{1}{2}\sigma_{n,k}^2t^2\right). \tag{18}$$

For n large enough (and fixed t), $\max\limits_{k\le n} t^2\sigma_{n,k}^2 \le \frac{1}{2}$ and then

$$e^{-\frac{1}{2}\sigma_{n,k}^2t^2-\frac{1}{4}\sigma_{n,k}^4t^4}\le 1-\frac{1}{2}\sigma_{n,k}^2t^2\le e^{-\frac{1}{2}\sigma_{n,k}^2t^2}.$$

Take product over $k\le n$, and observe that $\sum_{k=1}^n\sigma_{n,k}^4\to 0$ (why?). Hence,

$$\prod_{k=1}^{n}\left(1-\frac{1}{2}\sigma_{n,k}^2t^2\right)\to e^{-\frac{\sigma^2t^2}{2}}.$$

From 18 and Lévy's continuity theorem, we get $\sum_{k=1}^n X_{n,k}\xrightarrow{d}N(0,\sigma^2)$. ■

19.2. **Proof of Lindeberg-Feller CLT by replacement method.**

Proof. As before, without loss of generality, we assume that on the same probability space as the random variables $X_{n,k}$ we also have the Gaussian random variables $Y_{n,k}$ that are independent among themselves and independent of all the $X_{n,k}$s and further satisfy $\mathbf{E}[Y_{n,k}]=\mathbf{E}[X_{n,k}]$ and $\mathbf{E}[Y_{n,k}^2]=\mathbf{E}[X_{n,k}^2]$.

Similarly to the earlier proof of CLT, fix n and define $U_k=\sum_{j=1}^{k-1}X_{n,j}+\sum_{j=k+1}^{n}Y_{n,j}$ and $V_k=\sum_{j=1}^{k}X_{n,j}+\sum_{j=k+1}^{n}Y_{n,j}$ for $0\le k\le n$. Then, $V_0=Y_{n,1}+\ldots+Y_{n,n}$ and $V_n=X_{n,1}+\ldots+X_{n,n}$. Also, $V_n\sim N(0,\sigma^2)$. Thus,

$$\begin{aligned}\mathbf{E}\left[f\left(V_n\right)\right]-\mathbf{E}[f(V_0)]&=\sum_{k=1}^{n}\mathbf{E}\left[f\left(V_k\right)-f\left(V_{k-1}\right)\right]\\&=\sum_{k=1}^{n}\mathbf{E}\left[f\left(V_k\right)-f\left(U_k\right)\right]-\sum_{k=1}^{n}\mathbf{E}\left[f\left(V_{k-1}\right)-f\left(U_k\right)\right].\end{aligned}\tag{19}$$

We expand $f(V_k)-f(U_k)$ by Taylor series, both of third order and second order and write

$$f(V_k)-f(U_k)=f'(U_k)X_{n,k}+\frac{1}{2}f''(U_k)X_{n,k}^2+\frac{1}{6}f'''(U_k^*)X_{n,k}^3,$$

$$f(V_k)-f(U_k)=f'(U_k)X_{n,k}+\frac{1}{2}f''(U_k^{\#})X_{n,k}^2$$

where U_k^* and $U_k^\#$ are between V_k and U_k. Write analogous expressions for $f(V_{k-1}) - f(U_k)$ (observe that $V_{k-1} = U_k + Y_{n,k}$) and subtract from the above to get

$$f(V_k) - f(V_{k-1}) = f'(U_k)(X_{n,k} - Y_{n,k}) + \frac{1}{2} f''(U_k)(X_{n,k}^2 - Y_{n,k}^2) + \frac{1}{6}(f'''(U_k^*)X_{n,k}^3 - f'''(U_k^{**})Y_{n,k}^3),$$

$$f(V_k) - f(V_{k-1}) = f'(U_k)(X_{n,k} - Y_{n,k}) + \frac{1}{2}(f''(U_k^\#)X_{n,k}^2 - f''(U_k^{\#\#})Y_{n,k}^2).$$

Use the first one when $|X_{n,k}| \le \delta$ and the second one when $|X_{n,k}| > \delta$ and take expectations to get

$$|\mathbf{E}[f(V_k)] - \mathbf{E}[f(V_{k-1})]| \le \frac{1}{2}\mathbf{E}[|f''(U_k)|] \left|\mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}| \le \delta}] - \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}| \le \delta}]\right| \tag{20}$$

$$+ \frac{1}{2}\left|\mathbf{E}[|f''(U_k^\#)| X_{n,k}^2 \mathbf{1}_{|X_{n,k}| > \delta}]\right| + \frac{1}{2}\left|\mathbf{E}[|f''(U_k^{\#\#})| Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}| > \delta}]\right| \tag{21}$$

$$+ \frac{1}{6}\left|\mathbf{E}[|f'''(U_k^*)| |X_{n,k}|^3 \mathbf{1}_{|X_{n,k}| \le \delta}]\right| + \frac{1}{6}\left|\mathbf{E}[|f'''(U_k^{**})| |Y_{n,k}|^3 \mathbf{1}_{|Y_{n,k}| \le \delta}]\right| \tag{22}$$

Since $\mathbf{E}[X_{n,k}^2] = \mathbf{E}[Y_{n,k}^2]$, the term in the first line (20) is the same as $\frac{1}{2}\mathbf{E}[|f''(U_k)|] \left|\mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] - \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\right|$ which in turn is bounded by

$$C_f\{\mathbf{E}[|X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] + \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\}.$$

The terms in (21) are also bounded by

$$C_f\{\mathbf{E}[|X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] + \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\}.$$

To bound the two terms in (22), we show how to deal with the first.

$$\left|\mathbf{E}[|f'''(U_k^*)| |X_{n,k}|^3 \mathbf{1}_{|X_{n,k}| \le \delta}]\right| \le C_f \delta \mathbf{E}[X_{n,k}^2].$$

The same bound holds for the second term in (22). Putting all this together, we arrive at

$$|\mathbf{E}[f(V_k)] - \mathbf{E}[f(V_{k-1})]| \le C_f\{\mathbf{E}[|X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] + \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\} + \delta\{\mathbf{E}[|X_{n,k}^2] + \mathbf{E}[Y_{n,k}^2]\}.$$

Add up over k and use (19) to get

$$\begin{aligned}\left|\mathbf{E}\left[f\left(V_n\right)\right] - \mathbf{E}[f(V_0)]\right| \le \delta \sum_{k=1}^{n} \mathbf{E}[|X_{n,k}^2] + \mathbf{E}[Y_{n,k}^2] \\ + C_f \sum_{k=1}^{n} \mathbf{E}[|X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] + \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}].\end{aligned}$$

As $n \to \infty$, the first term on the right goes to $2\delta\sigma^2$. The second term goes to zero. This follows directly from the assumptions for the terms involving X whereas for the terms involving Y (which are Gaussian), it is a matter of checking that the same conditions do hold for Y.

Consequently, we get $\limsup \left|\mathbf{E}[f(V_0)] - \mathbf{E}[f(V_n)]\right| \leq 2\sigma^2\delta$. As δ is arbitrary, we have shown that for any $f \in C_b^{(3)}(\mathbb{R})$, we have

$$\mathbf{E}[f(X_{n,1} + \ldots + X_{n,n})] \to \mathbf{E}[f(Z)]$$

where $Z \sim N(0, \sigma^2)$. This completes the proof that $X_{n,1} + \ldots + X_{n,n} \xrightarrow{d} N(0, \sigma^2)$. ■

20. SUMS OF MORE HEAVY-TAILED RANDOM VARIABLES

Let X_i be an i.i.d sequence of real-valued r.v.s. If the second moment is finite, we have see that the sums S_n converge to Gaussian distribution after shifting (by $n\mathbf{E}[X_1]$) and scaling (by $\sqrt{n}$). What if we drop the assumption of second moments? Let us first consider the case of Cauchy random variables to see that such results may be expected in general.

Example 55. *Let X_i be i.i.d Cauchy(1), with density $\frac{1}{\pi(1+x^2)}$. Then, one can check that $\frac{S_n}{n}$ has exactly the same Cauchy distribution! Thus, to get distributional convergence, we just write $\frac{S_n}{n} \xrightarrow{d} C_1$. If X_i were i.i.d with density $\frac{a}{\pi(a^2+(x-b)^2)}$ (which can be denoted $C_{a,b}$ with $a > 0$, $b \in \mathbb{R}$), then $\frac{X_i - b}{a}$ are i.i.d C_1, and hence, we get*

$$\frac{S_n - nb}{an} \xrightarrow{d} C_1.$$

This is the analogue of CLT, except that the location change is nb instead of $n\mathbf{E}[X_1]$, scaling is by n instead of $\sqrt{n}$ and the limit is Cauchy instead of Normal.

This raises the following questions.

(1) For general i.i.d sequences, how are the location and scaling parameter determined, so that $b_n^{-1}(S_n - a_n)$ converges in distribution to a non-trivial measure on the line?

(2) What are the possible limiting distributions?

(3) What are the *domains of attraction* for each possible limiting distribution, e.g., for what distributions on X_1 do we get $b_n^{-1}(S_n - a_n) \xrightarrow{d} C_1$?

For simplicity, let us restrict ourselves to symmetric distributions, i.e., $X \overset{d}{=} -X$. Then, clearly no shifting is required, $a_n = 0$. Let us investigate the issue of scaling and what might be the limit.

It turns out that for each $\alpha \leq 2$, there is a unique (up to scaling) distribution μ_α such that $X + Y \overset{d}{=} 2^{\frac{1}{\alpha}} X$ if $X, Y \sim \mu$ are independent. This is known as the symmetric α-stable distribution and has characteristic function $\psi_\alpha(t) = e^{-c|t|^\alpha}$. For example, the normal distribution corresponds to $\alpha = 2$ and the Cauchy to $\alpha = 1$. If X_i are i.i.d μ_α, then is is easy to see that $n^{-1/\alpha} S_n \xrightarrow{d} \mu_\alpha$. The fact is that there is a certain domain of attraction for each stable distribution, and for i.i.d random variables from any such distribution $n^{-1/\alpha} S_n \xrightarrow{d} \mu_\alpha$.

21. Appendix: Characteristic functions and their properties

Definition 56. Let μ be a probability measure on $\mathbb{R}$. The function $\psi_\mu : \mathbb{R}^d \to \mathbb{R}$ define by $\psi_\mu(t) := \int_{\mathbb{R}} e^{itx} d\mu(x)$ is called the *characteristic function*[5] or the *Fourier transform* of μ. If X is a random variable on a probability space, we sometimes say "characteristic function of X" to mean the c.f of its distribution (thus $\psi_X(t) = \mathbf{E}[e^{itX}]$). We also write $\hat{\mu}$ instead of ψ_μ.

There are various other "integral transforms" of a measure that are closely related to the c.f. For example, if we take $\psi_\mu(it)$ is the moment generating function of μ (if it exists). For μ supported on $\mathbb{N}$, its so called generating function $F_\mu(t) = \sum_{k\geq 0} \mu\{k\} t^k$ (which exists for $|t| < 1$ since μ is a probability measure) can be written as $\psi_\mu(-i \log t)$ (at least for $t > 0$!) etc. The characteristic function has the advantage that it exists for all $t \in \mathbb{R}$ and for all finite measures μ.

The importance of c.f comes from the following facts.

(A) It transforms well under certain operations of measures, such as shifting a scaling and under convolutions.

(B) The c.f. determines the measure. Further, the smoothness of the characteristic function encodes the tail decay of the measure, and vice versa.

(C) $\hat{\mu}_n(t) \to \hat{\mu}(t)$ pointwise, if and only if $\mu_n \xrightarrow{d} \mu$. This is the key property that was used in proving central limit theorems.

(D) There exist necessary and sufficient conditions for a function $\psi : \mathbb{R} \to \mathbb{C}$ to be the c.f o f a measure. Because of this and part (B), sometimes one defines a measure by its characteristic function.

(A) Transformation rules and some examples.

Theorem 57. *Let X, Y be random variables.*

(1) For any $a, b \in \mathbb{R}$, we have $\psi_{aX+b}(t) = e^{ibt}\psi_X(at)$.

(2) If X, Y are independent, then $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$.

Proof. (1) $\psi_{aX+b}(t) = \mathbf{E}[e^{it(aX+b)}] = \mathbf{E}[e^{itaX}]e^{ibt} = e^{ibt}\psi_X(at)$.

(2) $\psi_{X+Y}(t) = \mathbf{E}[e^{it(X+Y)}] = \mathbf{E}[e^{itX}e^{itY}] = \mathbf{E}[e^{itX}]\mathbf{E}[e^{itY}] = \psi_X(t)\psi_Y(t)$.

■

We give some examples.

(1) If $X \sim \text{Ber}(p)$, then $\psi_X(t) = pe^{it} + q$ where $q = 1 - p$. If $Y \sim \text{Binomial}(n, p)$, then, $Y \overset{d}{=} X_1 + \ldots + X_n$ where X_k are i.i.d Ber(p). Hence, $\psi_Y(t) = (pe^{it} + q)^n$.

[5] In addition to the usual references, Feller's *Introduction to probability theory and its applications: vol II*, chapter XV, is an excellent resource for the basics of characteristic functions. Our presentation is based on it too.

(2) If $X \sim \text{Exp}(\lambda)$, then $\psi_X(t) = \int_0^\infty \lambda e^{-\lambda x} e^{itx} dx = \frac{\lambda}{\lambda - it}$. If $Y \sim \text{Gamma}(\nu, \lambda)$, then if ν is an integer, then $Y \overset{d}{=} X_1 + \ldots + X_\nu$ where X_k are i.i.d Exp(λ). Therefore, $\psi_Y(t) = \frac{\lambda^\nu}{(\lambda - it)^\nu}$. This is true even if ν is not an integer, but the proof would have to be a direct computation.

(3) $Y \sim \text{Normal}(\mu, \sigma^2)$. Then, $Y = \mu + \sigma X$, where $X \sim N(0,1)$ and by the transofrmatin rules, $\psi_Y(t) = e^{i\mu t}\psi_X(\sigma t)$. Thus it suffices to find the c.f of $N(0,1)$. Denote it by ψ.

$$\psi(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{itx} e^{-\frac{x^2}{2}} dx = e^{-\frac{t^2}{2}} \left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}} dx \right).$$

It appears that the stuff inside the brackets is equal to 1, since it looks like the integral of a normal density with mean it and variance σ^2. But if the mean is complex, what does it mean?! Using contour integration, one can indeed give a rigorous proof that the stuff inside brackets is indeed equal to 1.

Alternately, one can justify differentiation under the integral sign to get

$$\psi'(t) = \frac{i}{\sqrt{2\pi}} \int x e^{itx} e^{-\frac{x^2}{2}} dx.$$

Then, justify differentiation under, in class using contour integration, which will not be repeated here. The final concusion is that $N(\mu, \sigma^2)$ has c.f $e^{it\mu - \frac{\sigma^2 t^2}{2}}$.

(B) Continuity properties. The following lemma gives some basic properties of a c.f.

Lemma 58. *Let $\mu \in \mathcal{P}(\mathbb{R})$. Then, $\hat{\mu}$ is a uniformly continuous function on $\mathbb{R}$ with $|\hat{\mu}(t)| \le 1$ for all t with $\hat{\mu}(0) = 1$. (equality may be attained elsewhere too).*

Proof. Clearly $\hat{\mu}(0) = 1$ and $|\hat{\mu}(t)| \le 1$. Further,

$$|\hat{\mu}(\lambda + h) - \hat{\mu}(\lambda)| \le \int |e^{ix(\lambda+h)} - e^{i\lambda x}| d\mu(x) = \int |e^{ihx} - 1| d\mu(x).$$

The last quantity does not depend on λ. Further, the integrand approaches 0 as $h \to 0$ while $|e^{ihx} - 1| \le 2$ gives the domination required to conclude that the integral goes to 0 as $h \to 0$. ■

The more we assume about the continuity/smoothness of the measure μ, the stronger the conclusion that can be drawn about the decay of $\hat{\mu}$. And conversely, if the tail of μ decays fast, the smoother $\hat{\mu}$ will be. We used this latter fact in the proof of central limit theorems.

Lemma 59. *Let $T_n(\theta) = \sum_{k=0}^n \frac{(i\theta)^k}{k!}$ be the nth order Taylor series for $e^{i\theta}$. Then, for any $n \ge 0$ and any $\theta \in \mathbb{R}$, we have*

$$|e^{i\theta} - T_n(\theta)| \le \begin{cases} \frac{|\theta|^{n+1}}{(n+1)!} \\ \frac{2|\theta|^n}{n!} \end{cases} \tag{23}$$

Proof. Observe that

$$\frac{d}{d\theta}e^{i\theta} = ie^{i\theta}, \quad \frac{d}{d\theta}T_n(\theta) = iT_{n-1}(\theta).$$

Therefore, for any $\theta \in \mathbb{R}$,

$$|e^{i\theta} - T_n(\theta)| \le \int_0^\theta |e^{i\varphi} - T_{n-1}(\varphi)| d\varphi.$$

Inductively, this gives the inequalities in (23). In the base case $n = 0$, we have $|e^{i\theta} - 1| = 2|\sin(\theta/2)| \le 2 \wedge |\theta|$. ■

Lemma 60. *Let X be a random variable with $\mathbf{E}|X|^p < \infty$ for some positive integer p. Then, as $\lambda \to 0$,*

$$\psi_X(\lambda) = \sum_{k=0}^{p} \frac{(i\lambda)^k}{k!}\mathbf{E}[X^k] + o(\lambda^{p+1}).$$

In particular, $\psi_X^{(k)}(0) = i^k \mathbf{E}[X^k]$.

Proof. Put $\theta = \lambda X$ in (23) and take expectations to get

$$\left|\mathbf{E}[e^{i\lambda X}] - \sum_{k=0}^{p} \frac{(i\lambda)^k}{k!}\mathbf{E}[X^k]\right| \le \frac{|\lambda|^p}{p!}\mathbf{E}\left[|X|^p \left(\frac{|\lambda X|}{p+1} \wedge 2\right)\right].$$

We need to show that $\mathbf{E}\left[|X|^p \left(\frac{|\lambda X|}{p+1} \wedge 2\right)\right] \to 0$ as $\lambda \to 0$. This follows from DCT since the integrand goes to zero almost surely as $\lambda \to 0$ and is bounded by $2|X|^p$ which is integrable. ■

We stated only what we need. But the theme here can be developed further.

- If $\mathbf{E}|X|^p < \infty$, a little more work shows that $\psi \in C^{(p)}(\mathbb{R})$ and $\psi^{(k)}(\lambda) = i^k \mathbf{E}[X^k e^{i\lambda X}]$ for $k \le p$.
- Conversely, if a characteristic function is differentiable p times, it can be shown that $\mathbf{E}|X|^{p-1} < \infty$.
- The above facts show that the decay of the tail of a measure is encoded in the smoothness of the characteristic function.
- Conversely, the decay of the characteristic function encodes the smoothness of the measure. For example, if μ has a density then $\hat{\mu}(\lambda) \to 0$ as $\lambda \to \pm\infty$ (Riemann-Lebesgue lemma). If the density is differentiable p times, then $\hat{\mu}(\lambda) = o(|\lambda|^{-p})$ etc. Conversely, we have seen that if $\hat{\mu}$ is integrable, then μ has a density (Fourier inversion formula). If it decays faster, one can deduce further smoothness of the density.

For proofs, consult, Feller's book.

(C) Inversion formulas.

Theorem 61. *If $\hat{\mu} = \hat{\nu}$, then $\mu = \nu$.*

Proof. Let θ_σ denote the $\mathrm{N}(0, \sigma^2)$ distribution and let $\varphi_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$ and $\Phi_\sigma(x) = \int_{-\infty}^{x} \varphi_\sigma(u)du$ and $\hat{\theta}_\sigma(t) = e^{-\sigma^2 t^2/2}$ denote the density and cdf and characteristic functions, respectively. Then, by Parseval's identity, we have for any α,

$$\begin{aligned} \int e^{-i\alpha t}\hat{\mu}(t)d\theta_\sigma(t) &= \int \hat{\theta}_\sigma(x-\alpha)d\mu(x) \\ &= \frac{\sqrt{2\pi}}{\sigma}\int \varphi_{\frac{1}{\sigma}}(\alpha - x)d\mu(x) \end{aligned}$$

where the last line comes by the explicit Gaussian form of $\hat{\theta}_\sigma$. Let $f_\sigma(\alpha) := \frac{\sigma}{\sqrt{2\pi}} \int e^{-i\alpha t}\hat{\mu}(t)d\theta_\sigma(t)$ and integrate the above equation to get that for any finite $a < b$,

$$\begin{aligned} \int_a^b f_\sigma(\alpha)d\alpha &= \int_a^b \int_{\mathbb{R}} \varphi_{\frac{1}{\sigma}}(\alpha - x)\; d\mu(x)\; d\alpha \\ &= \int_{\mathbb{R}} \int_a^b \varphi_{\frac{1}{\sigma}}(\alpha - x)\; d\alpha\; d\mu(x) \qquad \text{(by Fubini)} \\ &= \int_{\mathbb{R}} \Big(\Phi_{\frac{1}{\sigma}}(b-x) - \Phi_{\frac{1}{\sigma}}(a-x)\Big)\, d\mu(x). \end{aligned}$$

Now, we let $\sigma \to \infty$, and note that

$$\Phi_{\frac{1}{\sigma}}(u) \to \begin{cases} 0 & \text{if } u < 0. \\ 1 & \text{if } u > 0. \\ \frac{1}{2} & \text{if } u = 0. \end{cases}$$

Further, $\Phi_{\sigma^{-1}}$ is bounded by 1. Hence, by DCT, we get

$$\lim_{\sigma\to\infty} \int_a^b f_\sigma(\alpha)d\alpha = \int \left[\mathbf{1}_{(a,b)}(x) + \frac{1}{2}\mathbf{1}_{\{a,b\}}(x)\right] d\mu(x) = \mu(a,b) + \frac{1}{2}\mu\{a,b\}.$$

Now we make two observations: (a) that f_σ is determined by $\hat{\mu}$, and (b) that the measure μ is determined by the values of $\mu(a,b) + \frac{1}{2}\mu\{a,b\}$ for all finite $a < b$. Thus, $\hat{\mu}$ determines the measure μ. ■

Corollary 62 (Fourier inversion formula). *Let $\mu \in \mathcal{P}(\mathbb{R})$.*

(1) For all finite $a < b$, we have

$$\mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\} = \lim_{\sigma\to\infty} \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-iat} - e^{-ibt}}{it} \hat{\mu}(t) e^{-\frac{t^2}{2\sigma^2}} dt \tag{24}$$

(2) If $\int_{\mathbb{R}} |\hat{\mu}(t)|dt < \infty$, then μ has a continuous density given by

$$f(x) := \frac{1}{2\pi}\int_{\mathbb{R}} \hat{\mu}(t)e^{-ixt}dt.$$

Proof. (1) Recall that the left hand side of (24) is equal to $\lim_{\sigma\to\infty} \int_a^b f_\sigma$ where

$$f_\sigma(\alpha) := \frac{\sigma}{\sqrt{2\pi}}\int e^{-i\alpha t}\hat{\mu}(t)d\theta_\sigma(t).$$

Writing out the density of θ_σ we see that

$$\begin{aligned}\int_a^b f_\sigma(\alpha)d\alpha &= \frac{1}{2\pi}\int_a^b\int_{\mathbb{R}} e^{-i\alpha t}\hat{\mu}(t)e^{-\frac{t^2}{2\sigma^2}}dtd\alpha \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_a^b e^{-i\alpha t}\hat{\mu}(t)e^{-\frac{t^2}{2\sigma^2}}d\alpha\, dt \quad \text{(by Fubini)} \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-iat}-e^{-ibt}}{it}\hat{\mu}(t)e^{-\frac{t^2}{2\sigma^2}}dt.\end{aligned}$$

Thus, we get the first statement of the corollary.

(2) With f_σ as before, we have $f_\sigma(\alpha) := \frac{1}{2\pi}\int e^{-i\alpha t}\hat{\mu}(t)e^{-\frac{t^2}{2\sigma^2}}dt$. Note that the integrand converges to $e^{-i\alpha t}\hat{\mu}(t)$ as $\sigma\to\infty$. Further, this integrand is bounded by $|\hat{\mu}(t)|$ which is assumed to be integrable. Therefore, by DCT, for any $\alpha\in\mathbb{R}$, we conclude that $f_\sigma(\alpha)\to f(\alpha)$ where $f(\alpha) := \frac{1}{2\pi}\int e^{-i\alpha t}\hat{\mu}(t)dt$.

Next, note that for any $\sigma > 0$, we have $|f_\sigma(\alpha)| \le C$ for all α where $C = \int|\hat{\mu}(t)|dt$. Thus, for finite $a < b$, using DCT again, we get $\int_a^b f_\sigma \to \int_a^b f$ as $\sigma\to\infty$. But the proof of Theorem 61 tells us that

$$\lim_{\sigma\to\infty}\int_a^b f_\sigma(\alpha)d\alpha = \mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}.$$

Therefore, $\mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\} = \int_a^b f(\alpha)d\alpha$. Fixing a and letting $b \downarrow a$, this shows that $\mu\{a\} = 0$ and hence $\mu(a,b) = \int_a^b f(\alpha)d\alpha$. Thus f is the density of μ.

The proof that a c.f. is continuous carries over verbatim to show that f is continuous (since f is the Furier trnasform of $\hat{\mu}$, except for a change of sign in the exponent). ■

An application of Fourier inversion formula Recall the Cauchy distribution μ with with density $\frac{1}{\pi(1+x^2)}$ whose c.f is not easy to find by direct integration (Residue theorem in complex analysis is a way to compute this integral).

Consider the seemingly unrelated p.m ν with density $\frac{1}{2}e^{-|x|}$ (a symmetrized exponential, this is also known as Laplace's distribution). Its c.f is easy to compute and we get

$$\hat{\nu}(t) = \frac{1}{2}\int_0^{\infty} e^{itx-x}dx + \frac{1}{2}\int_{-\infty}^{0} e^{itx+x}dx = \frac{1}{2}\left(\frac{1}{1-it}+\frac{1}{1+it}\right) = \frac{1}{1+t^2}.$$

By the Fourier inversion formula (part (b) of the corollary), we therefore get

$$\frac{1}{2}e^{-|x|} = \frac{1}{2\pi}\int \hat{\nu}(t)e^{itx}dt = \frac{1}{2\pi}\int \frac{1}{1+t^2}e^{itx}dt.$$

This immediately shows that the Cauchy distribution has c.f. $e^{-|t|}$ without having to compute the integral!

(D) Continuity theorem. Now we come to the key result that was used in the proof of central limit theorems. This is the equivalence between convergence in distribution and pointwise convergence of characteristic functions.

Theorem 63. *Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$.*

(1) If $\mu_n \xrightarrow{d} \mu$ then $\hat{\mu}_n(t) \to \hat{\mu}(t)$ pointwise for all t.

(2) If $\hat{\mu}_n(t) \to \psi(t)$ pointwise for all t, then $\psi = \hat{\mu}$ for some $\mu \in \mathcal{P}(\mathbb{R})$ and $\mu_n \xrightarrow{d} \mu$.

Proof. (1) If $\mu_n \xrightarrow{d} \mu$, then $\int f d\mu_n \to \int f d\mu$ for any $f \in C_b(\mathbb{R})$ (bounded continuous function). Since $x \to e^{itx}$ is a bounded continuous function for any $t \in \mathbb{R}$, it follows that $\hat{\mu}_n(t) \to \hat{\mu}(t)$ pointwise for all t.

(2) Now suppose $\hat{\mu}_n(t) \to \hat{\mu}(t)$ pointwise for all t. We first claim that the sequence $\{\mu_n\}$ is tight. Assuming this, the proof can be completed as follows.

Let μ_{n_k} be any subsequence that converges in distribution, say to ν. By tightness, $\nu \in \mathcal{P}(\mathbb{R})$. Therefore, by the first part, $\hat{\mu}_{n_k} \to \hat{\nu}$ pointwise. But obviously, $\hat{\mu}_{n_k} \to \hat{\mu}$ since $\hat{\mu}_n \to \hat{\mu}$. Thus, $\hat{\nu} = \hat{\mu}$ which implies that $\nu = \mu$. That is, any convergent subsequence of $\{\mu_n\}$ converges in distribution to μ. This shows that $\mu_n \xrightarrow{d} \mu$.

It remains to show tightness. From Lemma 64 below, as $n \to \infty$,

$$\mu_n\left([-2/\delta, 2/\delta]^c\right) \;\; \leq \;\; \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\hat{\mu}_n(t))dt \longrightarrow \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\hat{\mu}(t))dt$$

where the last implication follows by DCT (since $1-\hat{\mu}_n(t) \to 1-\hat{\mu}(t)$ for each t and also $|1-\hat{\mu}_n(t)| \leq 2$ for all t). Further, as $\delta \downarrow 0$, we get $\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\hat{\mu}(t))dt \to 0$ (because, $1-\hat{\mu}(0) = 0$ and $\hat{\mu}$ is continuous at 0).

Thus, given $\epsilon > 0$, we can find $\delta > 0$ such that $\limsup_{n\to\infty} \mu_n\left([-2/\delta, 2/\delta]^c\right) < \epsilon$. This means that for some finite N, we have $\mu_n\left([-2/\delta, 2/\delta]^c\right) < \epsilon$ for all $n \geq N$. Now, find

$A > 2/\delta$ such that for any $n \leq N$, we get $\mu_n\left([-2/\delta, 2/\delta]^c\right) < \epsilon$. Thus, for any $\epsilon > 0$, we have produced an $A > 0$ so that $\mu_n\left([-A, A]^c\right) < \epsilon$ for all n. This is the definition of tightness. ■

Lemma 64. *Let $\mu \in \mathcal{P}(\mathbb{R})$. Then, for any $\delta > 0$, we have*

$$\mu\left(\left[-\frac{2}{\delta}, \frac{2}{\delta}\right]^c\right) \leq \frac{1}{\delta}\int\limits_{-\delta}^{\delta}(1-\hat{\mu}(t))dt.$$

Proof. We write

$$\begin{aligned}\int_{-\delta}^{\delta}(1-\hat{\mu}(t))dt &= \int_{-\delta}^{\delta}\int_{\mathbb{R}}(1-e^{itx})d\mu(x)dt \\ &= \int_{\mathbb{R}}\int_{-\delta}^{\delta}(1-e^{itx})dtd\mu(x) \\ &= \int_{\mathbb{R}}\left(2\delta - \frac{2\sin(x\delta)}{x}\right)d\mu(x) \\ &= 2\delta\int_{\mathbb{R}}\left(1-\frac{\sin(x\delta)}{x\delta}\right)d\mu(x).\end{aligned}$$

When $\delta|x| > 2$, we have $\frac{\sin(x\delta)}{x\delta} \leq \frac{1}{2}$ (since $\sin(x\delta) \leq 1$). Therefore, the integrand is at least $\frac{1}{2}$ when $|x| > \frac{2}{\delta}$ and the integrand is always non-negative since $|\sin(x)| \leq |x|$. Therefore we get

$$\int_{-\delta}^{\delta}(1-\hat{\mu}(t))dt \geq \delta\mu\left([-2/\delta, 2/\delta]^c\right). \qquad ■$$