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].

[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. THE BASIC SET UP FOR PROBABILITY

A *random experiment* is an undefined but intuitively unambiguous term that conveys the idea of an "experiment" that can have one of multiple outcomes, and which one actually occurs is unpredictable. The first question in making a theory of probability is to give a mathematical definition that can serve as a model for the real-world notion of a random experiment.

In basic probability class we have already seen how to do this, provided the number of outcomes is finite or countably infinite. This is how it is done.

Definition 1: Discrete probability space

A discrete probability space is a pair (Ω, p), where Ω is a non-empty countable set and $p : \Omega \to [0,1]$ is a function such that $\sum_{\omega\in\Omega} p(\omega) = 1$. Then define $\mathbf{P} : 2^{\Omega} \to [0,1]$ by $\mathbf{P}(A) = \sum_{\omega\in A} p(\omega)$.

The set Ω is called the *sample space* (the collection of all possible outcomes), $p(\omega)$ are called *elementary probabilities*, subsets of Ω are called *events*, and $\mathbf{P}(A)$ is said to be the *probability of the event* A. The way this mathematical notion is supposed to represent a random experiment is familiar. We just illustrate with a few examples.

Example 1: A coin is tossed n times

Then $\Omega = \{0,1\}^n$ where if $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ denotes the outcome where the ith toss is a head if $\omega_i = 1$ and a tail if $\omega_i = 0$. Further, $p(\omega) = p^{\omega_1+\ldots+\omega_n}(1-p)^{n-\omega_1-\ldots-\omega_n}$ (this assignment incorporates the idea that distinct tosses are 'independent'). An example of the event of getting k heads exactly, i.e., $A = \{\omega : \omega_1 + \ldots + \omega_n = k\}$, which has probability $\mathbf{P}(A) = \binom{n}{k} p^k (1-p)^{n-k}$.

Example 2: r balls are thrown into n bins at random

Then $\Omega = [n]^r$ where $[n] = \{1, \ldots, n\}$. Here $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ denotes the outcome where the ith ball goes into the bin numbered ω_i. Elementary probabilities are defined by $p(\omega) = n^{-r}$. An example of an event is that the first bin is empty, i.e., $A = \{\omega : \omega_i \neq 1 \text{ for all i}\}$, and it has probability $\mathbf{P}(A) = (n-1)^r/n^r$.

But when the number of possible outcomes is uncountable, this framework does not suffice. Three examples:

(1) A glass rod falls and breaks into two pieces.

(2) A fair coin is tossed infinitely many times.

(3) A dart is thrown at a circular dart board.

If Ω denotes the sample space (the set of all possible outcomes), then in the above cases it must respectively be equal to

(1) $[0,1]$, where we think of the glass rod as the line segment $[0,1]$ and the outcome denoting the point in $[0,1]$ where the breakage occurs,

(2) $\{0,1\}^{\mathbb{N}}$, where $\omega = (\omega_1, \omega_2, \ldots)$ denotes the outcome where the kth toss turns up ω_k (always 1 denotes heads and 0 denotes tails),

(3) $\{(x,y) : x^2 + y^2 \leq 1\}$, where the point (x,y) denotes the location where the dart hits the dartboard.

In all three cases Ω is uncountable. We also agree on the probabilities of many events, for example that $[0.1, 0.35]$ and $\{\omega \in \{0,1\}^{\mathbb{N}} : \omega_1 = 1,\ \omega_2 = 0\}$ and $\{(x,y) : x > 0 > y\}$ in the three examples all have probability $\frac{1}{4}$. But where that comes from? If any elementary probability has to be assigned to singletons, it can only be zero, and there is no unambiguous meaning to adding uncountably many zeros to get $\frac{1}{4}$. So we need a new framework.

The first example is clearly the same as the issue of assigning lengths to subsets of the line, and in measure theory class we have seen that it can be done satisfactorily by giving up the idea of assigning length to every subset. As recompense, we get a notion of length that is not just finitely, but countably additive. This framework exactly fits our need.

Definition 2: Probability space

A probability space is a triple $(\Omega, \mathcal{F}, \mathbf{P})$ where

- Ω is a non-empty set,
- $\mathcal{F}$ is a sigma algebra of subsets of Ω. That is, $\mathcal{F} \subseteq 2^{\Omega}$; $\emptyset \in \mathcal{F}$; $A \in \mathcal{F} \implies A^c \in \mathcal{F}$; $A_n \in \mathcal{F} \implies \cup_n A_n \in \mathcal{F}$.
- $\mathbf{P}$ is a probability measure on $\mathcal{F}$. That is $\mathbf{P} : \mathcal{F} \to [0,1]$ and $\mathbf{P}(\sqcup A_n) = \sum_n \mathbf{P}(A_n)$ if $A_n \in \mathcal{F}$ are pairwise disjoint, and $\mathbf{P}(\Omega) = 1$.

Observe that n will always indicate a countable indexing (may start at 0 or 1 or vary over all integers). For $A \in \mathcal{F}$, we say that $\mathbf{P}(A)$ is the probability of A. We do not talk of the probability of sets not in the sigma algebra. This framework will form the basis of all probability.

To return to the modeling of random experiments, what the sample space should be is usually clear, as we have seen. What sigma-algebra to take? Except for the trivial sigma-algebras 2^{Ω} and $\{\emptyset, \Omega\}$, all sigma-algebras of interest arise as follows.

Definition 3: Generated sigma-algebra

Let $\mathcal{S}$ be a collection of subsets of Ω. The smallest sigma-algebra containing $\mathcal{S}$, also called the sigma-algebra generated by $\mathcal{S}$, exists and is defined as

$$\sigma(\mathcal{S}) = \bigcap_{\mathcal{F} \supseteq \mathcal{S}} \mathcal{F},$$

where the intersection is over all sigma-algebras that contain $\mathcal{S}$.

Into $\mathcal{S}$ we put in all subsets which we definitely wish to define probabilities for, and then take $\sigma(\mathcal{S})$ as our sigma-algebra. For example, in the stick-breaking example, we may take $\mathcal{S}$ to be the collection of all intervals in $[0,1]$. That is called the Borel sigma-algebra on $[0,1]$ and denoted $\mathcal{B}$ or $\mathcal{B}_{[0,1]}$. This is one of the most important sigma-algebras for us, so let us define it in general.

Definition 4: Borel sigma-algebra

Let X be a metric space. The smallest sigma-algebra containing all open sets is called the Borel sigma-algebra of X and denoted $\mathcal{B}_X$.

Many different collections of subsets can give rise to the same sigma-algebra. For example, the collection of closed subsets also generates $\mathcal{B}_X$. If $X = \mathbb{R}$, the collection of intervals, the collection of intervals with rational end-points, the collection of compact sets, all these generate $\mathcal{B}_{\mathbb{R}}$ (exercise!).

Now that we are clear how the sigma-algebra associated to a random experiment is obtained, the question remains of the probability measure. We have Ω, a collection of subsets $\mathcal{S}$, and the sigma-algebra $\sigma(\mathcal{S})$. By symmetry considerations or experiments or something else, we know what probability of events in $\mathcal{S}$ ought to be. So the primary question of designing a probability space reduces to this:

Question 1: Extension of probability

Given $P : \mathcal{S} \to [0,1]$, does there exist a probability measure $\mathbf{P}$ on $\sigma(\mathcal{S})$ such that $\mathbf{P}(A) = P(A)$ for $A \in \mathcal{S}$. If so, is it unique?

The answer to this comes from the construction of measures in measure theory. As it turns out, for our purposes it suffices to assume the existence of Lebesgue measure, and everything else follows from that.

Example 3: Break a stick at random

Here $\Omega = [0,1]$, the sigma algebra is $\mathcal{B}$ the collection of all Borel subsets of $[0,1]$ and the probability measure is λ, the Lebesgue measure on $[0,1]$. It is a non-trivial fact that there is a unique measure λ on $\mathcal{B}$ such that $\lambda([a,b]) = b - a$ whenever $[a,b] \subseteq [0,1]$.

Similarly the dart throwing can be captured by taking the sample space to be $\mathbb{D} = \{(x, y) : x^2 + y^2 < 1\}$ and the Borel sigma algebra of $\mathbb{D}$ and the two-dimensional Lebesgue measure on $\mathbb{D}$ (normalized by $1/\pi$). How to make sense of tossing infinitely many coins? We could invoke yet another theorem in measure theory, or more precisely the method of construction of measures via outer measures etc. Conveniently for us, we can use the stick-breaking probability space and create many other probability spaces, including the one for tossing a coin infinitely many times. Let us introduce this notion first.

Definition 5: Measurable function

Let $\mathcal{F}$ be a sigma-algebra on X and let $\mathcal{G}$ be a sigma-algebra on Y. A map $T : X \to Y$ is said to be *measurable* if $T^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{G}$.

Lemma 1: Push-forward measure

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{G}$ be a sigma-algebra on Λ. Suppose $T : \Omega \to \mathcal{G}$ is a measurable function. Then, $\mathbf{Q} : \mathcal{G} \to [0, 1]$ defined by $\mathbf{Q}(A) = \mathbf{P}(T^{-1}(A))$ is a probability measure on $(\Lambda, \mathcal{G})$.

Proof. If $A_n \in \mathcal{G}$ are pairwise disjoint, then so are $B_n := T^{-1}(A)$ which are in $\mathcal{F}$. Further, $T^{-1}(\cup_n A_n) = \cup_n B_n$, hence

$$\mathbf{Q}(\cup_n A_n) = \mathbf{P}(T^{-1}(\cup_n A_n)) = \sum_n \mathbf{P}(B_n) = \sum_n \mathbf{Q}(A_n).$$

Of course $T^{-1}(\Lambda) = \Omega$, hence $\mathbf{Q}(\Lambda) = \mathbf{P}(\Omega) = 1$. ■

We say that $\mathbf{Q}$ is the push-forward of $\mathbf{P}$ under T, and sometimes denote it as $\mathbf{Q} = \mathbf{P} \circ T^{-1}$.

Example 4: Tossing a coin infinitely many times

Here $\Omega = \{0, 1\}^{\mathbb{N}}$. In $\mathcal{S}$, we include all sets that are defined by finitely many co-ordinates. These sets of the form

$$A = \{\omega = (\omega_1, \omega_2, \ldots) \in \Omega : \omega_{i_1} = \varepsilon_1, \ldots, \omega_{i_n} = \varepsilon_n\} \tag{1}$$

for some $n \geq 1$ and some $1 \leq i_1 < \ldots < i_n$ and some $\varepsilon_1, \ldots, \varepsilon_n \in \{0, 1\}$, are called *finite dimensional cylinder sets* and the corresponding sigma-algebra $\mathcal{C} = \sigma(\mathcal{S})$ is called the cylinder sigma-algebra.

Define $T : [0, 1] \to \{0, 1\}^{\mathbb{N}}$ by $T(x) = (x_1, x_2, \ldots)$ where $x = \sum_{n \geq 1} x_n 2^{-n}$ is the binary expansion of x. To avoid ambiguity, for dyadic rational $x = k/2^n$, we take the expansion that has infinitely many ones. We claim that T is measurable. Indeed,

$$T^{-1}(\{\omega = (\omega_1, \omega_2, \ldots) \in \Omega : \omega_1 = \varepsilon_1, \ldots, \omega_n = \varepsilon_N)$$

is an interval of length 2^{-N}, and for any $A \in \mathcal{S}$, we can write $T^{-1}(A)$ as a union of such intervals. For example, if A is as in (1), then by taking $N = i_n$ and both possibilities for ω_i for $i \in [n] \setminus \{i_1, \ldots, i_n\}$, we see that $T^{-1}(A)$ is a union of 2^{N-n} pairwise disjoin intervals each of length 2^{-N}.

As T is measurable, we can define $\mathbf{P} = \lambda \circ T^{-1}$ as a probability measure on $\mathcal{C}$. Is this the probability measure we want? If we take an element of $\mathcal{S}$, say A as in (1), from the earlier discussion

$$\mathbf{P}(A) = \lambda(T^{-1}(A)) = 2^{N-n} \times \frac{1}{2^N} = \frac{1}{2^n},$$

which is the probability we wanted to assign to A.

In fact, as it happens, every probabilty space of interest to probabilists can be got this way by pushing forward Lebesgue measure on $[0, 1]$ by a measurable mapping.

Theorem 2: Borel isomorphism theorem

Let (X, d) be a complete and separable metric space and let μ be a probability measure on $\mathcal{B}_X$. Then there is a measurable $T : [0, 1] \to X$ such that $\lambda \circ T^{-1} = \mu$.

We shall not prove this theorem, but what we primarily need is a very important case of interest, when $X = \mathbb{R}^{\mathbb{N}}$ and μ is an infinite product of measures on $\mathbb{R}$. This is intimately connected to one of the most important notions in probability, namely *independence*. Instead of repeating, we refer the reader to sections 28–30 (also 27 if not familiar with finite product measures and 31–32 to go a little beyond the bare minimum needed) of Part-1 of these lecture notes. In section 24 there is a brief introduction to conditional probability.

Remark 1: History

Immediately after the initial works of Borel and Lebesgue on measure and integral, it was realized that measure theory could provide the foundation for probability theory. But it was only after the notions of independence and conditional probability could be satisfactorily captured under this framework that this became universally accepted. Many people made contributions to the former, but it was Kolmogorov's brilliant capturing of conditional probability under measure theoretic framework that is usually marked as the foundation of axiomatic definition of probability.

3. Some basic tools in probability

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

3.1. **First moment method.** 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 3: First moment method or Markov's inequality

Let $X \geq 0$ be a r.v. For any $t > 0$, we have $\mathbf{P}(X \geq t) \leq t^{-1}\mathbf{E}[X]$.

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

Thus, a positive random variable is unlikely to be more than a few multiples of its mean, e.g. there is only 10% chance of it being more than 10 times the mean. Trivial though it seems, Markov's inequality is very useful, particularly as it can be applied to various functions of the random variable of interest. Here are two instances.

(1) If X has finite variance,

$$\mathbf{P}(|X - \mathbf{E}[X]| \geq t) = \mathbf{P}(|X - \mathbf{E}[X]|^2 \geq t^2) \leq t^{-2}\text{Var}(X),$$

which is called *Chebyshev's inequality*. Higher the moments that exist, better the asymptotic tail bounds that we get, for example, $\mathbf{P}(|X - \mathbf{E}[X]| \geq t) \leq t^{-2p}\mathbf{E}[|X - \mathbf{E}[X]|^{2p}]$.

(2) 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 $e^{\lambda X}$.

3.2. **Second moment method.** The first moment method says that a positive random variable is likely to be less than a few multiples of the mean. Can we say the converse, i.e., a random variable is likely to be larger than a fraction of its mean? If the expectation is large, is the random variable

likely to be large? This is not true, for example, if $Y_n \sim (1-\frac{1}{n})\delta_0 + \frac{1}{n}\delta_{n^2}$, then $\mathbf{E}[Y_n] \to \infty$ but $\mathbf{P}\{Y_n > 0\} \to 0$.

What more information about a random variable will allow us to get the desired conclusion? Here is a natural approach using Chebyshev's inequality: If X is a non-negative random variable

$$\mathbf{P}\left(X \geq \frac{1}{2}\mathbf{E}[X]\right) \geq 1 - \mathbf{P}\left(|X - \mathbf{E}[X]| \geq \frac{1}{2}\mathbf{E}[X]\right) \geq 1 - 4\frac{\text{Var}(X)}{\mathbf{E}[X]^2}.$$

Thus, if the variance is bounded by $\frac{1}{5}\mathbf{E}[X]^2$, we get a non-trivial lower bound for the probability. More generally, if $\text{Var}(X) < (1-\delta)^2\mathbf{E}[X]^2$, then we get a lower bound for the probability that $X \geq \delta\mathbf{E}[X]$. Observe that in the example given above, $\text{Var}(Y_n) \asymp n^3$ is way larger than $\mathbf{E}[Y_n]^2 \asymp n^2$, hence the method does not work.

Thus, a control on the variance in terms of the square of the mean, allows us to say that a positive random variable is at least a fraction of its mean (with considerable probability). The following inequality is a variant of the same idea. It is better, as it gives a lower bound even if we only know that $\text{Var}(X) \leq 100\mathbf{E}[X]^2$.

Lemma 4: Second moment method or Paley-Zygmund inequality

For any non-negative r.v. X, and any $0 \leq \alpha \leq 1$, we have

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

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

Proof. $\mathbf{E}[X]^2 = \mathbf{E}[X\mathbf{1}_{X>0}]^2 \leq \mathbf{E}[X^2]\mathbf{E}[\mathbf{1}_{X>0}] = \mathbf{E}[X^2]\mathbf{P}(X > 0)$. 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,

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

The remaining conclusions follow easily. ■

3.3. **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 5: Borel Cantelli lemmas

Let A_n be events on a common probability space.

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

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

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 $\mathbf{P}(A_n \text{ i.o.}) = 1$. ■

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!

3.4. **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 6

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 even occurred or not. For example, A could be the event that infinitely many X_k are positive.

Theorem 6: 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 1

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

Remark 2: Reformulation on product space

We may reformulate Kolmogorov's zero-one law as follows. Let $(\Omega_k, \mathcal{F}_k)$ be measure spaces and consider $\Omega = \Omega_1 \times \Omega_2 \times \ldots$ endowed with the product sigma-algebra $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \ldots$. Let $\Pi_k : \Omega \mapsto \Omega_k$ be projection maps. Let $\mathcal{G}_k = \sigma\{\Pi_k, \Pi_{k+1}, \ldots\}$ and let $\mathcal{T} = \cap_k \mathcal{G}_k$. Kolmogorov's zero-one law is the statement that under any product probability measure on Ω, the sigma-algebra $\mathcal{T}$ is trivial (check the equivalence of this statement with the earlier one).

3.5. **Ergodicity of i.i.d. sequence.** We now prove another zero-one law now, which covers more events, but for i.i.d. sequences only. We formulate it in the language of product spaces first. Let $(\Omega, \mathcal{F})$ be a measure space and consider the product space $\Omega^{\mathbb{N}}$ with the product sigma algebra $\mathcal{F}^{\otimes\mathbb{N}}$. Let Pi_k be the projection onto the kth co-ordinate. For $k \in \mathbb{N}$, let $\theta_k : \Omega^{\mathbb{N}} \mapsto \Omega^{\mathbb{N}}$ denote the shift map defined by $\Pi_n \circ \theta_k = \Pi_{n+k}$ for all $n \geq 1$. In other words, $(\theta_k\omega)(n) = \omega(n + k)$ where $\omega = (\omega(1), \omega(2), \ldots)$.

Definition 7: Invariant sigma-algebra

An event $A \in \mathcal{F}^{\otimes\mathbb{N}}$ is said to be invariant if $\omega \in A$ if and only $\theta_k\omega \in A$ for any $k \geq 1$. The collection of all invariant events forms a sigma algebra that is called the invariant sigma

algebra and denoted $\mathcal{I}$. An invariant random variable is one that is measurable with respect to $\mathcal{I}$.

Note that a random variable X is invariant if and only if $X \circ \theta_k = X$ for all $k \geq 1$. We could also have taken this as the definition of an invariant random variable and then defined A to be an invariant event if $\mathbf{1}_A$ is an invariant random variable.

Example 5

Let A be the set of all ω such that $\lim_{n\to\infty} \omega_n = 0$ and let B be the set of all ω such that $|\omega_k| \leq 1$ for all $k \geq 1$. Then A is an invariant event as well as a tail event while B is an invariant event but not a tail event.

Exercise 2

In the setting above, show that $\mathcal{T} \subseteq \mathcal{I}$.

Lemma 7: Ergodicity of i.i.d. measures

Let $\mathbf{P}$ be a probability measure on $(\Omega, \mathcal{F})$. Then the invariant sigma algebra $\mathcal{I}$ on $\Omega^{\mathbb{N}}$ is trivial under $\mathbf{P}^{\otimes\mathbb{N}}$.

Proof. Let $\mu = \mathbf{P}^{\otimes\mathbb{N}}$. Suppose $A \in \mathcal{I}$. Since $\mathcal{A} := \bigcup_n \sigma\{\Pi_1, \ldots, \Pi_n\}$ is an algebra that generates the sigma algebra $\mathcal{F}^{\otimes\mathbb{N}}$, for any $\varepsilon > 0$, there is some $B \in \mathcal{A}$ such that $\mu(A\Delta B) < \varepsilon$. Let N be large enough that $B \in \sigma\{\Pi_1, \ldots, \Pi_N\}$. Then $\theta_N B \in \sigma\{\Pi_{N+1}, \ldots, \Pi_{2N}\}$. Under the product measure, Π_ks are independent, hence $\mu(B \cap \theta_N(B)) = \mu(B)\mu(\theta_N(B))$. But $\mu = \mu(B) = \mu(\theta_N(B))$ (because the measure is an i.i.d. product measure and hence invariant under the shift θ_N). Thus, $\mu(B \cap \theta_N B) = \mu(B)^2$. Now, $\mu(B\Delta A) < \varepsilon$ and hence

$$|\mu(B \cap \theta_N(B)) - \mu(A \cap \theta_N(A))| \leq \mu(B\Delta A) + \mu((\theta_N B)\Delta(\theta_N A)) \leq 2\varepsilon,$$
$$|\mu(B)^2 - \mu(A)^2| \leq |\mu(B) - \mu(A)||\mu(B) + \mu(A)| \leq 2\varepsilon.$$

This shows that $\mu(A \cap \theta_N A)$ and $\mu(A)^2$ are within 4ε of each other. But $A \in \mathcal{I}$, meaning that $\theta_N A = A$. Therefore, $\mu(A)$ is within 4ε of $\mu(A)^2$. As ε is arbitrary, $\mu(A) = \mu(A)^2$. This forces that $\mu(A) = 0$ of $\mu(A) = 1$. ■

3.6. **Bernstein/Hoeffding inequality.** Chebyshev's inequality tells us that the probability for a random variable to differ from its mean by k multiples of its standard deviation is at most $1/k^2$. Its power comes from its generality, but the bound is rather weak. If we know more about the random variable under consideration, we can improve upon the bound considerably. Here is one such inequality that is very useful. Sergei Bernstein was the first to exploit the full power of the

Chebyshev inequality (by applying it to powers or exponential of a random variable), but the precise lemma given here is due to Hoeffding.

Lemma 8: Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent random variables having zero mean. Assume that $|X_k| \leq a_k$ $a.s.$ for some positive numbers a_k. Then, writing $S = X_1 + \ldots + X_n$ and $A = \sqrt{a_1^2 + \ldots + a_k^2}$, we have $\mathbf{P}\{S \geq tA\} \leq e^{-\frac{1}{2}t^2}$ for any $t > 0$.

Before going to the proof, let us observe the following simple extensions.

(1) Applying the same to $-X_k$s, we can get the two-sided bound $\mathbf{P}\{|S| \geq tA\} \leq 2e^{-t^2/2}$.

(2) If $|X_k| \leq a_k$ are independent but do not necessarily have mean zero, then we can apply Hoeffding's inequality to $Y_k = X_k - \mathbf{E}[X_k]$. Since $|X_k| \leq a_k$, we also have $|\mathbf{E}[X_k]| \leq a_k$ and hence $|Y_k| \leq 2a_k$. This gives a conclusion that is slightly weaker but qualitatively no different: With $S = X_1 + \ldots + X_n$,

$$\mathbf{P}\left\{S - \mathbf{E}[S] \geq t\sqrt{a_1^2 + \ldots + a_n^2}\right\} \leq e^{-\frac{1}{8}t^2}.$$

Proof. Fix $\theta > 0$ and observe that

$$\mathbf{P}\{S \geq tA\} = \mathbf{P}\{e^{\theta S} \geq e^{\theta tA}\} \leq e^{-\theta tA}\mathbf{E}[e^{\theta S}] = e^{-\theta tA}\mathbf{E}\left[\prod_{k=1}^{n} e^{\theta X_k}\right]. \tag{2}$$

The inequality in the middle is Markov's, applied to $e^{\theta S}$. Since $x \mapsto e^{\theta x}$ is convex, on the interval $[-a_k, a_k]$, it lies below the line $x \mapsto \frac{a_k - x}{2a_k}e^{-\theta a_k} + \frac{x + a_k}{2a_k}e^{\theta a_k}$. Since $-a_k < X_k < a_k$, we get that $e^{\theta X_k} \leq \alpha_k + \beta_k X_k$, where $\alpha_k = \frac{1}{2}(e^{\theta a_k} + e^{-\theta a_k})$ and $\beta_k = \frac{1}{2a_k}(e^{\theta a_k} - e^{-\theta a_k})$. Plug this into (2) to get

$$\mathbf{P}\{S \geq tA\} \leq e^{-\theta tA}\mathbf{E}\left[\prod_{k=1}^{n}(\alpha_k + \beta_k X_k)\right] = e^{-\theta tA}\prod_{k=1}^{n}\alpha_k$$

since all terms in the expansion of the product that involve at least one X_ks vanishes upon taking expectation (as they are independent and have zero mean). We now wish to optimize this bound over θ, but that is too complicated (note that α_ks depend on θ). We simplify the bound by observing that $\alpha_k \leq e^{\theta^2 a_k^2/2}$. This follows from the following observation:

$$\begin{aligned}
\frac{1}{2}(e^y + e^{-y}) &= \sum_{n=0}^{\infty}\frac{y^{2n}}{(2n)!} \quad \text{(the odd powers cancel)} \\
&\leq \sum_{n=0}^{\infty}\frac{y^{2n}}{2^n\, n!} \quad \text{(as } (2n)! \geq 2n \times (2n-2) \times \ldots \times 2 = 2^n\, n!\text{)} \\
&= e^{y^2/2}.
\end{aligned}$$

Consequently, we get that $\prod_{k=1}^{n} \alpha_k \le e^{\theta^2 A^2/2}$. Thus, $\mathbf{P}\{S \ge tA\} \le e^{-\theta t A + \frac{1}{2}\theta^2 A^2}$. Now it is easy to see that the bound is minimized when $\theta = t/A$ and that gives the bound $e^{-t^2/2}$. ■

Clearly the Hoeffding bound is much better than the bound $1/t^2$ got by a direct application of Chebyshev's inequality. It is also a pleasing fact that $e^{-t^2/2}$ is a bound for the tail of the standard Normal distribution. In many situations, we shall see later that a sum of independent random variables behaves like a Gaussian, but that is a statement of convergence in distribution which does not say anything about the tail behaviour at finite n. Hoeffding's inequality is a non-asymptotic statement showing that S behaves in some ways like a Gaussian.

3.7. **Lovász's local lemma.** One of the recurring difficulties in probability is to get lower bounds of probabilities of events. In many cases, one can find many events whose simultaneous occurrence would imply the occurrence of the event of interest. One may also able to get a bound on the individual probabilities, but how to get a lower bound for the probability of their intersection? Two very simple bounds are

(1) $\mathbf{P}(A_1 \cap \ldots \cap A_n) = 1 - \mathbf{P}(A_1^c \cup \ldots \cup A_n^c) \ge 1 - \sum_{k=1}^{n}(1 - \mathbf{P}(A_k))$, by the union bound.

(2) $\mathbf{P}(A_1 \cap \ldots \cap A_n) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_n)$ if A_is are independent.

The first one is often too weak (entirely useless if $\sum_{k=1}^{n} \mathbf{P}(A_k) < n-1$) and the second is often inapplicable because the assumption of independence is too strong. The following lemma is one of several such statements that relaxes the independence assumption but still gives a positive lower bound.

Lemma 9: Lovász's local lemma

Let $A_1, \ldots, A_n$ be events in a common probability space. Assume that each A_k is independent of all except at most d of the other A_is. Further assume that $\mathbf{P}(A_k) \ge 1 - p$ for all k. If $4dp < 1$, then $\mathbf{P}(A_1 \cap \ldots \cap A_n) \ge (1-2p)^n$. In particular, the intersection has strictly positive probability.

Proof. We write

$$\mathbf{P}(A_1 \cap \ldots \cap A_n) = \prod_{k=1}^{n-1} \frac{\mathbf{P}(A_1 \cap \ldots \cap A_{k+1})}{\mathbf{P}(A_1 \cap \ldots A_k \cap A_{k+1}^c)}.$$

Fix k and consider the kth term in the product. Let $S \subseteq \{1, \ldots, k\}$ be the set of indices i for which A_i is not independent of A_{k+1}. Then $|S| \le d$ and ■

3.8. **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 it is not clear how to get a bound for the maximum. Kolmogorov found an amazing inequality - there seems to be no reason to expect a priori that such an inequality must hold!

Lemma 10: Kolmogorov's maximal inequality

Let X_n be independent random variables with finite variance and $\mathbf{E}[X_n] = 0$ for all n. Then, $\mathbf{P}\{\max_{k\le n} |S_k| > t\} \le t^{-2} \sum_{k=1}^n \text{Var}(X_k)$.

Observe that the right hand side is the bound that Chebyshev's inequality gives for the probability that $|S_n| \ge t$. Here the same quantity is giving an upper bound for the (presumably) much larger probability that one of $|S_1|, \ldots, |S_n|$ is greater than or equal to t.

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,

$$\mathbf{P}(\max_{k\le n} |S_k| > t) = \mathbf{P}(|S_\tau| > t) \le t^{-2}\mathbf{E}[S_\tau^2]. \tag{3}$$

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

$$\begin{aligned}
\mathbf{E}[S_n^2] &= \mathbf{E}\left[(S_\tau + (S_n - S_\tau))^2\right] \\
&= \mathbf{E}[S_\tau^2] + \mathbf{E}\left[(S_n - S_\tau)^2\right] + 2\mathbf{E}[S_\tau(S_n - S_\tau)] \\
&\ge \mathbf{E}[S_\tau^2] + 2\mathbf{E}[S_\tau(S_n - S_\tau)]. \qquad (4)
\end{aligned}$$

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

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

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 (4), this implies that $\mathbf{E}[S_n^2] \ge \mathbf{E}[S_\tau^2]$. Plug this into (3) to get $\mathbf{P}(\max_{k\le n} S_k > t) \le t^{-2}\mathbf{E}[S_n^2]$. ■

Remark 3

In proving this theorem, Kolmogorov implicitly introduced *stopping times* and *martingale property* (undefined terms for now). When martingales were defined later by Doob, the same proof could be carried over to what is called Doob's maximal inequality. In simple language, it just means that Kolmogorov's maximal inequality remains valid if instead of independence of X_ks, we only assume that $\mathbf{E}[X_k \mid X_1, \ldots, X_{k-1}] = 0$.

3.9. **Coupling of random variables.** *Coupling* is the name probabilists give to constructions of random variables on a common probability space with given marginals and joint distribution according to the need at hand. We illustrate it with a few examples.

Getting bounds on the distance between two measures: Suppose μ and ν are two probability measures on $\mathbb{R}$ and we wish to get an upper bound on their Lévy-Prohorov distance. One way is to use the definition and work with the measures. Here is another: Suppose we are able to construct two random variables X, Y on some probability space such that $X \sim \mu$, $Y \sim \nu$ and $|X - Y| \leq r$ $a.s$. Then we can claim that $d(\mu, \nu) \leq r$. Indeed,

$$F_\nu(t) = \mathbf{P}\{Y \leq t\} \geq \mathbf{P}\{X \leq t - r\} = F_\mu(t - r)$$

and similarly $F_\mu(t) \geq F_\nu(t - r)$.

Similar ideas can be used for other distances. For example, on a finite set $[n] = \{1, 2, \ldots, n\}$, let μ, ν be two probability measures. Their total variation distance is defined as $d_{TV}(\mu, \nu) = \max_{A \subseteq [n]} |\mu(A) - \nu(A)|$. One way to get a bound on the total variation distance is to construct two random variables X, Y on some probability space such that $X \sim \mu$, $Y \sim \nu$ and $\mathbf{P}\{X \neq Y\} = r$. Then $d_{TV}(\mu, \nu) \leq r$. Indeed, for any A, we have

$$\mu(A) = \mathbf{P}\{X \in A\} \leq \mathbf{P}\{Y \in A\} + \mathbf{P}\{Y \notin A, X \in A\} \leq \nu(A) + \mathbf{P}\{X \neq Y\}.$$

Getting the inequality with μ and ν reversed, we see that $d_{TV}(\mu, \nu) \leq \mathbf{P}\{X \neq Y\}$. It is not hard (in fact a nice exercise) to show that there is a coupling (X, Y) that achieves equality, i.e., $\mathbf{P}\{X \neq Y\} = d_{TV}(\mu, \nu)$.

Proving inequalities between numbers by coupling: Sometimes to show that $a \leq b$, it turns out to be convenient to construct random variables X, Y such that $X \leq Y$ $a.s.$ and $\mathbf{E}[X] = a$ and $\mathbf{E}[Y] = b$. That of course implies $a \leq b$ but the interesting point is that it can often be done by this method but not directly. Coupling method has been effectively used to show that a set is non-empty by showing that is has positive probability under some measure! This is called the *probabilistic method*.

Illustration of coupling: Let $X \sim \text{Bin}(100, 3/4)$ and $Y \sim \text{Bin}(100, 1/2)$. Then it must be true that $\mathbf{P}\{X \geq 71\} \geq \mathbf{P}\{Y \geq 71\}$, but can you show it by writing out the probabilities? It is possible, but here is a less painful way. Let $U_1, \ldots, U_{100}$ be i.i.d. Unif$[0, 1]$ random variables on some probability space. Let $X' = \sum_k \mathbf{1}_{U_k \leq 3/4}$ and $Y' = \sum_k \mathbf{1}_{V_k \leq 1/2}$. Then $X' \geq Y'$, hence the event $\{Y' \geq 71\}$ is a subset of $\{X' \geq 71\}$ showing that $\mathbf{P}\{X' \geq 71\} \geq \mathbf{P}\{Y' \geq 71\}$. But x' has the same distribution as X and Y' has the same distribution as Y, showing the inequality we wanted!

More generally, if $X \sim \mu$ and $Y \sim \nu$ and $X \geq Y$ $a.s.$, then $F_\mu(t) \leq F_\nu(t)$ for all $t \in \mathbb{R}$. If the latter relationship holds, we say that ν is stochastically dominated by μ.

Exercise 3

If ν is stochastically dominated by μ, show that there is a coupling of $X \sim \mu$ with $Y \sim \nu$ in such a way that $X \geq Y$ $a.s.$

Other instances of coupling: If you have studied Markov chains, then you would have perhaps seen a proof of convergence to stationarity by a coupling method due to Doeblin. In this method, two Markov chains are run, one starting from the stationary distribution and another starting at an arbitrary state. It is shown that the two Markov chains eventually meet. Once they meet, when they separate, it is impossible to tell which is which (by Markov property), hence the second chain "must have reached stationarity too".

4. 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.

4.1. **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,

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

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!

4.2. **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 11: Coupon collector problem

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

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

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 11. 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 (**??**) to get

$$ne^{-\frac{t}{n}-\frac{t}{n^2}} \leq \mathbf{E}[S_{t,n}] = n\left(1-\frac{1}{n}\right)^t \leq ne^{-\frac{t}{n}}. \tag{5}$$

and

$$\mathbf{E}[S_{t,n}^2] = n\left(1-\frac{1}{n}\right)^t + n(n-1)\left(1-\frac{2}{n}\right)^t \leq ne^{-\frac{t}{n}} + n(n-1)e^{-\frac{2t}{n}}. \tag{6}$$

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

$$\mathbf{P}(T_n > n\log n + n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \leq \mathbf{E}[S_{t,n}] \leq ne^{-\frac{n\log n + n\theta_n}{n}} \leq e^{-\theta_n} = o(1). \tag{7}$$

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 (6) we get $\mathbf{E}[S_{t,n}^2] \leq e^{\theta_n} + e^{2\theta_n}$. The first inequality in (5) gives $\mathbf{E}[S_{t,n}] \geq e^{\theta_n - \frac{\log n - \theta_n}{n}}$. Thus,

$$\mathbf{P}(T_n > n\log n - n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \geq \frac{\mathbf{E}[S_{t,n}]^2}{\mathbf{E}[S_{t,n}^2]} \geq \frac{e^{2\theta_n - 2\frac{\log n - \theta_n}{n}}}{e^{\theta_n} + e^{2\theta_n}} = 1 - o(1) \tag{8}$$

as $n \to \infty$. From (7) and (8), 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 11. 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 $\xi \sim \text{Geo}(p)$ (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,

$$\begin{aligned}
\mathbf{E}[T_n] &= \sum_{k=1}^{n} \frac{n}{n-k+1} = n \log n + O(n), \\
\text{Var}(T_n) &= n \sum_{k=1}^{n} \frac{k-1}{(n-k+1)^2} = n \sum_{j=1}^{n} \frac{n-j}{j^2} \\
&\leq Cn^2
\end{aligned}$$

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

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

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

4.3. **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 12: The fundamental theorem on Branching processes

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]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 $k \geq 2$, 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 $v_j \leq \eta((v_1, \ldots, v_{j-1}))$ for $2 \leq j \leq k$. Also define $Z_{k-1}(\eta) = \#\{(v_1, \ldots, v_k) \in T\}$ for $k \geq 2$ 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.

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

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

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

(1) By Markov's inequality, $\mathbf{P}(Z_n > 0) \leq \mathbf{E}[Z_n] = m^n \to 0$. 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

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

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 4: The "critical" case $m = 1$

Strictly speaking, the fundamental theorem of branching processes also asserts that extinction occurs almost surely when $m = 1$. However, to get it by these methods, one will have to refine the first moment method as follows. If X is a random variable taking values in $\mathbb{N}$,

then $\mathbf{P}\{X \geq 1\} \leq \mathbf{E}[X]/\mathbf{E}[X|X \geq 1]$, where the denominator on the right is a conditional expectation. Clearly the bound is at least as good as Markov's inequality, but it can be much better in some situations. For example, in the branching process with $m = 1$, one can show that $\mathbf{E}[Z_n|Z_n \geq 1] \to \infty$ as $n \to \infty$ (intuitively, if the branching process has to survive as long as n generations, it has to do it by spawning many offsprings). Since $\mathbf{E}[Z_n] = 1$, this shows that $\mathbf{P}\{Z_n \geq 1\} \to 0$, proving almost sure extinction.

4.4. **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 $n \geq 1$ 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 13: Hardy-Ramanujan

With the above setting, for any $\delta > 0$, as $n \to \infty$ we have

$$\mu_n \left\{ k \in [n] : \left| \frac{\nu(k)}{\log\log n} - 1 \right| > \delta \right\} \to 0. \tag{9}$$

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, $\nu(k) = \sum_{p \leq k} X_p(k)$. We define $\psi(k) := \sum_{p \leq \sqrt[4]{k}} X_p(k)$. Then, $\psi(k) \leq \nu(k) \leq \psi(k) + 4$ since there can be at most four primes larger than $\sqrt[4]{k}$ that divide k. From this, it is clearly enough to show (9) 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 $\mathbf{E}_n[X_p X_q] = \frac{\lfloor \frac{n}{pq} \rfloor}{n}$. Observe that $\frac{1}{p} - \frac{1}{n} \leq \frac{\lfloor \frac{n}{p} \rfloor}{n} \leq \frac{1}{p}$ and $\frac{1}{pq} - \frac{1}{n} \leq \frac{\lfloor \frac{n}{pq} \rfloor}{n} \leq \frac{1}{pq}$.

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

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

4.5. **A random graph question.** The complete graph K_n has vertex set $[n] = \{1, 2, \ldots, n\}$ and edge set $E = \{\{i, j\} : 1 \le i < j \le n\}$. We now define a random graph model as a random sub-graph of K_n. This model has been studied extensively by probabilists in the last fifty years.

Definition 8: Erdös-Rényi random graph

Fix $0 < p < 1$. Let $X_{i,j}$, $1 \le i < j \le n$, be i.i.d. Ber(p) random variables. Let G be the graph with vertex set $[n]$ and edge-set $\{\{i, j\} : X_{i,j} = 1\}$. Then G is called the *Erdös-Rényi random graph with parameters n and p* and denoted $\mathcal{G}(n, p)$.

There are many interesting questions about $\mathcal{G}(n, p)$. Here we ask only one: *Is $\mathcal{G}(n, p)$ connected?* If $p = 1$, the answer is clearly yes, and if $p = 0$, the answer is clearly no. It is not hard to see that (use coupling!) to show that the probability that $\mathcal{G}(n, p)$ is connected increases with p. Where

does the change from disconnected to connected take place? The answer is given in the following theorem.

Theorem 14: Connectivity threshold for Erdös-Renyi random graph

Fix $\delta > 0$ and let $p_n^{\pm} = (1 \pm \delta)\frac{\log n}{n}$. Then, as $n \to \infty$,

$$\mathbf{P}\{\mathcal{G}(n, p_n^+) \text{ is connected } \} \to 1 \quad \text{and} \quad \mathbf{P}\{\mathcal{G}(n, p_n^-) \text{ is connected } \} \to 0.$$

Unlike in the other problems, here the second moment method is easier, because we show disconnection by showing that there is at least one isolated vertex (i.e., a vertex that is not connected to any other vertex). To show connectedness, we must go over all proper subsets of vertices.

Proof that $\mathcal{G}(n, p_n^-)$ is unlikely to be connected. Let Y be the number of isolated vertices, i.e., $Y = \sum_{i=1}^n Y_i$, where Y_i is the indicator of the event that vertex i is not connected to any other vertex. Then,

$$\mathbf{E}[Y] = \sum_{i=1}^n \mathbf{E}[Y_i] = n(1-p)^{n-1} \geq ne^{-np-np^2}$$

if $p < \frac{1}{2}$ (so that $1 - p \geq e^{-p-p^2}$). Further, $Y_i Y_j = 1$ if and only if all the $2n-3$ edges coming out of i or j (including the one connecting i and j) are absent (i.e., $X_{i,k}, X_{j,k}$ are all 0). Therefore,

$$\begin{aligned}
\mathbf{E}[Y^2] &= \sum_{i=1}^n \mathbf{E}[Y_i] + 2\sum_{i<j} \mathbf{E}[Y_i]\mathbf{E}[Y_j] \\
&= n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3} \\
&\leq ne^{-p(n-1)} + n^2 e^{-(2n-3)p}.
\end{aligned}$$

When $p = p_n^-$, by the second moment method that

$$\mathbf{P}\{Y \geq 1\} \geq \frac{\mathbf{E}[Y]^2}{\mathbf{E}[Y^2]} \geq \frac{n^2 e^{2np-2np^2}}{ne^{-p(n-1)} + n^2 e^{-(2n-3)p}} = \frac{e^{-2np^2}}{\frac{1}{n}e^{p(n+1)} + e^{3p}}$$

which goes to 1 as $n \to \infty$ (as $p_n \to 0$ and $\frac{1}{n}e^{np_n} \to 0$). When $Y \geq 1$, $\mathcal{G}(n,p)$ is disconnected, completing the proof. ■

Proof that $\mathcal{G}(n, p_n^+)$ is unlikely to be disconnected. We get a crude estimate as follows. Suppose $A \subseteq [n]$. Then A is disconnected from A^c if and only if $X_{i,j} = 0$ for all $i \in A$ and all $j \in A^c$. This has probability $(1-p)^{|A|(n-|A|)}$. If the graph is disconnected, then there must be some such set A with $|A| \leq n/2$. Thus, by the union bound,

$$\mathbf{P}\{\mathcal{G}(n,p) \text{ is not connected}\} \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p)^{k(n-k)}.$$

Now, we set $p = p_n^+$ and divide the sum into $k \leq \varepsilon n$ and $k > \varepsilon n$.

In the second sum, we use the simple bounds $\binom{n}{k} \le 2^n$ and $k(n-k) \ge \varepsilon(1-\varepsilon)n^2$. Since $1-p \le e^{-p}$, and there are at most n terms, we get (recall the definition of p_n^+)

$$\sum_{k>\varepsilon n} \binom{n}{k}(1-p)^{k(n-k)} \le n2^n e^{-\varepsilon(1-\varepsilon)(1+\delta)n\log n}.$$

Obviously this goes to zero as $n \to \infty$ (for any choice of $\varepsilon > 0$, which will be made later).

The sum over $k \le \varepsilon$ is handled by setting $\binom{n}{k} \le n^k$ and $1-p \le e^{-p}$. We get

$$\begin{aligned}\sum_{1\le k\le \varepsilon n} \binom{n}{k}(1-p)^{k(n-k)} &\le \sum_{k\le\varepsilon n} e^{-k[(n-k)p-\log n]} \\ &\le \sum_{1\le k\le\varepsilon n} e^{-k\log n[(1+\delta)(1-\frac{k}{n})-1]} \\ &\le \sum_{k=1}^{\infty} e^{-k\log n[(1+\delta)(1-\varepsilon)-1]}.\end{aligned}$$

If $\varepsilon > 0$ is chosen small enough that $(1+\delta)(1-\varepsilon) - 1 \ge \frac{1}{2}\delta$, then the above sum becomes a geometric series whose sum is

$$\frac{e^{-\frac{1}{2}\delta\log n}}{1-e^{-\frac{1}{2}\delta\log n}} \le \frac{1}{2}n^{-\delta/2},$$

the inequality holding for large n. Thus, $\mathbf{P}\{\mathcal{G}(n,p_n^+) \text{ is connected }\} \to 1$. ■

4.6. **A probabilistic version of Fermat's last theorem.** Fermat's last theorem is the statement that there are no strictly positive integers a, b, c such that $a^p + b^p = c^p$, if $p \ge 3$ is an integer. For $p = 2$ there are solutions of course, e.g., $3, 4, 5$. What is the intuition behind why it fails for larger p? There are more squares than cubes than fourth powers and so on (in the sense that the number of p-th powers below N grows like $N^{1/p}$). In a sparser sequence, there should be less coincidences of the kind where sum of two terms is another term. Here is a way to make a random version of the question that shows that $p = 3$ is precisely where there is a change of behaviour!

Fix $\alpha > 0$ and let $\xi_n \sim \text{Ber}(n^{-\alpha})$ be independent. This gives us a random subset of positive integers $\mathcal{S}_\alpha = \{n : \xi_n = 1\}$. By considering the summability of $\mathbf{P}\{\xi_n = 1\}$, from the Borel-Cantelli lemmas we see that $\mathcal{S}_\alpha$ is a finite set w.p.1. if and only if $\alpha > 1$. Hence let us fix $\alpha \le 1$ and observe that $|\mathcal{S}_\alpha \cap [N]| = \xi_1 + \ldots + \xi_N$. Therefore,

$$\mathbf{E}[|\mathcal{S}_\alpha \cap [N]|] = \sum_{k=1}^{N} \frac{1}{k^\alpha} \sim \begin{cases} \frac{1}{1-\alpha}N^{1-\alpha} & \text{if } \alpha < 1, \\ \log N & \text{if } \alpha = 1. \end{cases}$$

Alternately, for $\alpha < 1$ the Nth term is of the order of N^p where $p = \frac{1}{1-\alpha}$. Thus $p > 3$ corresponds to $\alpha < \frac{2}{3}$.

Theorem 15: Erdös–Ulam

If $\alpha < \frac{2}{3}$, then with probability 1, there are at most finitely many triples $(a,b,c) \in \mathcal{S}_\alpha^3$ such that $a < b < c$ and $a + b = c$. If $\alpha \geq \frac{2}{3}$, then with probability 1, there are infinitely many such triples.

Just to avoid some computations, we have not allowed $a = b$ in our solution space. It does not make a difference to the result if allowed. The proof will proceed by computing the first and second moment of the random variable T_N denoting the number of solution triples with $c \leq N$.

Proof. Fix any $1 \leq a < b < c = (a+b)$. The probability that (a,b,c) is in $\mathcal{S}_\alpha^3$ is $1/(ab(a+b))^\alpha$. As $a + b \geq \sqrt{ab}$,

$$\begin{aligned}\mathbf{E}[T_N] &\leq \sum_{1 \leq a < b < N} \frac{1}{(ab)^{\frac{3\alpha}{2}}} \qquad \text{(because } a+b \geq \sqrt{ab}\text{)} \\ &\leq \left(\sum_{k=1}^{\infty} \frac{1}{k^{\frac{3\alpha}{2}}}\right)^2\end{aligned}$$

This sum finite if $\alpha > \frac{2}{3}$. Since the total number of solutions T is the increasing limit of T_N, MCT shows that $\mathbf{E}[T] < \infty$ and hence $T < \infty$ a.s. This proves the first statement.

For the second statement, we work out the case $\alpha = \frac{2}{3}$ and leave $\alpha < \frac{2}{3}$ as an (easier) exercise.

$$\mathbf{E}[T_N] = \sum_{c=1}^{N} \frac{1}{c^{\frac{2}{3}}} \sum_{a < \frac{c}{2}} \frac{1}{(a(c-a))^{\frac{2}{3}}}.$$

The inner sum can be written as

$$\frac{1}{c^{\frac{1}{3}}} \times \frac{1}{c} \sum_{a < \frac{c}{2}} \frac{1}{(\frac{a}{c}(1-\frac{a}{c}))^{\frac{2}{3}}} \sim \frac{1}{c^{\frac{1}{3}}} \int_0^{1/2} \frac{dx}{x^{\frac{2}{3}}(1-x)^{\frac{2}{3}}}.$$

for c large. Denoting the integral as C (and a small argument needed to ignore small c), we get $\mathbf{E}[T_N] \sim C \sum_{c=1}^{N} \frac{1}{c} \sim C \log N$. This expectation goes to infinity and hence $\mathbf{E}[T] = \infty$. But to say that T is infinite a.s., we compute the second moment of T_N.

$$\mathbf{E}[T_N^2] = \sum_{c,c'=1}^{N} \sum_{a \leq c,\ a' \leq c'} \mathbf{E}[\xi_a \xi_{c-a} \xi_c \xi_{a'} \xi_{c'-a'} \xi_{c'}].$$

When the two triples are disjoint, the expectations factor and hence we can write

$$\begin{aligned}\mathbf{E}[T_N^2] &= \mathbf{E}[T_N]^2 + \sum_{*} \mathbf{E}[\xi_a \xi_{c-a} \xi_c \xi_{a'} \xi_{c'-a'} \xi_{c'}] - \mathbf{E}[\xi_a \xi_{c-a} \xi_c] \mathbf{E}[\xi_{a'} \xi_{c'-a'} \xi_{c'}] \\ &\leq \mathbf{E}[T_N]^2 + \sum_{*} \mathbf{E}[\xi_a \xi_{c-a} \xi_c \xi_{a'} \xi_{c'-a'} \xi_{c'}]\end{aligned}$$

where the asterisk indicates summing over pairs of triples such that $\{a, c-a, c\} \cap \{a', c'-a', c'\} \neq \emptyset$. We show that this entire sum is $O(\log N)$, which then shows that the standard deviation of T_N is

$O(\sqrt{\log N})$. As $\mathbf{E}[T_N] \sim C \log N$, by Chebyshev inequality we get

$$\mathbf{P}\{T_N \leq (1-\delta)C \log N\} \leq \frac{\mathrm{Var}(T_N)}{C^2\delta^2 \log^2 N} \to 0$$

as $N \to \infty$. This shows that $T = \infty$ a.s. and in fact gives a more quantitative statement about how many solutions there are.

It remains to show that the asterisked sum is $O(\log N)$. Now we must divide into several cases.

(1) $a = a'$, $c = c'$: The triples are the same and summing over all such cases we just get $\mathbf{E}[T_N] \sim C \log N$.

(2) $c = c'$: If $a = a'$, then the triples would be the same, and that has been taken care of in the first case. Therefore, $\xi_a, \xi_{c-a}, \xi_{a'}, \xi_{c-a'}\xi_c$ are all independent and the expectation of their product is $1/(caa'(c-a)c-a'))^{2/3}$. Summing over all such cases (we overestimate by also including the terms $a = a'$) we get

$$\sum_{c=1}^{N} \frac{1}{c^{\frac{2}{3}}} \sum_{a,a' < \frac{c}{2}} \frac{1}{(a(c-a))^{\frac{2}{3}}(a'(c-a'))^{\frac{2}{3}}}$$

The inner sum factors and can be written as

$$\left(\frac{1}{c^{4/3}} \sum_{a=1}^{c/2} \frac{1}{(\frac{a}{c}(1-\frac{a}{c}))^{\frac{2}{3}}} \right)^2 \sim \frac{C^2}{c^{2/3}}$$

for the same C as before. Hence the total contribution of these terms is $\sum_{c=1}^{N} c^{-4/3}$ which stays bounded as $N \to \infty$.

(3) $c < c'$: It is possible that $a = a'$ or $c - a = c' - a'$ but not both (otherwise we would have $c = c'$) in which case there are five distinct indices. But it is also possible to have $a' = c - a$, $c' - a' = c$ which fixed $a' = c' - c$ and $a = 2c - c'$ (if at all possible) in which case there are four distinct indices a, a', c, c'. Hence, we can bound the total sum by

$$\sum_{c < c' \leq N} \frac{1}{(cc')^{2/3}} \left(2 \sum_{a < c/2, a' < c'/2} \frac{1}{(aa'(c-a))^{2/3}} + \frac{1}{(c'-c)^{2/3}} \right).$$

The first inner sum can be approximated as

$$\sim \frac{cc'}{c^{4/3}c'^{2/3}} \left(\int_0^{1/2} \frac{dx}{x^{2/3}(1-x)^{2/3}} \right) \left(\int_0^{1/2} \frac{dx}{x^{2/3}} \right) \sim C_1 c^{-1/3} c'^{1/3}.$$

The second inner sum can be approximated as

$$\sim \frac{1}{c^{1/3}} \int_0^{1/2} \frac{dx}{x^{2/3}(1-x)^{2/3}} = Cc^{-1/3}.$$

Thus, the total sum is

$$\leq C_2 \sum_{c<c'\leq N} \frac{1}{(cc')^{2/3}}(c^{-1/3}c'^{1/3} + c^{-1/3})$$

$$= C_3 \sum_{c=1}^{N} \frac{1}{c} \sum_{c'>c} \frac{1}{c'^{1/3}}$$

by dividing into the two case, $c \neq c'$ and $c = c'$. ■

5. 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 $\limsup \frac{S_n}{n}$. Let us give a couple more.

5.1. **Random series.** Let X_n be independent random variables. The event that the series $\sum_n X_n$ converges is clearly a tail event, hence has probability zero or one. Is it zero or one? Depends on the variables.

Let $X_n \sim \text{Ber}(p_n)$. Then the series converges if and only if $X_n = 0$ for all but finitely many n. By the Borel-Cantelli lemma,

$$\mathbf{P}\{X_n = 1 \text{ i.o.}\} = \begin{cases} 0 & \text{if } \sum_n p_n < \infty, \\ 1 & \text{if } \sum_n p_n = \infty. \end{cases}$$

Thus, the series $\sum_n X_n$ converges almost surely if $\sum_n p_n < \infty$ and diverges almost surely if $\sum_n p_n = \infty$.

Since $p_n = \mathbf{E}[X_n]$, this may give the impression that what matters is the sum of expectations. Not entirely correct. For example, let X_n be independent with $\mathbf{P}\{X_n = 1\} = \mathbf{P}\{X_n = -1\} = p_n/2$ and $\mathbf{P}\{X_n = 0\} = 1 - p_n$. Then again, the random series converges if and only if $X_n \neq 0$ only finitely often. Again by Borel-Cantelli lemma, this is equivalent to the convergence of $\sum_n p_n$. Here $\mathbf{E}[X_n] = 0$ for all n, what p_n measures is the variance.

In general, Kolmogorov (after Khinchine and others) found a complete and satisfactory answer to the general question. His answer is that the random series converges almost surely if and only if three (non-random) series constructed from the distributions of X_ns converge. We shall prove Kolmogorov's three series theorem later.

5.2. **Random series of functions.** One can similarly ask about convergence of $\sum_n X_n u_n$, where X_n are independent random variables and u_n are elements of a Banach space. In particular, let $f_n : [0,1] \mapsto \mathbb{R}$ be given continuous functions and consider the series $\sum_n X_n f_n(t)$. The following events are clearly tail events.

- The event C that the series converges uniformly on $[0,1]$.

- The event ND that the sum is a nowhere differentiable function (it makes sense to ask this only if $\mathbf{P}(C) = 1$).

Again, whether these events have probability 0 or 1 depends on the variables X_ns and the functions f_ns. For example, if $f_n(t) = \sin(\pi nt)/n$ and X_n are i.i.d. $N(0,1)$, then Wiener showed that $\mathbf{P}(C) = 1$ and $\mathbf{P}(\text{ND}) = 1$.

We shall see this in the next part of the course on Brownian motion. For now, you may simply compare it with Weierstrass' nowhere differentiable function $\sum_n \sin(3^n\pi t)/3^n$. In contrast, the random series does not require such rapid increase of frequencies. However, although $\mathbf{P}(C \cap \text{ND}) = 1$, it is not easy to produce a *particular sequence* $x_n \in \mathbb{R}$ such that the function $\sum_n x_n \frac{\sin(\pi nt)}{n}$ converges uniformly but gives a nowhere differentiable function!

5.3. **Random power series.** Let X_n be i.i.d. Exp(1). As a special case of the previous examples, 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 number 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 $t = 1 + \varepsilon$ with $\varepsilon > 0$, this decays very fast and is summable. Hence, $|X_n|^{\frac{1}{n}} \leq 1 + \varepsilon$ $a.s.$. and hence $R \leq 1 + \varepsilon$ $a.s$. 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 $|X_n|^{\frac{1}{n}} \geq 1 - \varepsilon$. 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 $\sum_{n=0}^{\infty} c_n X_n z^n$.

5.4. **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!

5.5. **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).

6. Weak law of large numbers

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.

[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 $\bigcap_{F\subseteq I, |F|<\infty} \sigma(\bigcup_{i\in I\setminus F} \mathcal{F}_i)$

Here and in the rest of the course S_n will denote the partial sum $X_1 + \ldots + X_n$. If we have several sequences $(X_n), (Y_n)$ etc., we shall distinguish them by writing S_n^X, S_n^Y and so on.

Theorem 16: 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

$$\mathbf{P}\left(\left| \frac{1}{n} S_n - \mathbf{E}[X_1] \right| > \delta \right) \to 0.$$

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. **Step 1:** First assume that X_i have finite variance σ^2. Without loss of generality, let $\mathbf{E}[X_1] = 0$ (or else replace X_i by $X_i - \mathbf{E}[X_1]$). By Chebyshev's inequality, $\mathbf{P}(|n^{-1} S_n| > \delta) \leq n^{-2}\delta^{-2}\text{Var}(S_n)$. By the independence of X_is, we see that $\text{Var}(S_n) = n\sigma^2$. Thus, $\mathbf{P}(|\frac{S_n}{n}| > \delta) \leq \frac{\sigma^2}{n\delta^2}$ which goes to zero as $n \to \infty$, for any fixed $\delta > 0$.

Step 2: Now let X_i have finite expectation (which we assume is 0), but not necessarily any higher moments. 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,

$$\mathbf{P}\left\{ \left| \frac{S_n^Y}{n} - \mu_n \right| > \delta \right\} \leq \frac{\text{Var}(Y_1)}{n\delta^2} \leq \frac{A_n \mathbf{E}[|X_1|]}{n\delta^2}. \tag{10}$$

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.

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

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

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

Now, we take $A_n = \alpha n$ with $\alpha := \delta^3 \mathbf{E}[|X_1|]^{-1}$. The first term clearly becomes less than δ. 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

$$\limsup_{n\to\infty} \mathbf{P}\left\{ \left| \frac{S_n}{n} \right| > 2\delta \right\} \leq \delta$$

which gives the desired conclusion. ■

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 $a_n = \mathbf{E}[X\mathbf{1}_{|X|\leq n}]$).

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

7. Applications of weak law of large numbers

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

7.1. Bernstein's proof of Weierstrass' approximation theorem.

Theorem 17: Weierstrass' approximation theorem

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

$$\begin{aligned}
\left| \mathbf{E}_p \left[f\left(\frac{S_n}{n}\right)\right] - f(p) \right| &\le \mathbf{E}_p \left[\left| f\left(\frac{S_n}{n}\right) - f(p) \right| \right] \\
&= \mathbf{E}_p \left[\left| f\left(\frac{S_n}{n}\right) - f(p) \right| \mathbf{1}_{|\frac{S_n}{n}-p| \le \delta} \right] + \mathbf{E}_p \left[\left| f\left(\frac{S_n}{n}\right) - f(p) \right| \mathbf{1}_{|\frac{S_n}{n}-p| > \delta} \right] \\
&\le \omega_f(\delta) + 2\|f\| \mathbf{P}_p \left\{ \left| \frac{S_n}{n} - p \right| > \delta \right\} \qquad (11)
\end{aligned}$$

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 $\mathrm{Var}_p(X_1) = p(1-p)$ to write

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

Plugging this into (11) and recalling that $Q_{f,n}(p) = \mathbf{E}_p\left[f\left(\frac{S_n}{n}\right)\right]$, we get

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

Since f is uniformly continuous (which is the same as saying that $\omega_f(\delta) \downarrow 0$ as $\delta \downarrow 0$), given any $\varepsilon > 0$, we can take $\delta > 0$ small enough that $\omega_f(\delta) < \varepsilon$. 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 $\|Q_{f,n} - f\| < 2\varepsilon$. ■

Remark 5

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.

7.2. **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,

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

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 5

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.

7.3. **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...

8. Modes of convergence

Before going to the strong law of large numbers which gives a different sense in which S_n/n is close to the mean of X_1, 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 9

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 $\mathbf{P}\{\omega : \lim X_n(\omega) = X(\omega)\} = 1$.

- $X_n \xrightarrow{P} X$ (X_n converges to X in probability) if $\mathbf{P}\{|X_n - X| > \delta\} \to 0$ as $n \to \infty$ for any $\delta > 0$.
- $X_n \xrightarrow{L^p} X$ (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.
- $X_n \xrightarrow{d} X$ (X_n converges to X in distribution) if the distribution of $\mu_{X_n} \xrightarrow{d} \mu_X$ 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.

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

Example 6

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

(1) $\mathbf{1}_{A_n} \xrightarrow{P} 0 \iff \lim_{n\to\infty} \mathbf{P}(A_n) = 0$,

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

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 18

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

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

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

Proof. Note that analogous to the example, in general

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

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

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 $\limsup |X_n - X| \leq \delta \; a.s.$ 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 $(X_1, X_2, \ldots)$ 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 $\mathbf{P}(|X_n| \geq \delta)$ is not summable for any $\delta > 0$. Almost sure convergence implies convergence in probability, but no rate of convergence is assured.

Exercise 6

(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 $X_nY_n \overset{P}{\to} XY$.

8.2. **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 19

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 $\delta > 0$

$$\mathbf{P}(X_n \leq t) \leq \mathbf{P}(X \leq t + \delta) + \mathbf{P}(X - X_n > \delta)$$

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

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, $X_n \xrightarrow{d} X$.

(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-degnerate 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 $T(U) \stackrel{d}{=} X$. 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 7

(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 $U \stackrel{d}{=} X$, $V \stackrel{d}{=} Y$ 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 $X_nY_n \xrightarrow{d} XY$.

8.3. **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 7

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 **??** 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.

9. Uniform integrability

Definition 10: Uniform integrability

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

Example 8

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 $\mathbf{E}[|X_i|^p] \leq M$ for all $i \in I$ for some $M > 0$, then

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

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

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 **??** 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 $\mathbf{E}[|X_i|] = \mathbf{E}[|X_i|\mathbf{1}_{|X_i|<A}] + \mathbf{E}[|X_i|\mathbf{1}_{|X_i|\geq A}] \leq A + 1$. Convince yourself that for any $p > 1$, there exist uniformly integrable families that are not bounded in L^p.

Exercise 8

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 20

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 8, $\{X_n\}$ 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 $\mathbf{E}[|Y_{n_k}|\mathbf{1}_{|Y_{n_k}|\geq k}] > \delta$. 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 $\varepsilon > 0$ and find $A > 0$ so that $\mathbf{E}[|Y_k|\mathbf{1}_{|Y_k|>A}] \leq \varepsilon$ for all k. Then,

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

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 $\limsup \mathbf{E}[|Y_k|] \leq \varepsilon$ for any ε and it follows that $Y_k \xrightarrow{L^1} 0$. ■

Corollary 21

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 20 and corollary 21, 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 6: Relationship to compactness

Like tightness of measures, uniform integrability is also related to a compactness question. On a Banach space X, there is the norm topology coming from the norm, and the weak topology induced by the dual space X^* (it is the smallest topology on X in which every element of X^* is continuous). In particular when $X = L^p(\mu)$ for a probability measure μ, what are the compact sets in the weak topology?

For $1 < p < \infty$, we know that L^p and L^q are duals of each other, where $\frac{1}{p} + \frac{1}{q} = 1$. Therefore, the weak topology on L^p is the same as the weak* topology on L^p when viewed as the dual of L^q. By the Banach-Alaoglu theorem, norm-bounded sets are pre-compact in the weak topology. Norm-boundedness is also necessary (why?), hence this gives a precise characterization for pre-compact sets in L^p with weak topology. This argument fails for L^1, since it is not the dual of a Banach space. The *Dunford-Pettis theorem* asserts that pre-compact subsets of $L^1(\mu)$ in this weak topology are precisely uniformly integrable subsets of $L^1(\mu)$.

10. Strong law of large numbers

If X_n are i.i.d with finite mean, then the weak law asserts that $n^{-1}S_n \xrightarrow{P} \mathbf{E}[X_1]$. The strong law strengthens it to almost sure convergence.

Theorem 22: Kolmogorov's strong law of large numbers

Let X_n be i.i.d with $\mathbf{E}[|X_1|] < \infty$. Then, as $n \to \infty$, we have $\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}[X_1]$.

The proof of this theorem is somewhat complicated. First of all, we should ask if WLLN implies SLLN? From Lemma 18 we see that this can be done if $\mathbf{P}\left(|n^{-1}S_n - \mathbf{E}[X_1]| > \delta\right)$ is summable, for every $\delta > 0$. Even assuming finite variance $\mathrm{Var}(X_1) = \sigma^2$, Chebyshev's inequality only gives a bound of $\sigma^2\delta^{-2}n^{-1}$ for this probability and this is not summable. Since this is at the borderline of summability, if we assume that pth moment exists for some $p > 2$, we may expect to carry out this proof. Suppose we assume that $\alpha_4 := \mathbf{E}[X_1^4] < \infty$ (of course 4 is not the smallest number bigger than 2, but how do we compute $\mathbf{E}[|S_n|^p]$ in terms of moments of X_1 unless p is an even integer?). Then, we may compute that (assume $\mathbf{E}[X_1] = 0$ without loss of generality)

$$\mathbf{E}\left[S_n^4\right] = n^2(n-1)^2\sigma^4 + n\alpha_4 = O(n^2).$$

Thus $\mathbf{P}\left(|n^{-1}S_n| > \delta\right) \leq n^{-4}\delta^{-4}\mathbf{E}[S_n^4] = O(n^{-2})$ which is summable, and by Lemma 18 we get the statement of SLLN under fourth moment assumption. This can be further strengthened to prove SLLN under the second moment assumption, which we first present since there is one idea (of working with subsequences) that will also be used in the proof of SLLN under just the first moment assumption[4].

Theorem 23: SLLN under second moment assumption

Let X_n be i.i.d with $\mathbf{E}[|X_1|^2] < \infty$. Then, $\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}[X_1]$ as $n \to \infty$.

[4]The idea of proving SLLN this way was told to me by Sourav Sarkar who came up with the idea when he was a B.Stat student. I have not seen it any book, although it is likely that the observation has been made before.

Proof. Assume $\mathbf{E}[X_1] = 0$ without loss of generality and let $\sigma^2 = \text{Var}(X_1)$. By Chebyshev's inequality, $\mathbf{P}\{|\frac{1}{n}S_n| \geq t\} \leq \frac{\sigma^2}{nt^2}$ since $\text{Var}(S_n) = n\sigma^2$. Now consider the sequence $n_k = k^2$. The bounds $\frac{\sigma^2}{tn_k^2}$ are summable, hence by the first Borel-Cantelli lemma, we see that $|\frac{1}{n_k}S_{n_k}| \leq \delta$ for all but finitely many k, almost surely. If this even be denoted E_δ, then $\mathbf{P}(E_\delta) = 1$, hence $\cap_{\delta \in \mathbb{Q}_+} E_\delta$ also has probability one, which is another way of saying that $\frac{1}{n_k}S_{n_k} \overset{a.s.}{\to} 0$.

This can be applied to the i.i.d. sequence X_n^+ and the i.i.d. sequence X_n^- (that two sequences are not independent of each other is irrelevant) to see that

$$\frac{1}{n_k}U_{n_k} \to \mathbf{E}[X_1^+] \quad \text{and} \quad \frac{1}{n_k}V_{n_k} \to \mathbf{E}[X_1^-], \quad \text{a.s.} \tag{12}$$

where U_n, V_n are partial sums of X_i^+ and X_i^-, respectively.

Now for any n, let k be such that $n_k \leq n < n_{k+1}$. Clearly $U_{n_k} \leq U_n < U_{n_{k+1}}$ and $V_{n_k} \leq V_n < V_{n_{k+1}}$, since the summands are non-negative (a similar assertion is false for S_n, which is why we break into positive and negative parts). Thus,

$$\frac{1}{n_{k+1}}U_{n_k} \leq \frac{1}{n}U_n \leq \frac{1}{n_k}U_{n_{k+1}}$$

and the analogous statement for V. Now, $n_{k+1}/n_k \to 1$, hence rewriting the above as

$$\frac{n_k}{n_{k+1}}\frac{1}{n_k}U_{n_k} \leq \frac{1}{n}U_n \leq \frac{n_{k+1}}{n_k}\frac{1}{n_{k+1}}U_{n_{k+1}},$$

we see that on the event in (12), we also have $\frac{1}{n}U_n \to \mathbf{E}[X_1^+]$ and $\frac{1}{n}V_n \to \mathbf{E}[X_1^-]$. Putting these together with the almost sure assertion of (12), and recalling that $S_n = U_n - V_n$, we conclude that $\frac{1}{n}S_n \overset{a.s.}{\to} \mathbf{E}[X_1^+] - \mathbf{E}[X_1^-] = \mathbf{E}[X_1]$. ■

Now we return to the more difficult question of proving the strong law under first moment assumptions[5]. We shall reuse the idea from the previous proof of (1) proving almost sure convergence along a subsequence $\{n_k\}$ and then (2) getting a conclusion about the whole sequence from the subsequence . However, since we do not have second moment, we cannot use Chebyshev to take the sequence $n_k = k^2$ in the first step. In fact, we shall have to take an exponentially growing sequence $n_k = \alpha^k$, where $\alpha > 1$. But this is a problem for the second step, since $n_{k+1}/n_k \to \alpha$ whereas the proof above works only if we have $n_{k+1}/n_k \to 1$. Fortunately, we shall be able to take α arbitrarily close to 1 and thus bridge this gap! Another point is that as before, using positive random variables is necessary to be able to sandwich S_n between S_{n_k} and $S_{n_{k+1}}$. This will also feature in the proof below.

Proof of Theorem 22. **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 $\mu = \mathbf{E}[X_1] < \infty$

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

(Caution: Don't also assume zero mean in addition to non-negativity!). One consequence of non-negativity is that

$$\frac{S_{N_1}}{N_2} \le \frac{S_n}{n} \le \frac{S_{N_2}}{N_1} \text{ if } N_1 \le n \le N_2. \tag{13}$$

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 24

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 $1 \le k \le n_j$ write $X_k = Y_k + Z_k$ where $Y_k = X_k \mathbf{1}_{X_k \le n_j}$ 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 $j \ge J_\delta$, we have $\mathbf{E}[Z_1] \le \delta$. Let $S^Y_{n_j} = \sum_{k=1}^{n_j} Y_k$ and $S^Z_{n_j} = \sum_{k=1}^{n_j} Z_k$. Since $S_{n_j} = S^Y_{n_j} + S^Z_{n_j}$ and $\mathbf{E}[X_1] = \mathbf{E}[Y_1] + \mathbf{E}[Z_1]$, we get

$$\begin{aligned}
\mathbf{P}\left\{ \left| \frac{S_{n_j}}{n_j} - \mathbf{E}[X_1] \right| > 2\delta \right\} &\le \mathbf{P}\left\{ \left| \frac{S^Y_{n_j}}{n_j} - \mathbf{E}[Y_1] \right| + \left| \frac{S^Z_{n_j}}{n_j} - \mathbf{E}[Z_1] \right| > 2\delta \right\} \\
&\le \mathbf{P}\left\{ \left| \frac{S^Y_{n_j}}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} + \mathbf{P}\left\{ \left| \frac{S^Z_{n_j}}{n_j} - \mathbf{E}[Z_1] \right| > \delta \right\} \\
&\le \mathbf{P}\left\{ \left| \frac{S^Y_{n_j}}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} + \mathbf{P}\left\{ \frac{S^Z_{n_j}}{n_j} \ne 0 \right\}. \qquad (14)
\end{aligned}$$

We shall show that both terms in (14) are summable over j. The first term can be bounded by Chebyshev's inequality

$$\mathbf{P}\left\{ \left| \frac{S^Y_{n_j}}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} \le \frac{1}{\delta^2 n_j} \mathbf{E}[Y_1^2] = \frac{1}{\delta^2 n_j} \mathbf{E}[X_1^2 \mathbf{1}_{X_1 \le n_j}]. \tag{15}$$

while the second term is bounded by the union bound

$$\mathbf{P}\left\{ \frac{S^Z_{n_j}}{n_j} \ne 0 \right\} \le n_j \mathbf{P}(X_1 > n_j). \tag{16}$$

The right hand sides of (15) and (16) 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 (15) and (16) are summable in j, and hence it also follows that $\mathbf{P}\left\{ \left| \frac{S_{n_j}}{n_j} - \mathbf{E}[X_1] \right| > 2\delta \right\}$ is summable. This happens for every $\delta > 0$ and hence Lemma 18 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 (13) 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 $\lim_{n\to\infty} \frac{S_n}{n} = \mathbf{E}[X_1]$ $a.s.$ ■

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 $a.s.$ What is this constant?

If X_i have finite mean and $a_n = n$, the strong law tells us that the constant is zero. What if we divide by something smaller, such as n^α for some $\alpha < 1$? To probe this question further, let us assume that X_i are i.i.d. $\text{Ber}_\pm(1/2)$ random variables. Then using higher moments (just as we did in proving strong law under fourth moment assumption), we can get better results. For example, from the fact that $\mathbf{E}[S_n^4] = n + 3n(n-1)$ (check!), we can see that $\limsup \frac{S_n}{a_n} = 0$ $a.s.$ if $a_n = n^\alpha$ with $\alpha > \frac{3}{4}$. More generally, we reason as follows. For a positive integer p,

$$\mathbf{P}\{S_n \ge t_n\} \le \mathbf{E}[S_n^{2p}] t_n^{-2p} \quad \le \quad C_p n^p t_n^{-2p}$$

where we used the fact that $\mathbf{E}[S_n^{2p}] \le C_p n^p$ for a constant C_p. Assuming this, we see that if $t_n = n^\alpha$ with $\alpha > \frac{1}{2}$, then we can choose a p large enough to make the probabilities summable. By Borel-Cantelli it follows that $\limsup n^{-\alpha} S_n \overset{a.s.}{\to} 0$ as $n \to \infty$.

To see that $\mathbf{E}[S_n^{2p}] \le C_p n^p$, expand S_n^{2p} as a sum of monomial terms $X_1^{k_1} \dots X_n^{k_n}$ where k_i are non-negative integers that sum to $2p$. When we take expectations, this factors as $\mathbf{E}[X_1^{k_1}] \dots \mathbf{E}[X_n^{k_n}]$. If any k_i is odd, then the product is zero. If all k_is are even, the product is 1. We need to count the number of monomials of the latter type: Since each k_i is even, there are at most p of them that are non-zero. These can be chosen in $\binom{n}{p} \le n^p$ ways. Once the indices are chosen, the number of monomials are at most the number of ways to distribute $2p$ balls into p bins. Let this number be C_p. With all the overcounting, we still get $\mathbf{E}[S_n^{2p}] \le C_p n^p$, as claimed.

Instead of using moments, one may use Hoeffding's inequality to see that $\limsup \frac{S_n}{a_n} = 0$ even if $a_n = C\sqrt{n \log n}$ for a large enough constant C (Exercise!). In the converge direction, one can show that $\limsup \frac{S_n}{\sqrt{n}} = +\infty$, $a.s.$ (let us accept this without proof for now). This motivates the question of what is the right order of (limsup) growth of S_n?

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

The sharp answer, due to Khinchine is one of the great results of probability theory.

Theorem 25: Khinchine's law of iterated logarithm

Let X_i be i.i.d. $\text{Ber}_{\pm}(1/2)$ random variables. Then,

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

By symmetry, the liminf of $S_n/\sqrt{2n\log\log n}$ is equal to -1 almost surely. From these two, one can also deduce that the set of all limit points of the sequence $\{S_n/\sqrt{2n\log\log n}\}$ is equal to $[-1,1]$, almost surely.

The law of iterated logarithms was extended to general distributions with finite variance by Hartman and Wintner (with intermediate improvements by Kolmogorov and perhaps others). Here we only prove the theorem for Bernoullis (the general case is more complicated and a clean way to do it is via Brownian motion in the next course).

Result 26: Hartman-Wintner law of iterated logarithm

Let X_i be i.i.d. with mean μ and finite, non-zero variance σ^2. Then,

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

12. Proof of LIL for Bernoulli random variables

Let $X_1, X_2, \ldots$ be i.i.d. $\text{Ber}_{\pm}(1/2)$ random variables. Theorem 25 follows from the following two statements. For any $\delta > 0$, we have

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \leq 1+\delta \quad a.s. \tag{17}$$

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \geq 1-\delta \quad a.s. \tag{18}$$

Taking intersection over a countable number of $\delta = \frac{1}{k}$, $k \geq 1$, we get the statement of LIL. To motivate the principal idea in the proof, consider the following toy situation.

Example 9: Borel-Cantelli after blocking

Let B_n be events in a probability space and let $A_1 = B_1$, $A_2 = A_3 = B_2$, $A_4 = A_5 = A_6 = B_3$ and so on (n many A_is are equal to B_n). To show that only finitely many A_ns occur $a.s.$, if we apply Borel-Cantelli lemma to A_ns directly, we get the sufficient condition $\sum n\mathbf{P}(B_n) < \infty$. This is clearly foolish, as the event $\{A_n \text{ i.o.}\}$ is the same as $\{B_n \text{ i.o.}\}$, and the latter has zero probability whenever $\sum \mathbf{P}(B_n) < \infty$, a much weaker condition!

What this suggests is that when we have a sequence of A_ns and want to show that $\mathbf{P}\{A_n \text{ i.o.}\} = 0$, it may be good to combine together those A_is that are close to each other. For example, we can take a subsequence $1 = n_1 < n_2 < \dots$ and set C_k to be the union of A_ns with $n_k \le n < n_{k+1}$. If only finitely many C_ks occur, the only finitely many A_ns occur, and thus it suffices to show that $\sum_k \mathbf{P}(C_k) < \infty$. The naive union bound $\mathbf{P}(C_k) \le \sum_{n=n_k}^{n_{k+1}} \mathbf{P}(A_n)$ takes us back to the condition $\sum_n \mathbf{P}(A_n) < \infty$, but the point is that there may be better bounds for $\mathbf{P}(C_n)$ than the union bound.

Proof of the upper bound (17). Write $a_n = \sqrt{2n \log\log n}$. We want to show that only finitely many of the events $A_n = \{S_n > a_n(1+\delta)\}$ occur, $a.s.$ We use blocking as follows. Fix $\lambda > 1$ and set $n_k = \lfloor \lambda^k \rfloor$. Define the events

$$C_k = \bigcup_{n=n_k}^{n_{k+1}-1} A_n = \{S_n > a_n(1+\delta) \text{ for some } n_k \le n < n_{k+1}\},$$

$$B_k = \bigcup_{n=n_k}^{n_{k+1}-1} A_n = \{S_n > a_{n_k}(1+\delta) \text{ for some } n_k \le n < n_{k+1}\}.$$

Then $C_k \subseteq B_k$ as a_n is increasing in n. Thus if we show that $\sum_k \mathbf{P}(B_k) < \infty$, it follows that only finitely many C_n occur $a.s.$ and hence only finitely many A_n occur $a.s.$ We claim that

$$\mathbf{P}(B_k) \le C_\lambda k^{-(1+\delta)^2/\lambda} \quad \text{where } C_\lambda < \infty \text{ for any } \lambda > 1. \tag{19}$$

Granting this, it is clear that choosing $1 < \lambda < (1+\delta)^2$ ensures summability of $\mathbf{P}(B_k)$. We give two proofs of the above inequality below, which completes the proof. ■

Proof of (19) via the reflection principle: We shall need the following lemma which is of interest in itself.

Lemma 27: Reflection principle/Ballot problem

Let X_k be i.i.d. $\text{Ber}_\pm(1/2)$ random variables. Then for any $a \ge 0$, we have

$$\mathbf{P}\{\max\{S_0, \dots, S_n\} \ge a\} = 2\mathbf{P}\{S_n \ge a\}.$$

The proof is given in many places, we omit it here. Chapter-3 of Feller's vol-1 is highly recommended.

Returning to the proof of (19), if B_k occurs, then there is some $n \le n_{k+1}$ (in fact some $n \ge n_k$) such that $S_n \ge a_{n_k}(1+\delta)$. The reflection principle in Lemma 27 applies to give the bound

$$\begin{aligned}\mathbf{P}(B_k) &\le 2\mathbf{P}\{S_{n_{k+1}} \ge a_{n_k}(1+\delta)\} \\ &\le 2e^{-\frac{(1+\delta)^2 a_{n_k}^2}{2n_{k+1}}} \quad \text{(by Hoeffding's inequality)}.\end{aligned}$$

The exponent is (omitting integer part for simplicity of notation)

$$\frac{(1+\delta)^2 2\lambda^k \log\log \lambda^k}{2\lambda^{k+1}} = \frac{(1+\delta)^2}{\lambda}\log(k\log\lambda) \tag{20}$$

from which (19) immediately follows. ■

Proof of (19) via a modified first moment method: Let $X_k = \sum_{n=n_k}^{n_{k+1}-1} \mathbf{1}_{S_n > a_{n_k}(1+\delta)}$, so that $B_k = \mathbf{1}_{X_k \geq 1}$. We use the following improvement of Markov's inequality.

$$\mathbf{P}(B_k) = \mathbf{P}\{X_k \geq 1\} \leq \frac{\mathbf{E}[X_k]}{\mathbf{E}[X_k \mid X_k \geq 1]}.$$

What we need is an upper bound for the numerator and a lower bound for the denominator.

To get an upper bound for $\mathbf{E}[X_k]$, use Hoeffding's inequality to write

$$\mathbf{E}[X_k] = \sum_{n=n_k}^{n_{k+1}-1} \mathbf{P}\{S_n > a_{n_k}(1+\delta)\} \leq \sum_{n=n_k}^{n_{k+1}-1} \exp\left\{-\frac{a_{n_k}^2(1+\delta)^2}{2n}\right\}$$
$$\leq (n_{k+1}-n_k)\exp\left\{-\frac{a_{n_k}^2(1+\delta)^2}{2n_{k+1}}\right\}$$

where we bounded all terms by the largest one (which is the last one).

Next we claim that $c(n_{k+1}-n_k)$ (for some $c>0$) is a lower bound for $\mathbf{E}[X_k \mid X_k \geq 1]$. The heuristic idea is that if $X_k \geq 1$, there is some $N \in [n_k, n_{k+1})$ for which $S_N \geq a_{n_k}(1+\delta)$. If we fix that N and regard it as given, then $S_n - S_N$ has a symmetric distribution about 0 for any n, hence $\mathbf{P}\{S_n - S_N \geq 0\} \geq \frac{1}{2}$, which would imply that $\mathbf{E}[X_k \mid X_k \geq 1] \geq \frac{1}{2}(n_{k+1}-n_k)$. This reasoning is faulty, as the way we choose N (which is a random variable) may invalidate the claim that $S_n - S_N$ has a symmetric distribution.

To make the reasoning precise, write $X_k = Y_k + Z_k$ where Y_k is the number of n in the first half of the interval $[n_k, n_{k+1})$ for which $S_n > a_{n_k}(1+\delta)$ and Z_k is the analogous number for the second half of $[n_k, n_{k+1})$. Then $X_k \mathbf{1}_{X_k \geq 1} \geq \frac{1}{2}(Y_k \mathbf{1}_{Z_k \geq 1} + Z_k \mathbf{1}_{Y_k \geq 1})$ and $\{X_k \geq 1\} \subseteq \{Y_k \geq 1\} \cup \{Z_k \geq 1\}$. Consequently,

$$\mathbf{E}[X_k \mid X_k \geq 1] = \frac{\mathbf{E}[X_k \mathbf{1}_{X_k \geq 1}]}{\mathbf{P}\{X_k \geq 1\}} \geq \frac{1}{2}\frac{\mathbf{E}[Y_k \mathbf{1}_{Z_k \geq 1}] + \mathbf{E}[Z_k \mathbf{1}_{Y_k \geq 1}]}{\mathbf{P}\{Z_k \geq 1\} + \mathbf{P}\{Y_k \geq 1\}}$$
$$\geq \frac{1}{2}\min\left\{\frac{\mathbf{E}[Y_k \mathbf{1}_{Z_k \geq 1}]}{\mathbf{P}\{Z_k \geq 1\}}, \frac{\mathbf{E}[Z_k \mathbf{1}_{Y_k \geq 1}]}{\mathbf{P}\{Y_k \geq 1\}}\right\}$$
$$= \frac{1}{2}\min\{\mathbf{E}[Y_k \mid Z_k \geq 1], \mathbf{E}[Z_k \mid Y_k \geq 1]\}.$$

In the second line we used the elementary inequality $\frac{a+b}{c+d} \geq \min\{\frac{a}{c}, \frac{b}{d}\}$ valid for any non-negative numbers a, b, c, d. Now consider the second term inside the minimum. Since $Y_k \geq 1$, condition on the location N in the first half of $[n_k, n_{k+1})$ where $S_n > a_{n_k}(1+\delta)$ and use the fact that $S_n - S_N$, $n \geq N$, is still a simple symmetric random walk, and hence for any n in the second half, has

probability $1/2$ or more to be non-negative. Therefore, $\mathbf{E}[Z_k \mid Y_k \geq 1] \geq \frac{1}{4}(n_{k+1} - n_k)$. Similarly (considering the random walk in backwards direction starting from n_{k+1}), reason that $\mathbf{E}[Y_k \mid Z_k \geq 1] \geq \frac{1}{4}(n_{k+1} - n_k)$. Putting all this together, $\mathbf{E}[X_k \mid X_k \geq 1] \geq \frac{1}{8}(n_{k+1} - n_k)$.

Thus, $\mathbf{P}(B_k) \leq \frac{(n_{k+1}-n_k)\exp\left\{-\frac{a_{n_k}^2(1+\delta)^2}{2n_{k+1}}\right\}}{\frac{1}{8}(n_{k+1}-n_k)} \leq 8e^{-\frac{a_{n_k}^2(1+\delta)^2}{2n_{k+1}}}$. By the computation shown in (20), this is of the form given in (19). ■

13. 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 28: Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent, and assume that $|X_k| \leq d_k$ w.p.1. For simplicity assume that $\mathbf{E}[X_k] = 0$. Then, for any $n \geq 1$ and any $t > 0$,

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

Remark 7

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 $|Y_k| \leq 2d_k$ (because $|X_k| \leq d_k$ implies that $|\mathbf{E}[X_k]| \leq d_k$ and hence $|X_k - \mathbf{E}[X_k]| \leq 2d_k$). Thus, applying the result to Y_ks, we get

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

Proof. Without loss of generality, take $\mathbf{E}[X_k] = 0$. Now, if $|X| \leq d$ 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

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

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

elementary inequality $\cosh(x) \leq \exp(x^2/2)$ to get

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

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

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

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 $\mathbf{P}(|X_1| \leq d) = 1$.

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

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

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.

14. 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 10

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 29: Khinchine

Let X_n be independent random variables with finite second moment. Assume that $\mathbf{E}[X_n] = 0$ for all n and that $\sum_n \text{Var}(X_n) < \infty$. 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

$$\mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } n \geq N\right) = \lim_{m\to\infty} \mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } N \leq n \leq N+m\right). \tag{21}$$

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 10, 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 (21), 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. ■

What to do if the assumptions are not exactly satisfied? First, suppose that $\sum_n \text{Var}(X_n)$ 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 29 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 $\sum_n \text{Var}(Y_n) < \infty$, then it follows from the argument in the previous paragraph and Theorem 29 that $\sum Y_n$ converges a.s. Thus we have proved

Lemma 30: Kolmogorov's three series theorem - part 1

Suppose X_n are independent random variables. Suppose for some $A > 0$, the following hold with $Y_n := X_n \mathbf{1}_{|X_n| \leq A}$.

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

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 gives a complete and 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.

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 31: Central limit theorem for i.i.d. variables

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. In the rest of the section, we discuss various aspects of the theorem, and in later sections we give proofs of this and even more general central limit theorems.

Why this scaling?: 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 are at least subsequences that have distributional limits.

Why Normal distribution?: 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 $(2n)^{-\frac{1}{2}}S_{2n} \xrightarrow{d} Y$ and further,

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

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

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

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

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

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 $Y_1 + Y_2 \stackrel{d}{=} \sqrt{2}Y$ (where Y_1, Y_2, Y are i.i.d. $N(0, \sigma^2)$) is not trivial. Here are two ways to do it.

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

(2) If we assume further that Y has all moments, write $\alpha_k = \mathbf{E}[Y^k]$ and observe that

$$2^{m/2}\alpha_m = \sum_{k=0}^{m} \binom{m}{k} \alpha_k \alpha_{m-k} \quad \text{for all } m \geq 1.$$

Starting from $\alpha_0 = 1$, one deduces that $\alpha_1 = 0$ (because $\sqrt{2}\alpha_2 = 2\alpha_1$) and α_2 is arbitrary. But then onwards, it is clear that α_ms can be inductively deduced in terms of α_2. We leave it as an exercise to show that

$$\alpha_m = \begin{cases} 0 & \text{if } m \text{ is odd} \\ \alpha_2 \times (m-1) \times (m-3) \times \ldots \times 3 \times 1 & \text{if } m \text{ is even.} \end{cases}$$

But these are precisely the moments of $N(0, \alpha_2)$ distribution. Does that imply that Y must have $N(0, \alpha_2)$ distribution? The answer is yes[6], thus justifying the appearance of the normal distribution.

Justification by example: Assuming that $S_n/\sqrt{n}$ has a distributional limit, we have justified that the limit must be Gaussian. There are specific examples where one may easily verify the statement of the central limit theorem directly (indeed, that was how the theorem was arrived at).

One is of course the Demoivre-Laplace limit theorem (CLT for Bernoulli random variables), 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.

[6] A beautiful part of classical analysis is the *moment problem*, which asks whether a given sequence of numbers $(\alpha_m)_{m\geq 1}$ forms the moment sequence of a probability measure on $\mathbb{R}$, and if so, whether the measure is unique. There are precise answers to both questions, and an easy part of the answer is that any measure for which $\int e^{tx} d\mu(x)$ is finite for $|t| \leq \delta$ for some $\delta > 0$, has a unique moment sequence (i.e., no other measure can have the same sequence of moments as μ). This certainly applies to the Gaussian distribution.

Instead, let us consider the slightly less familiar case of exponential distribution. If X_i are i.i.d Exp(1) so that $\mathbf{E}[X_1] = 1$ and $\mathrm{Var}(X_1) = 1$. Then $S_n \sim \mathrm{Gamma}(n, 1)$ 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 (Scheffe's lemma) convergence of densities implies convergence in distribution and thus we get CLT for sums of exponential random variables.

Exercise 9

Prove the CLT for $X_1 \sim \mathrm{Ber}(p)$. Note that this also implies CLT for $X_1 \sim \mathrm{Bin}(k, p)$.

Justification under stronger hypotheses: 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. $\mathrm{Ber}_{\pm}(1/2)$. Then S_n has a symmetric distribution and hence all odd moments are zero (but first, $|S_n| \leq n$, hence all moments exist). For even moments,

$$\mathbf{E}[S_n^{2p}] = \sum_{1 \leq k_i \leq n} \mathbf{E}[X_{k_1} \dots X_{k_n}].$$

Fix $k = (k_1, \dots, k_{2p})$ and consider the corresponding summand. The expectation factors as a product of $\mathbf{E}[X^{\ell_i}]$, $1 \leq i \leq n$, where ℓ_i is the number of j for which $k_j = i$. Unless each ℓ_i is even, the summand vanishes. The terms for which each ℓ_i contribute 1 each, and these terms may be divided into two parts.

First, those in which each ℓ_i is 0 or 2. The number of ways to ways to choose the p indices i for which $\ell_i = 2$ is $n(n-1)\dots(n-p+1)$, and the number of ways that these indices may be chosen is $(2p-1)(2p-3)\dots(3)(1)$.

Next those terms in which at least one ℓ_i is equal to 4. The contribution is afgThere are n ways to choose that i, and there are

16. Strategies of proof of central limit theorem

In the next few sections, we shall prove CLT as stated in Theorem 31 as well as a more general CLT for triangular arrays to be stated in Theorem 36. 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 31 (assuming that X_1 has finite third moment) by both approaches. Then we prove the more general Theorem 36 (which implies Theorem 31 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 32

$Y_n \xrightarrow{d} Y$ 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 $k \geq 1$ define a function $f_k \in C_b(\mathbb{R})$ such that $0 \leq f_k \leq 1$, $f_k(x) = 1$ for $x \leq t$ and $f_k(x) = 0$ for $x \geq t + \frac{1}{k}$. 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 $F_Y(t) \leq \mathbf{E}[f_k(Y)] \leq F_Y(t + \frac{1}{k})$ and $F_{Y_n}(t) \leq \mathbf{E}[f_k(Y_n)] \leq F_{Y_n}(t + \frac{1}{k})$. Hence, $\limsup_{n\to\infty} F_{Y_n}(t) \leq F_Y(t + \frac{1}{k})$. This being true for every k, we let $k \to \infty$ and get $\limsup_{n\to\infty} F_{Y_n}(t) \leq F_Y(t)$. Similarly, use the function $g_k(x) := f_k(x + \frac{1}{k})$ to get

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

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

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 33

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, $Y_n \xrightarrow{d} Y$.

Proof. Repeat the proof given for Lemma 32 but take f_k to be a smooth function such that $0 \leq f_k \leq 1$, $f_k(x) = 1$ for $x \leq t$ and $f_k(x) = 0$ for $x \geq t + \frac{1}{k}$. ■

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 34: 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, $Y_n \xrightarrow{d} Y$.

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.

16.1. **General approach to proving central limit theorem(s).** The statement of central limit theorem is that $\mathbf{E}[\varphi(S_n/\sqrt{n})] \to \mathbf{E}[\varphi(Z)]$ whenever $\varphi \in \{\mathbf{1}_{(-\infty,t]} : t \in \mathbb{R}\}$ (and Z denotes a $N(0,1)$ random variable). We do not have a direct handle on the expectations of indicator variables. The point of the previous discussions is that we can replace them by a suitable class of nicer functions.

Characteristic functions: For example, to invoke Lemma 34, we only need to prove $\mathbf{E}[e_t(S_n/\sqrt{n})] \to \mathbf{E}[e_t(Z)]$ where $e_t(x) = e^{itx}$. The usefulness of this comes from the fact that $e_t(S_n/\sqrt{n}) = \prod_{k=1}^n e_t(X_{n,k})$ where $X_{n,k} = X_k/\sqrt{n}$ and by the independence assumption, the expectation factors as $\prod_{k=1}^n \mathbf{E}[e_t(X_{n,k})]$. How this is handled will be seen later in the proof. We should also know what it is supposed to converge to, namely $\mathbf{E}[e_t(Z)]$. It is shown in the appendix 21 that $\mathbf{E}[e_t(Z)] = e^{-t^2/2}$. Thus the proof of CLT reduces to showing that

$$\prod_{k=1}^{n} \mathbf{E}[e_t(X_{n,k})] \to e^{-\frac{1}{2}t^2} \quad \text{as } n \to \infty.$$

Invariance principle: The other method of proof that we show is to use Lemma 33. Then we need to show that $\mathbf{E}[f(S_n/\sqrt{n})] \to \mathbf{E}[f(Z)]$ for all $f \in C_b^{(3)}(\mathbb{R})$. Unlike for complex exponentials, we do not have any formula[7] for $\mathbf{E}[f(Z)]$ for general f. Our approach will be to show that if X_i are i.i.d. and Y_i are i.i.d., both having zero means and unit variances, then $\mathbf{E}[f(S_n^X/\sqrt{n})] \approx \mathbf{E}[f(S_n^Y/\sqrt{n})]$ for large n. If Y_i are i.i.d. $N(0,1)$, the right hand side is precisely $\mathbf{E}[f(Z)]$, and from that we shall be able to prove that the left hand side converges to $\mathbf{E}[f(Z)]$, for $f \in C_b^{(3)}$.

[7]However as shown in the appendix, we do have the identity $\mathbf{E}[Zf(Z)] = \mathbf{E}[f'(Z)]$ for all nice enough f. Further, it can be shown that if a random variable Z satisfies this for a large class of f, then $Z \sim N(0,1)$. Charles Stein found a proof of central limit theorem by showing (a) If $W = S_n/\sqrt{n}$, then $\mathbf{E}[Wf(W)] \approx \mathbf{E}[f'(W)]$ for large enough n, and (b) this approximate identity implies that W has approximately $N(0,1)$ distribution. . This is known as *Stein's method*, and has some advantages over the usual proofs.

17. Central limit theorem - two proofs assuming third moments

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

Theorem 35

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 31.

17.1. Proof via characteristic functions.

We shall need the following facts.

Exercise 10

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

Proof of Theorem 35. By Lévy's continuity theorem (Lemma 34), 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,

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

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 10 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

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

Add up over k from 1 to n to get

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

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 33 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 36: Lindeberg-Feller CLT

Suppose $X_{n,k}$, $k \le n$, $n \ge 1$, are random variables. We assume that

(1) For each n, the random variables $X_{n,1}, \dots, X_{n,n}$ 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 **??** *from Theorem 36.* Let $X_{n,k} = n^{-\frac{1}{2}} X_k$ for $k = 1, 2, \dots, n$. 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 11

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

To fit it in the framework of Theorem 36, 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 36 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 11

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

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

Example 12

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 $X_{n,k} = n^{-\frac{3}{2}} k X_k$. Let us check the crucial third condition of Theorem 36.

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

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 12

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 13

If $z_k, w_k \in \mathbb{C}$ and $|z_k|, |w_k| \le \theta$ for all k, then $\left| \prod_{k=1}^{n} z_k - \prod_{k=1}^{n} w_k \right| \le \theta^{n-1} \sum_{k=1}^{n} |z_k - w_k|$.

Proof of Theorem 36. The characteristic function of $S_n = X_{n,1} + \ldots + X_{n,n}$ is given by $\psi_n(t) = \prod_{k=1}^{n} \mathbf{E}\left[e^{itX_{n,k}}\right]$. Again, we shall use the Taylor expansion of e^{itx}, but we shall need both the second and first order expansions.

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

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

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

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

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

where, $R_{n,k}(t) := \frac{t^2}{2}\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|>\delta}X_{n,k}^2\left(1 - e^{itX_{n,k}^+}\right)\right] - \frac{it^3}{6}\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|\le\delta}X_{n,k}^3e^{itX_{n,k}^*}\right]$. 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

$$|R_{n,k}(t)| \le t^2\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|>\delta}X_{n,k}^2\right] + \frac{|t|^3\delta}{6}\mathbf{E}\left[X_{n,k}^2\right]. \tag{22}$$

We want to apply Exercise 13 to $z_k = \mathbf{E}\left[e^{itX_{n,k}}\right]$ and $w_k = 1 - \frac{1}{2}\sigma_{n,k}^2t^2$. Clearly $|z_k| \le 1$ by properties of c.f. If we prove that $\max_{k\le n}\sigma_{n,k}^2 \to 0$, then it will follow that $|w_k| \le 1$ and hence with $\theta = 1$ in Exercise 13, we get

$$\begin{aligned} \limsup_{n\to\infty}\left|\prod_{k=1}^{n}\mathbf{E}\left[e^{itX_{n,k}}\right] - \prod_{k=1}^{n}\left(1 - \frac{1}{2}\sigma_{n,k}^2t^2\right)\right| &\le \limsup_{n\to\infty}\sum_{k=1}^{n}|R_{n,k}(t)| \\ &\le \frac{1}{6}|t|^3\sigma^2\delta \quad \text{(by 22)} \end{aligned}$$

To see that $\max_{k\le n}\sigma_{n,k}^2 \to 0$, 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 $\max_{k\le n}\sigma_{n,k}^2 \to 0$ as $n \to \infty$. As $\delta > 0$ is arbitrary, we get

$$\lim_{n\to\infty}\prod_{k=1}^{n}\mathbf{E}\left[e^{itX_{n,k}}\right] = \lim_{n\to\infty}\prod_{k=1}^{n}\left(1 - \frac{1}{2}\sigma_{n,k}^2t^2\right). \tag{23}$$

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

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

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

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

From 23 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,

$$\begin{aligned}\mathbf{E}\left[f\left(V_n\right)\right] - \mathbf{E}[f(V_0)] &= \sum_{k=1}^n \mathbf{E}\left[f\left(V_k\right) - f\left(V_{k-1}\right)\right] \\ &= \sum_{k=1}^n \mathbf{E}\left[f\left(V_k\right) - f\left(U_k\right)\right] - \sum_{k=1}^n \mathbf{E}\left[f\left(V_{k-1}\right) - f\left(U_k\right)\right].\end{aligned} \tag{24}$$

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

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

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

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

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

$$|\mathbf{E}[f(V_k)] - \mathbf{E}[f(V_{k-1})]| \le \frac{1}{2}\mathbf{E}[|f''(U_k)|] \left|\mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|\le\delta}] - \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|\le\delta}]\right| \tag{25}$$

$$+ \frac{1}{2}\left|\mathbf{E}[|f''(U_k^\#)|X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}]\right| + \frac{1}{2}\left|\mathbf{E}[|f''(U_k^{\#\#})|Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\right| \tag{26}$$

$$+ \frac{1}{6}\left|\mathbf{E}[|f'''(U_k^*)||X_{n,k}|^3 \mathbf{1}_{|X_{n,k}|\le\delta}]\right| + \frac{1}{6}\left|\mathbf{E}[|f'''(U_k^{**})||Y_{n,k}|^3 \mathbf{1}_{|Y_{n,k}|\le\delta}]\right| \tag{27}$$

Since $\mathbf{E}[X_{n,k}^2] = \mathbf{E}[Y_{n,k}^2]$, the term in the first line (25) 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 (26) 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 (27), we show how to deal with the first.

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

The same bound holds for the second term in (27). Putting all this together, we arrive at

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

Add up over k and use (24) to get

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

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 13

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., $X \stackrel{d}{=} -X$. 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) symmetric distribution μ_α such that $X + Y \stackrel{d}{=} 2^{\frac{1}{\alpha}} X$ 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 11

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* 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[8].

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

(B) The characteristic function 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) Some basic observations.

Theorem 37

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

■

Lemma 38

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

Proof. Clearly $\hat{\mu}(0) = 1$ and $|\hat{\mu}(t)| \leq \int |e^{itx}| d\mu(x) = 1$. Further,

$$|\hat{\mu}(t+h) - \hat{\mu}(t)| \leq \int |e^{i(t+h)x} - e^{itx}| d\mu(x) = \int |e^{ihx} - 1| d\mu(x).$$

As $h \to 0$, the integrand $|e^{ihx} - 1| \to 0$ and is also bounded by 2. Hence by the dominated convergence theorem, the integral goes to zero as $h \to 0$. The uniformity is clear as there is no dependence on t. ■

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.

[8] 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.

Theorem 39

Let $\mu \in \mathcal{P}(\mathbb{R})$. If μ has finite k^{th} moment for some $k \in \mathbb{N}$, then $\hat{\mu} \in C^{(k)}(\mathbb{R})$ and $\hat{\mu}^{(k)}(t) = \int_{\mathbb{R}} (ix)^k e^{itx} d\mu(x)$.

Theorem 40

Let $\mu \in \mathcal{P}(\mathbb{R})$. Assume that μ has density f with respect to Lebesgue measure.

(1) (Riemann-Lebesgue lemma). $\hat{\mu}(t) \to 0$ as $t \to \pm\infty$.

(2) If $f \in C^{(k)}$, then $\hat{\mu}(t) = o(|t|^{-k})$ as $t \to \pm\infty$.

For proofs, consult, Feller's book.

(B) Examples. We give some examples.

(1) If $\mu = \delta_0$, then $\hat{\mu}(t) = 1$. More generally, if $\mu = p_1\delta_{a_1} + \ldots + p_k\delta_{a_k}$, then $\hat{\mu}(t) = p_1 e^{ita_1} + \ldots + p_k e^{ita_k}$.

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

(3) If $X \sim$ Exp(λ), then $\psi_X(t) = \int_0^\infty \lambda e^{-\lambda x} e^{itx} dx = \frac{\lambda}{\lambda - it}$. If $Y \sim$ Gamma(ν, λ), then if ν is an integer, then $Y \stackrel{d}{=} X_1 + \ldots + X_\nu$ 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.

(4) Laplace distribution having density $\frac{1}{2}e^{-|x|}$ on all of $\mathbb{R}$ has characteristic function $\frac{1}{1+t^2}$. This is similar to the previous example and left as an exercise.

(5) $Y \sim$ Normal(μ, σ^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[9].

[9]Here is the argument: Fix $R > 0$ and let $\gamma(u) = u$ and $\eta(t) = u + it$ for $-R \le u \le R$ and let $\eta'_x(s) = x + is$ for $0 \le s \le t$. The integral that we want is the limit of the contour integrals $\int_\eta e^{-\frac{1}{2}z^2} dz$ as $R \to \infty$. Since the integrand has no poles, this is the same as the integral $\int_\gamma + \int_{\eta'_R} - \int_{\eta'_{-R}}$ of $e^{-z^2/2}$. The integral over γ converges to $\int_{\mathbb{R}} e^{-x^2/2} dx$ which is $\sqrt{2\pi}$. The integrals over η'_R and η'_{-R} converge to zero as $R \to \infty$. This is because the absolute value of the integrand is $e^{-\frac{1}{2}(R^2+s^2)} \le e^{-R^2/2}$ for any $0 \le s \le t$. Thus the two integrals are bounded in absolute value by $e^{-R^2/2}|t|$ which goes to 0 as $R \to \infty$.

Alternately, one can obtain the characteristic function as follows.

Stein's equation: Let $f : \mathbb{R} \mapsto \mathbb{R}$ be any reasonable function (C_b^1 is more than needed). $\mathbf{E}[Zf(Z)] = \mathbf{E}[f'(Z)]$ (this is called *Stein's equation*).

To see this, integrate by parts to get

$$\mathbf{E}[f'(Z)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f'(x) e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int f(x) x e^{-x^2/2} dx = \mathbf{E}[Zf(Z)].$$

since the boundary terms vanish (provided f is grows slowly enough at $\pm\infty$). Take $f(x) = e^{itx}$ with a fixed $t \in \mathbb{R}$ to get $it\mathbf{E}[e^{itX}] = \mathbf{E}[Xe^{itX}] = \frac{1}{i}\frac{d}{dt}\mathbf{E}[e^{itX}]$ (where the last inequality is by differentiation under the expectation, which can be justified easily by dominated convergence theorem). Thus, $\psi'(t) = -\psi(t)$, which gives $\psi(t) = Ce^{-t^2/2}$. As $\psi(0) = 1$, we must have $C = 1$.

The final conclusion is that $N(\mu, \sigma^2)$ has characteristic function $e^{it\mu - \frac{\sigma^2 t^2}{2}}$.

(6) Let μ be the standard Cauchy measure $\frac{1}{\pi(1+x^2)}dx$. Let $t > 0$ and consider $\psi(t) = \frac{1}{\pi}\int \frac{e^{itx}}{1+x^2}dx$. We use contour integration. Let $\gamma(u) = u$ for $-R \le u \le R$ and $\eta(u) = Re^{is}$ for $0 \le s \le \pi$. Then by the residue theorem

$$\frac{1}{\pi}\int_\gamma \frac{e^{itz}}{1+z^2}dz + \frac{1}{\pi}\int_\eta \frac{e^{itz}}{1+z^2}dz = \frac{1}{\pi} \times 2\pi i \text{Res}\left(\frac{e^{itz}}{1+z^2}, i\right) = e^{-t}.$$

However, on η, the integrand is bounded by $\frac{e^{-t\,\mathrm{Im}\,z}}{|1+z^2|} \le \frac{1}{R^2-1}$, since $t > 0$. The length of the contour is πR, hence the total integral over η is $O(1/R)$ as $R \to \infty$. Thus, $\frac{1}{\pi}\int_\gamma \frac{e^{itx}}{1+x^2}dx$ converges to e^{-t} for $t > 0$. By the symmetry of the underlying measure, $\psi(-t) = \psi(t)$, whence we arrive at $\psi(t) = e^{-|t|}$.

(C) Inversion formulas.

Theorem 41

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

Proof. Let θ_σ denote the 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 $a < b$,

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

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 μ. ■

We can continue the reasoning in the above proof to get a formula for recovering a measure from its characteristic function.

Corollary 42: Fourier inversion formula

Let $\mu \in \mathcal{P}(\mathbb{R})$.

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

$$\mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\} = \lim_{\sigma\to\infty} \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-iat} - e^{-ibt}}{it} \hat{\mu}(t) e^{-\frac{t^2}{2\sigma^2}} dt \tag{28}$$

(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 (28) 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 41 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 Fourier transform 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 43: Lévy's continuity theorem

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 and ψ is continuous at 0, 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 $\hat{\mu}_n(t) \to \psi(t)$ pointwise for all t and ψ is continuous at zero. 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[10]. From Lemma 44 below, as $n \to \infty$,

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

where the last implication follows by DCT (since $1-\hat{\mu}_n(t) \to 1-\psi(t)$ for each t and also $|1-\hat{\mu}_n(t)| \leq 2$ for all t). Further, as $\delta \downarrow 0$, we get $\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\psi(t))dt \to 0$ (because, $1-\hat{\mu}(0) = 0$ and ψ is continuous at 0). Thus, given $\varepsilon > 0$, we can find $\delta > 0$ such that $\limsup_{n\to\infty} \mu_n\left([-2/\delta, 2/\delta]^c\right) < \varepsilon$. This means that for some finite N, we have $\mu_n\left([-2/\delta, 2/\delta]^c\right) < \varepsilon$ for all $n \geq N$. Now, find $A > 2/\delta$ such that for any $n \leq N$, we get $\mu_n\left([-2/\delta, 2/\delta]^c\right) < \varepsilon$. Thus, for any $\varepsilon > 0$, we have produced an $A > 0$ so that $\mu_n\left([-A, A]^c\right) < \varepsilon$ for all n. This is the definition of tightness. ■

[10] I would like to thank Pablo De Nápoli for pointing out a flaw in the statement and proof of the second part.

Lemma 44

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

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

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

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

(D) Positive semi-definiteness. What functions arise as characteristic functions of probability measures on $\mathbb{R}$? If $\varphi(t) = \int e^{itx}d\mu(x)$ for a probability measure μ, then $\varphi(-t) = \overline{\varphi(t)}$ for all $t \in \mathbb{R}$. Further, for any $m \geq 1$ and any complex numbers $c_1, \ldots, c_m$ and any real numbers $t_1, \ldots t_m$, we must have

$$\begin{aligned}0 \leq \int\left|\sum_{k=1}^{m}c_k e^{it_k x}\right|^2 d\mu(x) &= \sum_{k,\ell=1}^{n}c_k\bar{c}_\ell\int e^{i(t_k-t_\ell)x}d\mu(x) \\ &= \sum_{k,\ell=1}^{n}c_k\bar{c}_\ell\varphi(t_k-t_\ell).\end{aligned}$$

This motivates the following definition.

Definition 12: Positive definite functions

A function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is said to be *positive definite* if the matrix $M_\varphi[t_1, \ldots, t_n] := (\varphi(t_j - t_k))_{1\leq j,k\leq n}$ is Hermitian and positive semi-definite for any $n \geq 1$ and any $t_1, \ldots, t_n \in \mathbb{R}$.

Thus characteristic functions are necessarily positive definite functions. We have also seen that they are continuous and take the value 1 at 0. These are all the properties that it takes to make a characteristic function.

Theorem 45: Bochner's theorem

A function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is a characteristic function of a Borel probability measure on $\mathbb{R}$ if and only if φ is continuous, positive definite and $\varphi(0) = 1$.

Before starting the proof, we make some basic observations about positive definite functions.

- If φ is positive definite, then $|\varphi| \leq 1$. Indeed, for any t, the positive semi-definiteness of $M_\varphi[0, t]$ shows that $1 - |\varphi(t)|^2 \geq 0$ (note that $\varphi(-t) = \overline{\varphi(t)}$ is part of the condition of positive definiteness).
- If φ and ψ are positive definite functions and $\theta(t) = \varphi(t)\psi(t)$, then θ is also positive definite. The matrix $C = M_\theta[t_1, \ldots, t_n]$ is the Hadamard product (entry-wise product) of $A = M_\varphi[t_1, \ldots, t_n]$ and $B = M_\psi[t_1, \ldots, t_n]$. It is a theorem of Schur that a Hadamard product of positive semi-definite matrices is also positive demi-definite. It is not hard to see: As A is positive semi-definite, we can find random variables $X_1, \ldots, X_n$ such that $a_{i,j} = \mathbf{E}[X_i X_j]$. Similarly $B = \mathbf{E}[Y_i Y_j]$ for some random variables $Y_1, \ldots, Y_n$. We can construct X_is and Y_js on the same probability space, so that $(X_1, \ldots, X_n)$ is independent of $(Y_1, \ldots, Y_n)$. Then, the covariance matrix of $Z_i = X_i Y_i$, $1 \leq i \leq n$, is precisely C. Hence C is positive semi-definite.
- For any nice function $c : \mathbb{R} \mapsto \mathbb{C}$, we have

$$\iint c(t)\overline{c(s)}\varphi(t-s)dtds \geq 0. \tag{29}$$

  This is just a continuum analogue of $\sum_{j,k} c_j \overline{c_k} \varphi(t_j - t_k)$ and can be got by approximating the integral by sums. We omit details.

Now we come to the proof of Bochner's theorem. What we need to prove is that given a continuous positive definite function φ satisfying $\varphi(0)$, there is a probability measure whose characteristic function it is. The idea is simple. We have already seen inversion formulas that recover a measure from its characteristic function. We just apply these inversion formulas to φ and then try to show that the object we get is a probability measure.

Proof of Bochner's theorem. Let φ be a continuous, positive-definite function such that $\varphi(0) = 1$.

Case: φ is absolutely integrable: Taking a cue from the Fourier inversion formula,

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi(t) e^{-itx} dt.$$

The integral is well-defined as φ is bounded. We want to show that f is a probability density. First we show that f is non-negative[11]. Fix an interval $I_M = [-M, M]$ and observe that

$$\begin{aligned} f(x) &= \frac{1}{2\pi(2M)} \int_{I_M} \int_{\mathbb{R}} e^{ix(t-s)} \varphi(t-s) dt ds \quad \text{(the inner integral does not depend on } s) \\ &= \frac{1}{2\pi(2M)} \int_{I_M} \int_{I_M} e^{ix(t-s)} \varphi(t-s) dt ds + \frac{1}{2\pi(2M)} \int_{I_M} \int_{I_M^c} e^{ix(t-s)} \varphi(t-s) dt ds. \end{aligned}$$

The first integral is positive by (29) (take $c(t) = e^{ixt} \mathbf{1}_{|t| \le M}$). As for the second integral, we claim that it goes to zero as $M \to \infty$. Indeed, fix $\delta > 0$ and observe that for $|s| \le (1-\delta)M$, the inner integral is less than $c_M := \int_{I_{\delta M^c}} |\varphi(u)| du$ (as $|t-s| \ge \delta M$ for any $|s| < (1-\delta)M$ and any $|t| > M$). If $|s| > (1-\delta)M$, we just use the trivial bound $C := \int_{\mathbb{R}} |\varphi|$ for the inner integral. Overall, the bound for the second term becomes

$$\frac{1}{2\pi(2M)} (2(1-\delta)M c_M + C\delta M) \le c_M + \delta C.$$

Let $M \to \infty$ and then $\delta \downarrow 0$ (or just take $\delta = \frac{1}{\sqrt{M}}$) to see that this goes to zero as $M \to \infty$. This proves that $f(x) \ge 0$ for all x.

The formula for f shows that it is the inverse Fourier transform (need an argument first showing integrability of f) of φ (up to a factor of $1/2\pi$). Applying the Fourier inversion formula, we see that $\varphi(t) = \int_{\mathbb{R}} f(x) e^{itx} dx$, showing that φ is the characteristic function of the measure $f(x)dx$. In particular, $\int_{\mathbb{R}} f(x)dx = \varphi(0) = 1$ showing that f is a probability density.

General case: For any $\sigma > 0$, define $\varphi_\sigma(t) = \varphi(t) e^{-\sigma^2 t^2/2}$ (the idea behind: If φ is the characteristic function of a random variable X, then φ_σ would be that of $X + \sigma Z$, where $Z \sim N(0,1)$). Since φ is bounded, φ_σ is absolutely integrable for any $\sigma > 0$. Further, φ_σ is continuous and positive definite by the Schur product theorem. Thus, by the first case, φ_σ is the characteristic function of a measure μ_σ (in fact, $d\mu_\sigma(x) = f_\sigma(x)dx$, where $f_\sigma(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \varphi_\sigma(t) dt$).

$\varphi_\sigma \to \varphi$ point-wise as $\sigma \downarrow 0$. By the second part of Lévy's continuity theorem, this shows that φ is a characteristic function of the probability measure μ which is the distributional limit of μ_σ as $\sigma \downarrow 0$. ■

[11] It may be easier to first see the following formal argument. Fix $x \in \mathbb{R}$ and use $c(t) = e^{ixt}$ in (29) to get

$$\begin{aligned} 0 \le \iint e^{ix(t-s)} \varphi(t-s) dt ds &= \int \left[\int e^{ixu} \varphi(u) du \right] ds \\ &= f(x) \left(\int 1 ds \right). \end{aligned}$$

Of course, the integral here is infinite, hence the proof is only formal, but it gives a hint why $f(x) \ge 0$. The actual proof makes this precise by integrating s over a finite interval.

Remark 8

Fourier analysis on general locally compact abelian groups goes almost in parallel to that on the real line. If G is a locally compact abelian group (eg., $\mathbb{R}^d$, $(S^1)^d$, $\mathbb{Z}^d$, finite abelian groups, their products), then the set of characters (continuous homomorphisms from G to S^1) form a collection $\hat{G}$ called the dual of G. It can be endowed with a topology (basically of point-wise convergence on G) and these characters form a dense set in $L^2(G)$ (w.r.t. Haar measure). For a measure μ on G, one defines its Fourier transform $\hat{\mu} : \hat{G} \mapsto \mathbb{C}$ by $\hat{\mu}(\chi) = \int_G \chi(x) d\mu(x)$. Plancherel's theorem, Lévy's theorem, Bochner's theorem all go through with minimal modification of language[a].

[a] A good resource is the book *Fourier analysis on groups* by Walter Rudin.