Probability theory

Manjunath Krishnapur

Contents

CHAPTER 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 characteristic functions, a tool of great importance, 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. 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.

[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.

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 $\sum\limits_{\omega \in \Omega} p(\omega) = 1$. Then define $\mathbf{P} : 2^{\Omega} \to [0,1]$ by $\mathbf{P}(A) = \sum\limits_{\omega \in A} p(\omega)$.

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

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

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. In the next section, a very short introduction to Expectation is given, but for the construction and details, refer to Part-1.

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.

2. User's guide to expectation

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let RV denote the set of all random variables and let RV_+ denote the set of all non-negative random variables on this probability space. Here is the fundamental fact:

Fact: There is a unique function $\mathbf{E} : \text{RV}_+ \to [0, \infty]$ such that

(1) *Linearity:* $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ and $\mathbf{E}[cX] = c\mathbf{E}[X]$ for all $X, Y \in \text{RV}_+$ and for all $c \geq 0$.

(2) *Positivity:* $\mathbf{E}[X] \geq 0$ with equality if and only if $X = 0$ a.s.

(3) *MCT (Monotone convergence theorem):* If $X_n, X \in \text{RV}_+$ and $X_n \uparrow X$ a.s., then $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$.

(4) $\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$ for all $A \in \mathcal{F}$.

We do not go into the construction of expectation (also called Lebesgue integral). But it is worth noting that accepting the above fact, one has the following explicit form: For any $X \in \text{RV}_+$,

$$\mathbf{E}[X] = \lim_{n \to \infty} \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mathbf{P}\left\{ \frac{k}{2^n} \leq X < \frac{k+1}{2^n} \right\}.$$

This is got by observing that $X_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\frac{k}{2^n} \leq X < \frac{k+1}{2^n}}$ are increase to X pointwise, and hence $\mathbf{E}[X] = \mathbf{E}[X_n]$ by the MCT. And $\mathbf{E}[X_n]$ can be got from linearity.

One may also take the above formula as the definition of expectation (it is not hard to see that the limit exists) and prove that it satisfies the four properties stated above.

For general $X \in \text{RV}$, we write it as $X = X_+ - X_-$ where $X_+ = X \vee 0$ and $X_- = (-X)_+ = -(X \wedge 0)$. If $\mathbf{E}[X_+]$ and $\mathbf{E}[X_-]$ are both finite, then we say that X has expectation (or that X is integrable) and define $\mathbf{E}[X] = \mathbf{E}[X_+] - \mathbf{E}[X_-]$. Observe that $X_+ + X_- = |X|$, hence integrability is equivalent to $\mathbf{E}[|X|] < \infty$. We also write $X \in L^1$ if $X$is integrable (although L^1 is a space defined via an equivalence relation). More generally, if $|X|^p$ is integrable, we write $X \in L^p$ (or $L^p(\mathbf{P})$ or $L^p(\Omega, \mathcal{F}, \mathbf{P})$ if we needed).

2.1. Limit properties. Apart from MCT we also have the following very important facts.

(1) *Fatou's lemma:* If $X_n \in \text{RV}_+$, then $\mathbf{E} \liminf \mathbf{E}[X_n] \geq \mathbf{E}[\liminf X_n]$.

(2) *DCT:* If $X_n \to X$ a.s., if $|X_n| \leq Y$ for some integrable Y, then X_n, X are integrable and $\mathbf{E}[X_n] \to \mathbf{E}[X]$. In fact, $\mathbf{E}[|X_n - X|] \to 0$.

Fatou's lemma follows directly from MCT by observing that $Y_n := \inf_{k \geq n} X_k$ increase to $Y := \liminf X_n$ and that $Y_n \leq X_n$. DCT follows by applying Fatou's lemma to $Y - X_n$ and to $Y + X_n$.

2.2. Inequalities. Cauchy-Schwarz, Hölder's and Minkowski's inequalities are important and repeatedly used. These are explained in Part-1 of these lecture notes. Another set of all important inequalities are those of Markov and Chebyshev, and their generalizations. These are explained in the following sections.

2.3. Connection to independence. In general, statements for events have analogous statements for random variables and vice versa. Here is an illustration of how this works for independence of sigma-algebras (which was defined in terms of events).

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Sub sigma-algebras $\mathcal{G}_1, \ldots, \mathcal{G}_m$ are independent if and only if $\mathbf{E}[X_1 \ldots X_m] = \mathbf{E}[X_1] \ldots \mathbf{E}[X_m]$ for any bounded random variables X_i such that X_i is $\mathcal{G}_i$ measurable.

CHAPTER 2

Some basic tools in probability

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

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 the sum of expectations) than to calculate probabilities (example, tail probability of a sum of random variables), inequalities that bound probabilities in terms of moments may be expected to be somewhat useful. In fact, they are extremely useful!

Lemma 3: First moment method or Markov's inequality

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

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 less than 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. Observe that in the following instances X is not assumed to be positive, but Markov's inequality is applied to positive functions of X.

(1) Markov's inequality asserts that the tail of a random variable with finite expectation must decay at least as fast as $1/t$. In fact, the proof shows that if X is integrable then

$$\mathbf{P}\{|X| \geq t\} \leq \frac{1}{t}\mathbf{E}[|X|\mathbf{1}_{|X| \geq t}] \;=\; o(1/t)$$

since $\mathbf{E}[|X|\mathbf{1}_{|X| \geq t}] \to 0$ by DCT.

(2) If X has finite variance, applying Markov's inequality to $(X-\mathbf{E}[X])^2$ gives

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

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

(3) If $\mathbf{E}[e^{\lambda X}]<\infty$ for some $\lambda>0$, we get $\mathbf{P}\{X>t\}=\mathbf{P}\{e^{\lambda X}>e^{\lambda t}\}\leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$. This is an even better bound as it decays exponentially as $t\to\infty$.

1.1. A different sort of strengthening of Markov's inequality. In many situations, the following strengthening turns out to be useful. If X is a positive random variable, then

$$\mathbf{P}\{X\geq t\}\leq\frac{\mathbf{E}[X]}{\mathbf{E}[X\,\big|\,X\geq t]}. \tag{2}$$

We have not yet defined conditional expectation. For now, it can be interpreted in the elementary fashion

$$\mathbf{E}[X\,\big|\,X\geq t]=\frac{\mathbf{E}[X\mathbf{1}_{X\geq t}]}{\mathbf{P}\{X\geq t\}}.$$

In particular, if X takes values in $\mathbb{N}$ with $p_k=\mathbf{P}\{X=k\}$, then

$$\mathbf{E}[X\,\big|\,X\geq t]=\frac{kp_k+(k+1)p_{k+1}+\dots}{p_k+p_{k+1}+\dots}.$$

In any case, it is clear that $\mathbf{E}[X\,\big|\,X\geq t]\geq t$, hence it is stronger than Markov's inequality. To give a caricature of its usefulness, imagine X to be the number of fruits in a mango tree in a desert. Most likely X is zero, but if it is above a moderate threshold, we guess that unlikely rains must have occurred and hence X is likely to be very large. That means that $\mathbf{E}[X\,\big|\,X\geq t]\gg t$.

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[1] $Y_n\sim(1-\frac{1}{n})\delta_0+\frac{1}{n}\delta_{n^2}$, then $\mathbf{E}[Y_n]\to\infty$ but $\mathbf{P}\{Y_n>0\}=\frac{1}{n^2}\to 0$.

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

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

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 $\mathrm{Var}(X)<(1-\delta)^2\mathbf{E}[X]^2$, then we get a lower bound for the probability that $X\geq$

[1] The measure δ_x puts mass 1 at the point x, hence $\mathbf{P}\{Y_n>0\}=\frac{1}{n^2}\to 0$.

$\delta\mathbf{E}[X]$. 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 non-trivial lower bound even if we only know that $\text{Var}(X) \le 100\mathbf{E}[X]^2$.

Lemma 4: Second moment method or Paley-Zygmund inequality

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

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

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

PROOF. $\mathbf{E}[X]^2 = \mathbf{E}[X\mathbf{1}_{X>0}]^2 \le \mathbf{E}[X^2]\mathbf{E}[\mathbf{1}_{X>0}] = \mathbf{E}[X^2]\mathbf{P}\{X>0\}$. Hence the second inequality follows. The first one is similar. Let $\mu = \mathbf{E}[X]$. By Cauchy-Schwarz,

$$\mathbf{E}[X\mathbf{1}_{X>\alpha\mu}]^2 \le \mathbf{E}[X^2]\mathbf{P}\{X>\alpha\mu\}.$$

Further, $\mu = \mathbf{E}[X\mathbf{1}_{X<\alpha\mu}]+\mathbf{E}[X\mathbf{1}_{X>\alpha\mu}] \le \alpha\mu + \mathbf{E}[X\mathbf{1}_{X>\alpha\mu}]$, whence, $\mathbf{E}[X\mathbf{1}_{X>\alpha\mu}] \ge (1-\alpha)\mu$. Thus,

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

The remaining conclusions follow easily. ■

Remark 2

Alternately, the first inequality can be derived by applying the second one to $Y = (X-\alpha\mu)_+$, as (1) $\mathbf{P}\{Y>0\} = \mathbf{P}\{X>\alpha\mu\}$, (2) $\mathbf{E}[Y] \ge \mathbf{E}[X-\alpha\mu] = (1-\alpha)\mu$ and (3) $\mathbf{E}[Y^2] \le \mathbf{E}[X^2]$.

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!

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 any random variable measurable with respect to such a sigma algebra is an almost sure constant.

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. Or that $\limsup X_n = 1$, etc.

Theorem 6: Kolmogorov's zero-one law

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{F}_n$ be independent sub sigma algebras. Then the tail sigma-algebra $\mathcal{T}$ is trivial.

PROOF. 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. ■

Corollary 7

If $X_1, X_2, \ldots$ are independent random variables, and Y is another random variables such that Y is a function of $(X_n, X_{n+1}, \ldots)$ for any n, then Y is a constant a.s.

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 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 3: Reformulation in terms of product measures

Let $(\Omega_k, \mathcal{F}_k, \mu_k)$ be probability spaces and consider $(\Omega = \times_i \Omega_i, \mathcal{F} = \otimes_i \mathcal{F}_i, \mu = \otimes_i \mu_i)$. The tail sigma-algebra of the sequence $\mathcal{G}_k = \sigma\{\Pi_k, \Pi_{k+1}, \ldots\}$ is trivial.

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 on the product space 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 8: 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 $\mu(A) = 1$. ■

Remark 4: Reformulation in terms of sequences of random variables

Let $X_1, X_2, \ldots$ be a sequence of random variables on a common probability space such that $(X_k, X_{k+1}, \ldots)$ has the same distribution as $(X_1, X_2, \ldots)$ for any k. Let Y be another random variables such that $Y = F(X_k, X_{k+1}, \ldots)$ for any $k \geq 1$ for some $F : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$. Then Y is an almost sure constant.

It is often more natural to consider the invariant sigma-algebra on the 2-sided infinite product $\Omega^{\mathbb{Z}}$ with shifts being defined in the obvious way. Under any i.i.d. product measure, the invariant sigma-algebra is trivial.

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 9: Hoeffding's inequality

Let $X_1,\ldots,X_n$ be independent random variables having zero mean. Assume that $|X_k| \le a_k$ $a.s.$ 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 $\mathbf{P}\left\{S \ge tA\right\} \le e^{-\frac{1}{2}t^2}$ 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 $\mathbf{P}\{|S| \ge tA\} \le 2e^{-t^2/2}$.

(2) If $|X_k| \le a_k$ 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 $|X_k| \le a_k$, we also have $|\mathbf{E}[X_k]| \le a_k$ and hence $|Y_k| \le 2a_k$. This gives a conclusion that is slightly weaker but qualitatively no different: With $S = X_1+\ldots+X_n$,
$$\mathbf{P}\left\{S - \mathbf{E}[S] \ge t\sqrt{a_1^2+\ldots+a_n^2}\right\} \le e^{-\frac{1}{8}t^2}.$$

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

$$\mathbf{P}\{S \ge tA\} = \mathbf{P}\{e^{\theta S} \ge e^{\theta t A}\} \le e^{-\theta t A}\mathbf{E}[e^{\theta S}] = e^{-\theta t A}\mathbf{E}\left[\prod_{k=1}^{n} e^{\theta X_k}\right]. \tag{3}$$

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 $e^{\theta X_k} \le \alpha_k + \beta_k X_k$, 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 (3) to get

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

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 $\alpha_k \le e^{\theta^2 a_k^2/2}$. This follows from the following observation:

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

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.

7. 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, $\mathbf{P}\left\{\max_{k \le n} |S_k| > t\right\} \le t^{-2} \sum_{k=1}^{n} \text{Var}(X_k)$.

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. 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,

$$\mathbf{P}(\max_{k\le n} |S_k| > t) = \mathbf{P}(|S_\tau| > t) \le t^{-2}\mathbf{E}[S_\tau^2]. \tag{4}$$

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

$$\begin{aligned} \mathbf{E}[S_n^2] &= \mathbf{E}\left[(S_\tau + (S_n - S_\tau))^2\right] \\ &= \mathbf{E}[S_\tau^2] + \mathbf{E}\left[(S_n - S_\tau)^2\right] + 2\mathbf{E}[S_\tau(S_n - S_\tau)] \\ &\ge \mathbf{E}[S_\tau^2] + 2\mathbf{E}[S_\tau(S_n - S_\tau)]. \qquad (5) \end{aligned}$$

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

Remark 5

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 $\mathbf{E}[X_k \mid X_1, \ldots, X_{k-1}] = 0$.

8. 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. 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". Here are some simpler general situations where the method is useful.

Proving inequalities between numbers by coupling: Suppose we wish to show that $a \leq b$. If we could find random variables X, Y on a common probability space such that $X \leq Y$ a.s., and $\mathbf{E}[X] = a$ and $\mathbf{E}[Y] = b$, then the inequality would follow. If the numbers are in $[0,1]$, this may be be possible to prove by finding events $A \subseteq B$ such that $\mathbf{P}(A) = a$ and $\mathbf{P}(B) = b$. What is called the *probabilistic method* is of this kind: We show that a set A (described in some way), is non-empty by showing that $\mathbf{P}(A) > 0$ under some probability measure $\mathbf{P}$.

Example 6

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 $Y' = \sum_k \mathbf{1}_{U_k \leq 1/2}$. 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 $X \geq Y$ $a.s.$

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 $|X - Y| \leq r$ with probability at least $1 - r$. Then we can claim that $d(\mu, \nu) \leq r$. Indeed,

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

and similarly $F_\mu(t) \geq F_\nu(t - r) - r$. It is a fact that if $d(\mu, \nu) = r$, then such a coupled pair of random variables does exist but it requires a bit of work (it is akin to Hall's marriage problem), so we skip it.

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 $d_{TV}(\mu, \nu) = \max\limits_{A \subseteq [n]} |\mu(A) - \nu(A)|$. 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 an easy fact that one can always couple random variables this way.

Exercise 4

Show that there is a coupling (X, Y) that achieves equality, i.e., $\mathbf{P}\{X \neq Y\} = d_{TV}(\mu, \nu)$.

Defining distances using coupling: The fact that Lévy distance and total variation distance can be rephrased in terms of coupling suggests that one can define other distances between probability measures by minimizing some cost over all possible couplings. The following is a very useful definition (we shall not use it in this course though).

Definition 8: Transportation distance

Let μ and ν be two measures on $\mathbb{R}^d$. For $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, define $T_c(\mu, \nu) := \inf\{\mathbf{E}[c(X, Y)] : X \sim \mu,\ Y \sim \nu\}$, where the infimum is over all couplings with the given marginals (and one can choose the probability space too).

Popular choices of the cost function are $c(x, y) = \|x - y\|$ (Euclidean distance) and $c(x, y) = \|x - y\|^2$. In the latter case, the transportation distance is widely referred to as *Wasserstein metric*, although it has been well-argued that it should be called *Kantorovich metric*.

CHAPTER 3

Applications of the tools

We illustrate the use of the tools introduced in the previous chapter. Simultaneously, this is an excuse to showcase a few probability situations of interest on their own. Further, coupon collector problem, branching processes, random walks, etc., are not only interesting on their own, they also appear embedded within various other problems. A good understanding of probability requires one to know these well.

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,

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

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

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

which converges to 1 as $m \to \infty$, because of the assumption that $\sum \mathbf{P}(A_k) = \infty$. This shows that $\mathbf{P}(\cup_{k\ge n} A_k) = 1$ 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!

2. Coupon collector problem

A bookshelf has (a large number) n books numbered $1, 2, \dots, n$. 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

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

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 $\log(1-x) = -x - x^2/2 - x^3/3 - \dots$ which is valid for $|x| < 1$. Hence, if $|x| < \frac{1}{2}$, then

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

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

PROOF OF THEOREM 11. Fix an integer $t \ge 1$ and let $X_{t,k}$ be the indicator that the k^{th} book is not picked up on the first t days. Then, $\mathbf{P}(T_n > t) = \mathbf{P}(S_{t,n} \ge 1)$ where $S_{t,n} = X_{t,1} + \dots + X_{t,n}$ 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 $k \ne \ell$, we also compute that thefirst two moments of $S_{t,n}$ and use (**??**) to get

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

and

$$\mathbf{E}[S_{t,n}^2] = n\left(1-\frac{1}{n}\right)^t + n(n-1)\left(1-\frac{2}{n}\right)^t \le ne^{-\frac{t}{n}} + n(n-1)e^{-\frac{2t}{n}}. \tag{7}$$

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

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

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

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

$$\mathbf{P}(T_n > n\log n - n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \geq \frac{\mathbf{E}[S_{t,n}]^2}{\mathbf{E}[S_{t,n}^2]} \geq \frac{e^{2\theta_n - 2\frac{\log n - \theta_n}{n}}}{e^{\theta_n} + e^{2\theta_n}} = 1 - o(1) \tag{9}$$

as $n \to \infty$. From (8) and (9), 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 $\xi \sim$ 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 Var$(\xi) = \frac{1-p}{p^2}$, by direct calculations. Therefore, remembering that $1 + \frac{1}{2} + \ldots + \frac{1}{n} = \log n + O(1)$, we get

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

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

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. $\blacksquare$

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[1].

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) 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 $\mathbf{P}(\text{extinction}) = 1$.

[1] 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 $m = \mathbf{E}[\xi] > 1$, 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 beyond the first A, a brutal form of family planning. If inspite of the violence, the population survives, then the original must also survive...) ■

The proof does not cover the critical case which may be skipped on first reading.

The critical case $m = 1$: This case is a little more delicate as $\mathbf{E}[Z_n] = 1$ stays constant. Here the strengthened form of Markov's inequality (2) comes in handy. The intuitive explanation why it can help is that if there is one survivor in the nth generation, then it is likely that there are many survivors. For simplicity we give a not entirely rigorous argument in a particular example.

A HEURISTIC PROOF OF EXTINCTION IN THE CRITICAL CASE FOR BINARY BRANCHING. Assume that $p_0 = p_2 = \frac{1}{2}$. Then $m = 1$. If $Z_n \geq 1$, pick an individual in the nth generation (this is where the argument is loose - one needs to specify how this individual is picked). Call this individual v_n and let her ancestors be $v_{n-1}, v_{n-2}, \ldots, v_0$ (where v_k belongs to the kth generation). Let M_k be the number of descendents of v_k that are alive in generation n, excluding those that are also descendents of v_{k+1}. Then,

$$Z_n = 1 + M_{n-1} + \ldots + M_0.$$

We claim that $\mathbf{E}[M_k] = 1$. Indeed, as v_k has at least one offspring (i.e., v_{k+1}), she must have exactly one more off-spring, call it v'_{k+1}. Then M_k is exactly the number of descendents of v'_{k+1} who are in the nth generation of the original process (which is the $n - k - 1$st generation of the tree under v'_{k+1}). But as the branching is critical, $\mathbf{E}[M_k] = 1$. This shows that $\mathbf{E}[Z_n \mid Z_n \geq 1] = n + 1$ and consequently, by the strengthening of Markov's inequality given above,

$$\mathbf{P}\{Z_n \geq 1\} \leq \frac{\mathbf{E}[Z_n]}{\mathbf{E}[Z_n \,\big|\, Z_n \geq 1]} = \frac{1}{n+1}$$

which converges to 0. ■

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

$$\mu_n\left\{k \in [n] : \left|\frac{\nu(k)}{\log\log n} - 1\right| > \delta\right\} \to 0. \tag{10}$$

PROOF. (**Turan**). Fix n and for any prime p define $X_p : [n] \to \mathbb{R}$ by $X_p(k) = \mathbf{1}_{p|k}$. Then, $\nu(k) = \sum\limits_{p \leq k} X_p(k)$. We define $\psi(k) := \sum\limits_{p \leq \sqrt[4]{k}} X_p(k)$. 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 (10) 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

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

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 5 below)). Thus, we get $\mathbf{E}_n[\psi] = \log\log n + O(n^{-\frac{3}{4}})$ and $\text{Var}_n[\psi] = \log\log n + O(1)$. Thus, by Chebyshev's inequality,

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

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

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

Exercise 5

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

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 9: 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, $Y_iY_j = 1$ 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,

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

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

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

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

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

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

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. ■

7. 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.

8. 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 $f_n(t) = \sin(\pi n t)/n$ 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 $\sum_n x_n \frac{\sin(\pi n t)}{n}$ converges uniformly but gives a nowhere differentiable function!

9. 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, $|X_n|^{\frac{1}{n}} \le 1 + \varepsilon$ $a.s.$. and hence $R \le 1 + \varepsilon$ $a.s.$ Take intersection over rational ε to get $R \le 1$ $a.s.$. 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 $|X_n|^{\frac{1}{n}} \ge 1 - \varepsilon$. Again, take intersection over rational ε to conclude that $R \ge 1$ $a.s.$ 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 $\sum_{n=0}^{\infty} c_n X_n z^n$.

10. Growth of a supercritical branching process

We showed that a super-critical branching process survives with strictly positive probability. One can ask how the generation sizes Z_n grow when the branching is supercritical. An important theorem of Kesten and Stigum asserts that under the extra condition that $\mathbf{E}[L \log_+ L] < \infty$, the generation sizes grow exponentially in the sense that

$$\mathbf{P}\left\{\limsup \frac{Z_n}{m^n} > 0\right\} = \mathbf{P}\{\text{non-extinction}\}.$$

Actually it says that with $\lim Z_n/m^n$ in place of $\limsup$ (the existence of the limit must be proved, of course), but we stick to the above form. Obviously the event on the left is contained in the event on the right, hence the asserion is really that whenever non-extinction occurs, it occurs by the Z_n grown exponentially fast.

We prove a very special case of this, as the main goal here is to illustrate the tools introduced in the previous chapter. Recall that the off-spring variable L has distribution $p_k = \mathbf{P}\{L = k\}$ and $m = \sum_k kp_k$ is its mean.

Theorem 16: Growth of supercritical branching process

Assume that $p_0 = 0$ and $m > 1$ and that $\sigma^2 := \text{Var}(L) < \infty$. Then, $\limsup m^{-n} Z_n > 0$ a.s.

PROOF. Under the assumption that $p_0 = 0$, extinction never occurs. Further, if

Let $W_n = Z_n/m^n$ and let $W = \limsup W_n$. Also recall the way we constructed a branching process from i.i.d. random variables $L_{n,k}$, $n, k \geq 1$ by using $L_{n,1}, L_{n,2} \ldots$ to determine the numbers of offsprings of those individuals in the $(n-1)$st generation.

First we claim that $\mathbf{P}\{W > 0\} > 0$.

The same proof that we used (second moment method) to show that non-extinction has strictly positive probability in fact shows that

$$\liminf \mathbf{P}\left\{Z_n \geq \frac{1}{2}m^n\right\} \geq \frac{1}{4 + \frac{4\sigma^2}{m-1}}.$$

Now let $W = \limsup Z_n/m^n$ and let NE be the event of non-extinction. Clearly $\{W > 0\} \subseteq \text{NE}$. What we need to show is that $\mathbf{P}\{W > 0\} = \mathbf{P}\{\text{NE}\}$, which then implies that $\mathbf{P}\{\{W > 0\} \cap \text{NE}\} = 0$ as claimed.

First we claim that $\mathbf{P}\{W > 0\} > 0$. As $\{W < \varepsilon\} \subseteq \cup_N \cap_{n \geq N} \{Z_n < \varepsilon m^n\}$, it follows that if $\mathbf{P}\{W > 0\} = 0$, then for any $\varepsilon > 0$, there is some $N < \infty$ such that $\mathbf{P}\{Z_n > \varepsilon m^n \text{ for some } n \geq N\} < \varepsilon$. ■

11. 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[2], $\mathbf{P}_p(A)$ 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 $p \le \frac{1}{2}$ 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; zero-one laws may tell us that the probability of an event is 0 or 1, but deciding between these two possibilities can be very difficult!

12. Random walk

Let X_i be i.i.d. $\text{Ber}_{\pm}(1/2)$ and let $S_n = X_1 + \ldots + X_n$ for $n \ge 1$ 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 $k \ge 2$. 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).

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

CHAPTER 4

Modes of convergence

1. A metric on the space of probability measures on $\mathbb{R}^d$

What kind of space is $\mathcal{P}(\mathbb{R}^d)$, the space of Borel on $\mathbb{R}^d$? It is clearly a convex set (this is true for the space of probability measures on any measurable space). We want to measure closeness of two probability distributions. Two possible definitions come to mind.

(1) For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, define $D_1(\mu,\nu) := \sup_{A \in \mathcal{B}_d} |\mu(A) - \nu(A)|$. Since μ and ν are functions on the Borel σ-algebra, this is just their supremum distance, usually called the *total variation distance*. It is easy to see that D_1 is indeed a metric on $\mathcal{P}(\mathbb{R}^d)$.

One shortcoming of this metric is that D_1 is too strong. If μ is a discrete measure and ν is a measure with density, then $D_1(\mu,\nu) = 1$. But if μ is uniform distribution on $[0,1]$ and μ_n is uniform distribution on the finite set $\{j/n : 1 \le j \le n\}$, then for large n we would like to think that μ and μ_n are close (after all, if we want a sample from μ, a random number generator will in fact give us a sample from ν for some large n, and we accept that). But in the metric D_1, they remain far apart.

(2) We can restrict the class of sets over which we take the supremum. For instance, taking all semi-infinite intervals, we define the *Kolmogorov-Smirnov* distance

$$D_2(\mu,\nu) = \sup_{x \in \mathbb{R}^d} |F_\mu(x) - F_\nu(x)|.$$

If two CDFs are equal, the corresponding measures are equal. Hence D_2 is also a genuine metric on $\mathcal{P}(\mathbb{R}^d)$.

Clearly $D_2(\mu,\nu) \le D_1(\mu,\nu)$, hence D_2 is weaker than D_1. Unlike with D_1, it is possible to have discrete measures converging in D_2 to a continuous one, see Exercise 6. But it is still too strong.

For example, if $a \neq b$ are points in $\mathbb{R}^n$, then it is easy to see that $D_1(\delta_a,\delta_b) = D_2(\delta_a,\delta_b) = 1$. Thus, even when $a_n \to a$ in $\mathbb{R}^d$, we do not get convergence of δ_{a_n} to δ_a in these metrics. This is an undesirable feature as we must accept errors in measurement, for example, a 10 digit number as an approximation to a real number. Alternately, let us just say that we would like the embedding $\mathbb{R} \mapsto \mathcal{P}(\mathbb{R})$ defined by $a \mapsto \delta_a$ to be continuous.

Thus, we would like a weaker metric, where more sequences converge. The problem with the earlier two definitions is that they compare closeness of $\mu(A)$ with $\nu(A)$. But we must allow for finite precision of measurement, meaning that we cannot be too sure if a number belongs to A or is close to it. The next definition allows for this imprecision.

Definition 10

For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, define the *Lévy distance* between them as (here $\mathbf{1} = (1, 1, \ldots, 1)$)

$$d(\mu, \nu) := \inf\{u > 0 : F_\mu(x + u\mathbf{1}) + u \geq F_\nu(x),\ F_\nu(x + u\mathbf{1}) + u \geq F_\mu(x)\ \forall x \in \mathbb{R}^d\}.$$

If $d(\mu_n, \mu) \to 0$, we say that μ_n converges in distribution or converges weakly to μ and write $\mu_n \overset{d}{\to} \mu$. [...breathe slowly and meditate on this definition for a few minutes...]

Remark 6

Although we shall not use it, in the same way one can define a metric on $\mathcal{P}(X)$ for a metric space X (it is called *Lévy-Prohorov distance*). For $\mu, \nu \in \mathcal{P}(X)$

$$d(\mu, \nu) := \inf\{t > 0 : \mu(A^{(t)}) + t \geq \nu(A) \text{ and } \nu(A^{(t)}) + t \geq \mu(A) \text{ for all closed } A \subseteq X\}.$$

Here $A^{(t)}$ is the set of all points in X that are within distance t of A. This makes it clear that we do not directly compare the measures of a given set, but if $d(\mu, \nu) < t$, it means that whenever μ gives a certain measure to a set, then ν should give nearly that much (nearly means, allow t amount less) measure to a t-neighbourhood of A.

As an example, if $a, b \in \mathbb{R}^d$, then check that $d(\delta_a, \delta_b) \leq (\max_i |b_i - a_i|) \wedge 1$. Hence, if $a_n \to a$, then $d(\delta_{a_n}, \delta_a) \to 0$. Recall that δ_{a_n} does not converge to δ_a in D_1 or D_2.

Exercise 6

Let $\mu_n = \frac{1}{n}\sum_{k=1}^n \delta_{k/n}$. Show directly by definition that $d(\mu_n, \lambda) \to 0$. Show also that $D_2(\mu_n, \lambda) \to 0$ but $D_1(\mu_n, \lambda)$ does not go to 0.

The definition is rather unwieldy in checking convergence. The following proposition gives the criterion for convergence in distribution in terms of distribution functions.

Proposition 17

$\mu_n \overset{d}{\to} \mu$ if and only if $F_{\mu_n}(x) \to F_\mu(x)$ for all continuity points x of F_μ.

PROOF. Suppose $\mu_n \overset{d}{\to} \mu$. Let $x \in \mathbb{R}^d$ and fix $u > 0$. Then for large enough n, we have $F_\mu(x + u\mathbf{1}) + u \geq F_{\mu_n}(x)$, hence $\limsup F_{\mu_n}(x) \leq F_\mu(x + u\mathbf{1}) + u$ for all $u > 0$. By right continuity of F_μ, we get $\limsup F_{\mu_n}(x) \leq F_\mu(x)$. Further, $F_{\mu_n}(x) + u \geq F_\mu(x - u\mathbf{1})$ for large n, hence $\liminf F_{\mu_n}(x) \geq$

$F_\mu(x-u)$ for all u. If x is a continuity point of F_μ, we can let $u \to 0$ and get $\liminf F_{\mu_n}(x) \geq F_\mu(x)$. Thus $F_{\mu_n}(x) \to F_\mu(x)$.

For the converse, for simplicity let $d = 1$. Suppose $F_n \to F$ at all continuity points of F. Fix any $u > 0$. Find $x_1 < x_2 < \ldots < x_m$, continuity points of F, such that $x_{i+1} \leq x_i + u$ and such that $F(x_1) < u$ and $1 - F(x_m) < u$. This can be done because continuity points are dense. Now use the hypothesis to fix N so that $|F_n(x_i) - F(x_i)| < u$ for each $i \leq m$ and for $n \geq N$. Henceforth, let $n \geq N$.

If $x \in \mathbb{R}$, then either $x \in [x_{j-1}, x_j]$ for some j or else $x < x_1$ or $x > x_1$. First suppose $x \in [x_{j-1}, x_j]$. Then

$$F(x+u) \geq F(x_j) \geq F_n(x_j) - u \geq F_n(x) - u, \qquad F_n(x+u) \geq F_n(x_j) \geq F(x_j) - u \geq F(x) - u.$$

If $x < x_1$, then $F(x+u) + u \geq u \geq F(x_1) \geq F_n(x_1) - u$. Similarly the other requisite inequalities, and we finally have

$$F_n(x+2u) + 2u \geq F(x) \text{ and } F(x+2u) + 2u \geq F_n(x).$$

Thus $d(\mu_n, \mu) \leq 2u$. Hence $d(\mu_n, \mu) \to 0$. ■

Example 7

Again, let $a_n \to a$ in $\mathbb{R}$. Then $F_{\delta_{a_n}}(t) = 1$ if $t \geq a_n$ and 0 otherwise while $F_{\delta_a}(t) = 1$ if $t \geq a$ and 0 otherwise. Thus, $F_{\delta_{a_n}}(t) \to F_{\delta_a}(t)$ for all $t \neq a$ (just consider the two cases $t < a$ and $t > a$). This example also shows the need for excluding discontinuity points of the limiting distribution function. Indeed, $F_{\delta_{a_n}}(a) = 0$ (if $a_n \neq a$) but $F_{\delta_a}(a) = 1$.

Observe how much easier it is to check the condition in the theorem rather than the original definition! Many books use the convergence at all continuity points of the limit CDF as the definition of convergence in distribution. But we defined it via the Lévy metric because we are familiar with convergence in metric spaces and this definition shows that convergence in distribution in not anything more exotic. On the other hand, giving the metric first is also misleading unless one understands that there are several alternate definitions that we could have given (see exercise at the end of the section), all of which give the same topology on $\mathcal{P}(\mathbb{R})$. The point to keep in mind is that the topology, however you define it, is *metrizable*.

Exercise 7

If $a_n \to 0$ and $b_n^2 \to 1$, show that $N(a_n, b_n^2) \xrightarrow{d} N(0,1)$ (recall that $N(a, b^2)$ is the Normal distribution with parameters $a \in \mathbb{R}$ and $b^2 > 0$).

Question: In class, Milind Hegde raised the following question. If we define (write in one dimension for notational simplicity)

$$d'(\mu,\nu) = \inf\{t > 0 : F_\mu(x+t) \geq F_\nu(x) \text{ and } F_\nu(x+t) \geq F_\mu(x) \text{ for all } x\},$$

how different is the resulting metric from the Lévy metric? In other words, is it necessary to allow an extra additive t to $F_\mu(x+t)$?

It does make a difference! Suppose μ,ν are two probability measures on $\mathbb{R}$ such that $\mu(K_0) = 1$ for some compact set K_0 and $\nu(K) < 1$ for all compact sets K. Then, if x is large enough so that $x > y$ for all $y \in K_0$, then $F_\nu(x+t) < 1 = F_\mu(x)$ for any $t > 0$. Hence, $d'(\mu,\nu) > t$ for any t implying that $d'(\mu,\nu) = \infty$.

Now, it is not a serious problem if a metric takes the value ∞. We can replace d' by $d''(\mu,\nu) = d'(\mu,\nu) \wedge 1$ or $d'''(\mu,\nu) = d(\mu,\nu)/(1 + d(\mu,\nu))$ which gives metrics that are finite everywhere but are such that convergent sequences are the same as in d' (i.e., $d'(\mu_n,\mu) \to 0$ if and only if $d''(\mu_n,\mu) \to 0$).

But the issue is that measures with compact support can never converge to a measure without compact support. For example, if X has exponential distribution and $X_k = X \wedge k$, then the distribution of X_k does not converge to the distribution of X in the metric d'. However, it is indeed the case that the convergence happens in the metric d. Thus the two metrics are not equivalent [1].

Here are other ways to have defined the Lévy metric. There is no natural way to choose between these definitions, underlining the point made earlier that the value of the Lévy distance is itself of no great significance, what matters is the topology, or which sequences converge to which measure. In fact, the Kolmogorov-Smirnov and total variation distances are more meaningful (and actually used!) when one really wants to measure distances.

[1] In class I wrongly claimed that for probability measures on a compact set in place of the whole real line, eg., $\mathcal{P}([-1,1])$, convergence in d' and in d are equivalent. Chirag Igoor showed me the following counter-example. Let $\mu = \delta_1$ and for each n define

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/n & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

Then, $F_n(x) \to F_\mu(x)$ for each x and hence the corresponding measures converge to μ in Lévy metric. But the convergence fails in d'. To see this, take any $x > 0$ and observe that if $F_\mu(0.5+t) \geq F_{\mu_n}(0.5)$, then we must have $t \geq 0.5$. As this is true for every n, it follows that μ_n does not converge to μ in d'. Another such example is $\mu_n = (1-n^{-1})\delta_0 + n^{-1}\delta_1$ and $\mu = \delta_0$.

Exercise 8

Show that each of the following is a metric that is equivalent to the Lévy metric (in the sense that $\mu_n \to \mu$ in one metric if and only if in the others).

(1) $\inf\{u > 0 : F_\mu(x + au\mathbf{1}) + bu \geq F_\nu(x),\ F_\nu(x + au\mathbf{1}) + bu \geq F_\mu(x)\ \forall x \in \mathbb{R}^d\}$ where $a, b > 0$ are fixed.

(2) $\inf\{u + v : u, v > 0$ and $F_\mu(x + u\mathbf{1}) + v \geq F_\nu(x),\ F_\nu(x + u\mathbf{1}) + v \geq F_\mu(x)\ \forall x \in \mathbb{R}^d\}$.

Equivalent forms of convergence in distribution. We have given two equivalent definitions of convergence in distribution. There are several others.

Theorem 18

Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R}^d)$. The following statements are equivalent.

(1) $\mu_n \overset{d}{\to} \mu$.

(2) $F_{\mu_n}(x) \to F_\mu(x)$ for all x where F_μ is continuous.

(3) $\liminf\limits_{n\to\infty} \mu_n(G) \geq \mu(G)$ for all open $G \subseteq \mathbb{R}^d$.

(4) $\limsup\limits_{n\to\infty} \mu_n(C) \leq \mu(C)$ for all closed $C \subseteq \mathbb{R}^d$.

(5) $\int f d\mu_n \to \int f d\mu$ for all bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$.

We have proved the equivalence of (1) and (2). It is also clear that (3) and (4) are equivalent (just take complements). Hence it suffices to show that $(2) \implies (3) \implies (5) \implies (2)$. For simplicity, we present the proof in one-dimension.

PROOF FOR $d = 1$. Assume (2). Let $G \subseteq \mathbb{R}$ be an open set. Then write it as $G = \sqcup_k (a_k, b_k)$. Choose intervals $(a_k', b_k') \subseteq (a_k, b_k)$ such that a_k', b_k' are continuity points of F_μ and $\mu(a_k', b_k') \geq \mu(a_k, b_k) - \varepsilon 2^{-k}$ (possible as there are at most countably many discontinuity points). Then

$$\mu_n(a_k, b_k) \geq F_{\mu_n}(b_k') - F_{\mu_n}(a_k') \to F_\mu(b_k') - F_\mu(a_k') = \mu(a_k', b_k').$$

Hence $\liminf \mu_n(a_k, b_k) \geq \mu(a_k, b_k) - \varepsilon 2^{-k}$. By Fatou's lemma applied to sums, we see that

$$\liminf \sum_k \mu_n(a_k, b_k) \ \geq\ \sum_k \mu(a_k, b_k) - \varepsilon 2^{-k} \ \geq\ \mu(G) - \varepsilon.$$

The left side is $\liminf \mu_n(G)$ and $\varepsilon > 0$ is arbitrary, hence $\liminf \mu_n(G) \geq \mu(G)$. This proves (3).

Assume (3) holds. Let $f \in C_b(\mathbb{R})$. Then $\{f > t\}$ is an open set for any $t \in \mathbb{R}$ and hence $\liminf \mu_n\{f > t\} \geq \mu\{f > t\}$ by assumption. By Fatou's lemma,

$$\liminf \int_0^\infty \mu_n\{f > t\}dt \geq \int_0^\infty \mu\{f > t\}dt.$$

If $f \geq 0$, then this is the same as saying $\liminf \int f d\mu_n \geq \int f d\mu$. For general bounded continuous f with $M = \|f\|_{\text{sup}}$, apply this to the positive functions $M - f$ and $M + f$ to conclude that $\int f d\mu_n \to \int f d\mu$.

Assume (5) holds. If $x < y$, let $\varphi_{x,y} : \mathbb{R} \to [0, 1]$ be a continuous function such that $\varphi_{x,y}(u) = 1$ for $u \leq x$ and $\varphi_{x,y}(u) = 0$ for $u \geq y$. Then

$$F_{\mu_n}(x) \leq \int \varphi_{x,y} d\mu_n \leq F_{\mu_n}(y), \qquad F_\mu(x) \leq \int \varphi_{x,y} d\mu \leq F_\mu(y).$$

As $\int \varphi_{x,y} d\mu_n \to \int \varphi_{x,y} d\mu$ by assumption, we see that

$$\limsup F_{\mu_n}(x) \leq F_\mu(y), \qquad \liminf F_{\mu_n}(y) \geq F_\mu(x).$$

This is true for all $x < y$. Let $y \downarrow x$ in the first inequality to get $\limsup F_{\mu_n}(x) \leq F_\mu(x)$ for all x. Let $x \uparrow y$ in the second inequality to get $\liminf F_{\mu_n}(y) \geq F_\mu(y-)$ for all y. Hence if x is a continuity point of F_μ, we have $\lim F_{\mu_n}(x) = F_\mu(x)$. ■

As we have seen, $\mu_n \overset{d}{\to} \mu$ does not imply that $\mu_n(A) \to \mu(A)$ in general. Sometimes it does, for example if $A = (-\infty, x]$ where $\mu\{x\} = 0$. Here is a generalization.

Exercise 9

Let $A \in \mathcal{B}(\mathbb{R})$. If $\mu_n \overset{d}{\to} \mu$ and $\mu(\partial A) = 0$, then show that $\mu_n(A) \to \mu(A)$.

Remark 7

The dual of $C_c(\mathbb{R})$ is the space of all signed measures on $\mathbb{R}$ with finite total variation. These are basically of the form $\theta = \mu - \nu$ where μ, ν are mutually singular positive measures and θ acts on f by $f \mapsto \int f d\mu - \int f d\nu$. The dual norm is $\|\theta\| = \mu(\mathbb{R}) + \nu(\mathbb{R})$. Convergence in weak-* sense in the dual space is defined by $\theta_n \to \theta$ if $\theta_n(f) \to \theta(f)$ for all f (i.e., pointwise convergence of linear functionals), though we are being a little loose in talking in terms of sequences (the dual with weak-* topology is generally not a metric space). That is essentially the definition of weak convergence of probability measures (point (5) in the theorem proved above), except that in this sense probability measures can converge to a sub-probability measure. For example, $0.5\delta_0 + 0.5\delta_n \to 0.5\delta_0$. But if we ask for $\theta_n(f) \to \theta(f)$ for all $f \in C_b(\mathbb{R})$, a larger space, then this leakage of mass to infinity cannot happen. Modulo this point, convergence in distribution is just weak-* convergence.

2. Ways to prove convergence in distribution

We end the chapter by outlining different ways in which to prove convergence in distribution. Suppose we need to show that $\mu_n \overset{d}{\to} \mu$.

(1) The most elegant of all ways is to find random variables X_n, X on some probability space such that $X_n \sim \mu_n$ and $X \sim \mu$ and $X_n \overset{a.s.}{\to} X$. This will follow from later sections in this chapter.

In fact, Skorohod's principle tells us that this can always be done, although it is not always clear how to find such random variables.

(2) Go by the book and show that $\int f d\mu_n \to \int f d\mu$ for all $f \in C_b(\mathbb{R})$ or any of the other equivalent conditions that were mentioned before. In practise, the smaller the class of functions for which we need to check this convergence, the better it is for us.

For example, if we know that $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$, then it suffices to show that convergence for $f \in C_c^\infty(\mathbb{R})$. To see this, go back to the proof of $(5) \implies (2)$ in the proof of Theorem **??**. Observe that we can choose $\varphi_{x,y}$ to be smooth, even with bounded derivatives. The rest of the proof remains the same.

(3) We shall later see that a surprisingly small class of functions suffices! Let $e_t(x) = e^{itx}$ for $t \in \mathbb{R}$. If $\int e_t d\mu_n \to \int e_t d\mu$ for all $t \in \mathbb{R}$, then $\mu_n \overset{d}{\to} \mu$. We shall prove this when we discuss characteristic functions.

3. Compact subsets in the space of probability measure on Euclidean spaces

Often we face problems like the following. A functional $L : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is given, and we would like to find the probability measure μ that minimizes $L(\mu)$. By definition, we can find nearly optimal probability measures μ_n satisfying $L(\mu_n) - \frac{1}{n} \leq \inf_\nu L(\nu)$. Then we might expect that *if* the sequence μ_n (or a subsequence of it) converged to a probability measure μ, then μ might be the optimal solution we are searching for. This motivates us to characterize compact subsets of $\mathcal{P}(\mathbb{R}^d)$, so that existence of convergent subsequences can be asserted.

Looking for a convergent subsequence: Let μ_n be a sequence in $\mathcal{P}(\mathbb{R}^d)$. We would like to see if a convergent subsequence can be extracted. Towards this direction, we prove the following lemma. We emphasize the idea of proof (a diagonal argument) which recurs in many contexts.

Lemma 19: Helly's selection principle

Let F_n be a sequence distribution functions on $\mathbb{R}^d$. Then, there exists a subsequence $\{n_\ell\}$ and a non-decreasing, right continuous functon $F : \mathbb{R}^d \to [0,1]$ such that $F_{n_\ell}(x) \to F(x)$ if x is a continuity point of F.

As before, we present the proof in one-dimension (just for notational simplicity).

PROOF. Fix a dense subset $S = \{x_1, x_2, \ldots\}$ of $\mathbb{R}$. Then, $\{F_n(x_1)\}$ is a sequence in $[0,1]$. Hence, we can find a subsequence $\{n_{1,k}\}_k$ such that $F_{n_{1,k}}(x_1)$ converges to some number $\alpha_1 \in [0,1]$. Then, extract a further subsequence $\{n_{2,k}\}_k \subseteq \{n_{1,k}\}_k$ such that $F_{n_{2,k}}(x_2) \to \alpha_2$, another number in $[0,1]$. Of course, we also have $F_{n_{2,k}}(x_1) \to \alpha_1$. Continuing this way, we get numbers $\alpha_j \in [0,1]$ and subsequences $\{n_{1,k}\} \supset \{n_{2,k}\} \supset \ldots \{n_{\ell,k}\} \ldots$ such that for each ℓ, as $k \to \infty$, we have $F_{n_{\ell,k}}(x_j) \to \alpha_j$ for each $j \leq \ell$.

The *diagonal subsequence* $\{n_{\ell,\ell}\}$ is ultimately the subsequence of each of the above obtained subsequences and therefore, $F_{n_{\ell,\ell}}(x_j) \to \alpha_j$ as $\ell \to \infty$, for each j. Henceforth, write n_ℓ instead of $n_{\ell,\ell}$.

To get a function on the whole line, set $F(x) := \inf\{\alpha_j : j \text{ for which } x_j > x\}$. F is well defined, takes values in $[0,1]$ and is non-decreasing. It is also right-continuous, because if $y_n \downarrow y$, then for any j for which $x_j > y$, it is also true that $x_j > y_n$ for sufficiently large n. Thus $\lim_{n\to\infty} F(y_n) \leq \alpha_j$. Take infimum over all j such that $x_j > y$ to get $\lim_{n\to\infty} F(y_n) \leq F(y)$. Of course $F(y) \leq \lim F(y_n)$ as F is non-decreasing. This shows that $\lim F(y_n) = F(y)$ and hence F is right continuous.

Lastly, we claim that if y is any continuity point of F, then $F_{n_\ell}(y) \to F(y)$ as $\ell \to \infty$. To see this, fix $\delta > 0$. Find i, j such that $y - \delta < x_i < y < x_j < y + \delta$. Therefore

$$\liminf F_{n_\ell}(y) \geq \lim F_{n_\ell}(x_i) = \alpha_i \geq F(y - \delta)$$

$$\limsup F_{n_\ell}(y) \leq \lim F_{n_\ell}(x_j) = \alpha_j \leq F(y + \delta).$$

In each line, the first inequalities are by the increasing nature of CDFs, and the second inequalities are by the definition of F. Thus

$$F(y-) \leq \liminf F_{n_\ell}(y) \leq \limsup F_{n_\ell}(y) \leq F(y)$$

for all $y \in \mathbb{R}$. If $F(y-) = F(y)$, then it follows that $\lim F_{n_\ell}(y)$ exists and equals $F(y)$. ■

The Lemma does not say that F is a CDF, because in general it is not!

Example 8

Consider δ_n. Clearly $F_{\delta_n}(x) \to 0$ for all x if $n \to +\infty$ and $F_{\delta_n}(x) \to 1$ for all x if $n \to -\infty$. Even if we pass to subsequences, the limiting function is identically zero or identically one, and neither of these is a CDF of a probability measure The problem is that mass escapes to infinity. To get weak convergence to a probability measure, we need to impose a condition to avoid this sort of situation.

Definition 11

A family of probability measure $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$ is said to be *tight* if for any $\varepsilon > 0$, there is a compact set $K_\varepsilon \subseteq \mathbb{R}^d$ such that $\mu(K_\varepsilon) \geq 1 - \varepsilon$ for all $\mu \in \mathcal{A}$.

Example 9

Suppose the family has only one probability measure μ. Since $[-n, n]^d$ increase to $\mathbb{R}^d$, given $\varepsilon > 0$, for a large enough n, we have $\mu([-n, n]^d) \geq 1 - \varepsilon$. Hence $\{\mu\}$ is tight. If the family is finite, tightness is again clear.

Take $d = 1$ and let μ_n be probability measures with $F_n(x) = F(x - n)$ (where F is a fixed CDF), then $\{\mu_n\}$ is not tight. This is because given any $[-M, M]$, if n is large enough, $\mu_n([-M, M])$ can be made arbitrarily small. Similarly $\{\delta_n\}$ is not tight.

We now characterize compact subsets of $\mathcal{P}(R^d)$ in the following theorem. As $\mathcal{P}(\mathbb{R}^d)$ is a metric space, compactness is equivalent to sequential compactness and we phrase the theorem in terms of sequential compactness.

Theorem 20

Let $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$. Then, the following are equivalent.

(1) Every sequence in $\mathcal{A}$ has a convergent subsequence in $\mathcal{P}(\mathbb{R}^d)$.

(2) $\mathcal{A}$ is tight.

PROOF. Let us take $d = 1$ for simplicity of notation.

(1) Assume that $\mathcal{A}$ is tight. Then any sequence $(\mu_n)_n$ in $\mathcal{A}$ is also tight. By Lemma 19, there is a subsequence $\{n_\ell\}$ and a non-decreasing right continuous function F (taking values in $[0, 1]$) such that $F_{n_\ell}(x) \to F(x)$ for all continuity points x of F.

Fix $A > 0$ such that $\mu_{n_\ell}[-A, A] \geq 1 - \varepsilon$ and such that A is a continuity point of F. Then, $F_{n_\ell}(-A) \leq \varepsilon$ and $F_{n_\ell}(A) \geq 1 - \varepsilon$ for every n and by taking limits we see that $F(-A) \leq \varepsilon$ and $F(A) \geq 1 - \varepsilon$. Thus $F(+\infty) = 1$ and $F(-\infty) = 0$. This shows that F is a CDF and hence $F = F_\mu$ for some $\mu \in \mathcal{P}(\mathbb{R}^d)$. By Proposition 17 it also follows that $\mu_{n_\ell} \overset{d}{\to} \mu$.

(2) Assume that $\mathcal{A}$ is not tight. Then, there exists $\varepsilon > 0$ such that for any k, there is some $\mu_k \in \mathcal{A}$ such that $\mu_k([-k, k]) < 1 - 2\varepsilon$. In particular, either $F_{\mu_k}(k) \leq 1 - \varepsilon$ or/and $F_{\mu_k}(-k) \geq \varepsilon$. We claim that no subsequence of $(\mu_k)_k$ can have a convergent subsequence.

To avoid complicating the notation, let us show that the whole sequence does not converge and leave you to rewrite the same for any subsequence. There are infinitely many k for which $F_{\mu_k}(-k) \geq \varepsilon$ or there are infinitely many k for which $F_{\mu_k}(k) \geq 1-\varepsilon$. Suppose the former is true. Then, for any $x \in \mathbb{R}$, since $-k < x$ for large enough k, we see that $F_{\mu_k}(x) \geq F_{\mu_k}(-k) \geq \varepsilon$ for large enough k. This means that if F_{μ_k} converge to some F (at continuity points of F), then $F(x) \geq \varepsilon$ for all x. Thus, F cannot be a CDF and hence μ_k does not have a limit. ■

Exercise 10

Adapt this proof to higher dimensions.

4. Modes of convergence of random variables

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 12

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

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

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

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

Example 10

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

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

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

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

Lemma 21

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

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

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

PROOF. Note that analogous to the example, in general

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

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

Thus, applying Fatou's lemma we see that a.s convergence implies convergence in probability. For the second part, observe that by the first Borel Cantelli lemma, if $\sum_n \mathbf{P}(|X_n - X| > \delta) < \infty$, then $\mathbf{P}(|X_n - X| > \delta \text{ i.o}) = 0$ and hence $\limsup |X_n - X| \leq \delta \ a.s.$ 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 11

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

4.2. 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 22

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

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

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

PROOF.

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

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

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

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

If $X_n = 1 - U$ and $X = U$, then $X_n \overset{d}{\to} X$ 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 $T(U) \overset{d}{=} X$. Then define $X_n = T(1-U)$. For all n the random variable X_n has the same distribution as X and hence $X_n \overset{d}{\to} X$. But X_n does not converge in probability to X (unless X is degenerate).

Exercise 12

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

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

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

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

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

Example 11

On $([0,1], \mathcal{B}, \lambda)$, define $X_n = 2^n \mathbf{1}_{[0,1/n]}$. Then, $X_n \overset{a.s.}{\to} 0$ 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 $X_n \overset{P}{\to} X$. Write

$$\mathbf{E}[|X_n - X|] = \int_0^\infty \mathbf{P}(|X_n - X| > t)dt.$$

For each t the integrand goes to zero because $X_n \overset{P}{\to} X$. Will the integral go to zero? The example of $X_n = n\mathbf{1}_{[0,1/n]}$ and $X = 0$ (on $([0,1], \mathcal{B}, \lambda)$) shows that it need not. 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. This will

5. Uniform integrability

Definition 13: 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 $i \in I$.

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 $\mathbf{E}[|X_i|] = \mathbf{E}[|X_i|\mathbf{1}_{|X_i|<A}] + \mathbf{E}[|X_i|\mathbf{1}_{|X_i|\geq A}] \leq A + 1$.

The converse is not true, as the example of $X_n = n\mathbf{1}_{[0,\frac{1}{n}]}$ on $([0,1],\mathcal{B},\lambda)$ shows. In this case, for any A, if n is large enough, then $\mathbf{E}[|X_n|\mathbf{1}_{|X_n|>A}] = 1$, hence the family is not uniformly integrable. However, this just misses uniform integrability.

Example 12

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 $\sup_{i\in I}\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>t}] < \varepsilon$.

Exercise 13

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 23

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

(1) $X_n \overset{L^1}{\to} X$.

(2) $X_n \overset{P}{\to} X$ and $\{X_n\}$ is u.i.

Proof. If $Y_n = X_n - X$, then $X_n \overset{L^1}{\to} X$ iff $Y_n \overset{L^1}{\to} 0$, while $X_n \overset{P}{\to} X$ iff $Y_n \overset{P}{\to} 0$ and by the first part of exercise 13, $\{X_n\}$ 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 $Y_n \overset{L^1}{\to} 0$. We already showed that $Y_n \overset{P}{\to} 0$. 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 $Y_n \overset{L^1}{\to} 0$.

Next suppose $Y_n \overset{P}{\to} 0$ 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 $Y_n \overset{P}{\to} 0$ 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 $Y_k \overset{L^1}{\to} 0$. ■

Corollary 24

Suppose X_n, X are integrable random variables and $X_n \overset{a.s.}{\to} X$. Then, $X_n \overset{L^1}{\to} X$ 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 23 and corollary 24, 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.

5.1. Relationship to compactness. The definition of uniform integrability is reminiscent of the definition of tightness. In fact, it can be recast in that fashion.

Exercise 14

Given random variables X_i, $i \in I$ on $(\Omega, \mathcal{F}, \mathbf{P})$, define the measures $\mu_i(A) = \int_A |X_i| d\mathbf{P}$ and let $\nu_i = \mu_i \circ X_i^{-1}$ be the push-forward measure on $\mathbb{R}$. Show that $\{X_i : i \in I\}$ is uniformly integrable if and only if the measures $\{\nu_i : i \in I\}$ is tight.[a]

[a] We defined tightness for probability measures. Here we obviously mean that given $\varepsilon > 0$ there is some M such that $\nu_i([-M, M]^c) < \varepsilon$ for all $i \in I$.

Tightness is the criterion for precompactness in the space of probability measures. Similarly, uniform integrability is also related to a compactness question.

To explain this, recall that 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 clearly necessary, 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 $L^1(\mu)$.

CHAPTER 5

Sums of independent random variables-I

1. 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.

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 25: 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 $\delta > 0$,

$$\mathbf{P}\left\{ \left| \frac{1}{n}S_n - \mathbf{E}[X_1] \right| > \delta \right\} \to 0 \qquad \text{as } n \to \infty.$$

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 $Y_n \overset{P}{\to} Y$. In this language, the conclusion of the weak law of large numbers is that $\frac{1}{n}S_n \overset{P}{\to} \mathbf{E}[X_1]$ (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,

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

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.

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

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

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

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

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

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 Var$(S_n^Y) = n$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 to a constant!

(3) The proof under finite variance assumption is the most useful one, as the minimality of assumptions is less important than the strength of the conclusion. For example, if we assume that X_i have exponential moments, one can get the deviation probability to decay exponentially. We shall see this later under the heading "concentration of measure".

(4) 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 $t\mathbf{P}\{|X| \ge t\} \to 0$ as $t \to \infty$ (in which case, the weak law holds with $a_n = \mathbf{E}[X\mathbf{1}_{|X|\le n}]$).

Note that the Cauchy distribution violates this condition.

Exercise 15

Find a distribution which satisfies the condition $t\mathbf{P}\{|X| \ge t\} \to 0$ but does not have finite expectation.

2. Applications of weak law of large numbers

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

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

Theorem 26: 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 $n \ge 1$, 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 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 $X_1, X_2, \dots$ i.i.d Ber(p) random variables. For any $p \in [0,1]$, we have

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

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 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 (2) 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 8

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 $\text{Bin}(n,p)$ 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.

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

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

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

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

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

Exercise 16

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.

2.3. 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...

3. 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 27: 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 21 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 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 $\mathbf{P}\left(|n^{-1}S_n| > \delta\right) \le n^{-4}\delta^{-4}\mathbf{E}[S_n^4] = O(n^{-2})$ which is summable, and by Lemma 21 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 the general SLLN[1].

[1] 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.

Theorem 28: SLLN under second moment assumption

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

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

$$\frac{1}{n_k}U_{n_k} \to \mathbf{E}[X_1^+] \quad \text{and} \quad \frac{1}{n_k}V_{n_k} \to \mathbf{E}[X_1^-], \quad \text{a.s.} \tag{3}$$

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 (3), 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 (3), 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. We give two proofs, one in this section and one in the next[2].

In the first proof, 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 for positive random variables. 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! As before,

[2] The proof given in this section is due to Etemadi. Most books in probability give this proof. The presentation is adapted from a blog article of Terence Tao.

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

$$\frac{S_{N_1}}{N_2} \leq \frac{S_n}{n} \leq \frac{S_{N_2}}{N_1} \text{ if } N_1 \leq n \leq N_2. \tag{4}$$

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 29

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

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

$$\begin{aligned}
\mathbf{P}\left\{ \left| \frac{S_{n_j}}{n_j} - \mathbf{E}[X_1] \right| > 2\delta \right\} &\leq \mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| + \left| \frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1] \right| > 2\delta \right\} \\
&\leq \mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} + \mathbf{P}\left\{ \left| \frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1] \right| > \delta \right\} \\
&\leq \mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} + \mathbf{P}\left\{ \frac{S_{n_j}^Z}{n_j} \neq 0 \right\}. \qquad (5)
\end{aligned}$$

We shall show that both terms in (5) are summable over j. The first term can be bounded by Chebyshev's inequality

$$\mathbf{P}\left\{ \left| \frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1] \right| > \delta \right\} \leq \frac{1}{\delta^2 n_j} \mathbf{E}[Y_1^2] = \frac{1}{\delta^2 n_j} \mathbf{E}[X_1^2 \mathbf{1}_{X_1 \leq n_j}]. \tag{6}$$

while the second term is bounded by the union bound

$$\mathbf{P}\left\{ \frac{S_{n_j}^Z}{n_j} \neq 0 \right\} \leq n_j \mathbf{P}(X_1 > n_j). \tag{7}$$

The right hand sides of (6) and (7) 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 (6) and (7) 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 21 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 (4) 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 $\lim_{n \to \infty} \frac{S_n}{n} = \mathbf{E}[X_1]\ a.s.$ ■

4. Another proof of the SLLN via a maximal inequality

Here we give another proof of the SLLN, much shorter and involving hardly any technicalities[3]. But the techniques used in the first proof are useful and worth keeping in mind.

Lemma 30: A maximal inequality

Let X_k be i.i.d. random variables with finite expectation. Then, for any $t > 0$,

$$\mathbf{P}\left\{ \sup_n \frac{1}{n} S_n > t \right\} \le \frac{1}{t} \mathbf{E}[|X_1|].$$

The proof will assume that we know the SLLN for bounded i.i.d. random variables. Indeed, we do know a simple proof under the fourth moment assumption by a direct application of the first Borel-Cantelli lemma.

[3] Sauditya Jaiswal suggested that we could prove the SLLN on these lines, using the maximal inequality. When he asked me about it, my first response was that we shall see this proof when we study reverse martingales. That is true, but then I found that Michael Steele has a beautiful exposition (*Explaining a mysterious maximal inequality—and a path to the law of large numbers. Amer. Math. Monthly 122 (2015), no. 5, 490–494.*) that gives an elementary proof of the maximal inequality and deduces the SLLN from it. It seems nice enough to include here.

PROOF OF SLLN ASSUMING LEMMA 30. Fix $A > 0$ and define $Y_n = X_n \mathbf{1}_{|X_n| \le A}$ and $Z_n = X_n \mathbf{1}_{|X_n| > A}$, so that $X_n = Y_n + Z_n$ and $S_n^X = S_n^Y + S_n^Z$. The two sums can be controlled separately as follows.

(1) $\frac{1}{n} S_n^Y \overset{a.s.}{\to} \mathbf{E}[X_1 \mathbf{1}_{|X_1| \le A}]$ by the SLLN for bounded random variables

(2) For any $\varepsilon > 0$, by Lemma 30,

$$\mathbf{P}\left\{\limsup \frac{1}{n} S_n^Z > \varepsilon\right\} \le \mathbf{P}\left\{\sup_n \frac{1}{n} S_n^Z > \varepsilon\right\} \le \frac{1}{\varepsilon}\mathbf{E}[|X_1| \mathbf{1}_{|X_1| > A}]$$

Putting these together, we have

$$\begin{aligned}\limsup_{n\to\infty} \frac{S_n^X}{n} &\le \limsup_{n\to\infty} \frac{S_n^Y}{n} + \limsup_{n\to\infty} \frac{S_n^Z}{n} \\ &\le \mathbf{E}[X_1 \mathbf{1}_{|X_1| \le A}] + \varepsilon \quad w.p. \ge 1 - \frac{1}{\varepsilon}\mathbf{E}[|X_1| \mathbf{1}_{|X_1| > A}].\end{aligned}$$

Now let $A \to \infty$ and then $\varepsilon \downarrow 0$ (and note that $\mathbf{E}[X_1 \mathbf{1}_{|X_1| \le A}] \to \mathbf{E}[X_1]$ and $\mathbf{E}[X_1 \mathbf{1}_{|X_1| > A}] \to 0$ by DCT) to get $\limsup \frac{S_n^X}{n} \le 0$ $a.s.$ Applying the same to $-X_i$ gives $\liminf \frac{S_n^X}{n} \ge 0$ $a.s.$ Hence $\frac{S_n}{n} \overset{a.s.}{\to} \mathbf{E}[X_1]$. ■

It remains to prove the maximal inequality.

PROOF OF LEMMA 30. Define

$$\begin{aligned}M_n &= \max\{0, X_1, X_1 + X_2, \dots, X_1 + \dots + X_n\}, \\ M_n' &= \max\{0, X_2, X_2 + X_3, \dots, X_2 + \dots + X_{n+1}\}.\end{aligned}$$

Observe that these quantities are positive. On the event $\{M_n > 0\}$, we can drop the zero from the maximum and write

$$\begin{aligned}M_n &= \max\{X_1, X_1 + X_2, \dots, X_1 + \dots + X_n\} \\ &= X_1 + \max\{0, X_2, \dots, X_2 + \dots + X_n\} \\ &\le X_1 + M_n'.\end{aligned}$$

Hence, $M_n - M_n' \le X_1$ on the event $M_n > 0$. On the event $M_n \le 0$ we have the trivial bound $M_n - M_n' \le 0$ (since $M_n' \ge 0$ anyway). Putting them together, $M_n - M_n' \le X_1 \mathbf{1}_{M_n > 0}$.

If X_k are i.i.d. with finite mean, we have $M_n \overset{d}{=} M_n'$ and hence have the same expectation (check that $\mathbf{E}[M_n]$ exists). Hence $\mathbf{E}[X_1 \mathbf{1}_{M_n > 0}] \ge 0$.

Fix $t > 0$ and apply this to $X_i - t$ to get $\mathbf{E}[(X_1 - t)\mathbf{1}_{M_n > t}] \ge 0$ which implies that

$$\mathbf{P}\{M_n > t\} \le \frac{1}{t}\mathbf{E}[X_1].$$

Let $n \to \infty$ and note that $M_n \uparrow \sup_n \frac{S_n}{n}$ to get the statement of the Lemma. ■

5. Beyond the law of large numbers

There are multiple ways in which we can go beyond the laws of large numbers. Here are some important ones that we shall not be going into in great detail (but will touch upon some).

(1) From Chebyshev inequality $\mathbf{P}\{|\frac{1}{n}S_n - \mu| \geq \delta\} = O(1/n)$ under second moment assumption. If we assume more on the random variables, can we improve the estimate? The best sort of estimate one can hope for in general are of the form $e^{-c_\delta n}$. These questions come under the topic of *concentration of measure*.

(2) In the cases where we get bounds such as $e^{-c_\delta n}$, one could ask for explicit form of c_δ. Since the inequality was written to be valid for all n, one may be forced to choose small c_δ to take care of small values of n. Something more fundamental may result from asking the inequality for large n. In other words, we ask for I_δ so that $e^{-n(I_\delta+\varepsilon)}\mathbf{P}\{|\frac{1}{n}S_n - \mu| \geq \delta\} \leq e^{-n(I_\delta-\varepsilon)}$ for any $\varepsilon > 0$ and for large enough n (how large depends on ε). These questions come under the topic of *large deviation theory*.

(3) In the previous points $\delta > 0$ was fixed, which means that the deviation of S_n from $n\mu$ is of the order of n. What kinds of bounds can one get for $\mathbf{P}\{|S_n - n\mu| \geq n^p\}$ for $p < 1$? As the standard deviation of S_n is of the order of $\sqrt{n}$, it makes sense to take $p > \frac{1}{2}$. These questions are often called *moderate deviations*.

(4) While SLLN says that $\frac{1}{n}S_n \overset{a.s.}{\to} 0$ (when $\mathbf{E}[X_i] = 0$), what happens if we divide by something less, such as $n^{0.9}$? For any $a_n \uparrow \infty$, by Kolmogorov's zero-one law one can see that $\limsup_{n\to\infty} \frac{S_n}{a_n}$ is a constant random variable. If this constant is zero, then a_n is too large, if the constant is ∞ then a_n is too small. Could we find a_n so that the constant is 1? It turns out that the right answer is $a_n = \sqrt{2n\log\log n}$ (when X_i have zero mean and unit variance), hence the relevant results is called the *law of iterated logarithm*.

More generally, whenever we have a sequence of random variables $\xi_n \overset{a.s.}{\to} 0$, one can ask for numbers $b_n \uparrow \infty$ so that $\limsup b_n\xi_n = 1$.

Of course there are many other directions, such as relaxing the assumptions of identical distribution or independence. But they are not well-suited to cover in class and we ignore such questions entirely. Instead, in this section we work out detailed estimates for the special case of Bernoulli random variables, answering the above questions in detail.

5.1. Bernoulli random variables. Let X_i be i.i.d. Ber$(1/2)$ random variables. Then S_n has the transformed Binomial distribution

$$p_n(k) := \mathbf{P}\{S_n = k\} = \binom{n}{k}\frac{1}{2^n} \quad 0 \leq k \leq n.$$

By Stirling's formula, we have the following estimate when n as well as k and $n-k$ are large:

$$
\begin{aligned}
p_n(k) &\sim \frac{n^{n+\frac{1}{2}}}{2^n k^{k+\frac{1}{2}}(n-k)^{n-k+\frac{1}{2}}\sqrt{2\pi}} \\
&= \frac{n^n}{2^n k^k (n-k)^{n-k}} \frac{\sqrt{n}}{\sqrt{2\pi}\sqrt{k(n-k)}} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi}\sqrt{k(n-k)}} \exp\left\{-n\left[\log 2 + \frac{k}{n}\log\frac{k}{n} + \frac{n-k}{n}\log\frac{n-k}{n}\right]\right\} \\
&= \frac{\sqrt{n}}{\sqrt{2\pi}\sqrt{k(n-k)}} e^{-nI(k/n)}
\end{aligned}
$$

where $I(x) = \log 2 + x\log x + (1-x)\log(1-x)$ for $x \in [0,1]$ (with the interpretation that $0\log 0 = 0$, by continuity). is called the Shannon entropy function. The precise meaning of the approximation in the first line is that given $\varepsilon > 0$, there exist N and K such that for all $n \geq N$ and $K \leq k \leq N-K$, we have

$$
\frac{(1-\varepsilon)}{2\sqrt{n}} e^{-nI(k/n)} \leq p_n(k) \leq (1+\varepsilon)e^{-nI(k/n)}. \tag{8}
$$

where we used the fact that $k(n-k)$ is largest when $k = n/2$ and smallest when $k = 1$ (we anyway have $k \geq K$) to to simplify the form of the bounds.

The properties of $x \mapsto I(x)$ play a key role in the estimates for the probabilities. It is symmetric about $x = 1/2$, attains its minimum value of 0 uniquely at $x = 1/2$, is convex, and is bounded between the parabolas $2(x-\frac{1}{2})^2 \leq I(x) \leq 3(x-\frac{1}{2})^2$ for $0 \leq x \leq 1$.

Large deviations: If $x > \frac{1}{2}$, then take $\varepsilon = 1/2$ (or any fixed number in $(0,1)$) and use (8) to get

$$
\begin{aligned}
\mathbf{P}\{S_n > nx\} &\geq p_n(\lceil nx \rceil) \geq \frac{1}{4\sqrt{n}} e^{-nI(x)}, \\
\mathbf{P}\{S_n > nx\} &= \sum_{k \geq nx} p_n(k) \ \leq \ ne^{-nI(x)}.
\end{aligned}
$$

In the second line, we bounded all terms by the largest one (i.e., $p_n(\lceil nx \rceil)$ and used the fact that $I(x)$ is increasing on $[1/2, 1]$. As $I(x) > 0$ for $x > \frac{1}{2}$, the polynomial factors outside are negligible compared to the exponential term and we can simply write $\mathbf{P}\{S_n > nx\} \approx e^{-nI(x)}$ in the sense that

$$
\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{S_n > nx\} = -I(x).
$$

This is the statement of the large deviation principle for Bernoullis.

Concentration inequalities: From the estimate above and the fact that $I(x) \geq 2(x-\frac{1}{2})^2$, we get

$$
\mathbf{P}\{S_n > nx\} \leq ne^{-nI(x)} \ \leq \ ne^{-2n(x-\frac{1}{2})^2}
$$

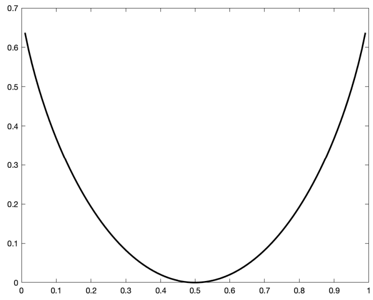

FIGURE 1. Graph of the function $x \mapsto I(x)$

We can get rid of the polynomial factor below and rewrite this as

$$\mathbf{P}\{S_n > nx\} \leq C_\varepsilon e^{(2-\varepsilon)n(x-\frac{1}{2})^2}$$

for any $\varepsilon > 0$ and $C_\varepsilon < \infty$ (required to take care of the case of small n). With more care, one can derive the following inequality of *Bernstein*

$$\mathbf{P}\{S_n > nx\} \leq 2e^{-2(x-\frac{1}{2})^2}$$

6. 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. 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,

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

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 $n^{-\alpha} S_n \overset{a.s.}{\to} 0$ 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 not zero. The subset of such indices 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 $a_n = h_n \sqrt{n \log n}$ for any sequence $h_n \to \infty$. In the converse 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? In other words, we want a deterministic sequence a_n such that $\limsup S_n/a_n$ is finite and strictly positive. Since the $\limsup$ is a constant a.s., we can scale by that and reformulate the question as follows.

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

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

Theorem 31: Khinchine's law of iterated logarithm

Let X_i be i.i.d. 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 (since the difference between successive terms is $1/\sqrt{2n\log\log n}$ that goes to zero) 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 32: Hartman-Wintner law of iterated logarithm

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

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

7. 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 31 follows from the following two statements. For any $\delta > 0$, we have

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \le 1+\delta \quad a.s. \tag{9}$$

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \ge 1-\delta \quad a.s. \tag{10}$$

Taking intersection over countably many values of δ, e.g., $\delta = \frac{1}{k}$, $k \ge 1$, we get the statement of LIL. To motivate the principal idea in the proof, consider the following toy situation.

Example 13: 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 naively, 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 $1 = n_1 < n_2 < \ldots$ 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 (9). 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

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

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

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

$$\mathbf{P}(D_k) \le C_\lambda k^{-(1+\delta)^2/\lambda} \quad \text{where } C_\lambda < \infty \text{ for any } \lambda > 1. \tag{11}$$

Granting this, it is clear that choosing $1 < \lambda < (1+\delta)^2$ ensures summability of $\mathbf{P}(D_k)$. We give two proofs of the inequality (11) below, which completes the proof. ■

Proof of (11) via the reflection principle: The following lemma is of interest in itself and useful.

Lemma 33: Reflection principle/Ballot problem

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

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

Equality holds if n and a have opposite parity.

Chapter-3 of Feller's vol-1 is highly recommended for more such beautiful combinatorial facts about simple symmetric random walks.

Proof. Break the event $\max\{S_0, \dots, S_n\} \ge a$ as a union of pairwise disjoint events

$$A_k = \{S_0 < a, \dots, S_{k-1} < a, S_k = a\}, \quad k = 1, \dots, n.$$

By the symmetry of $S_n - S_k$ and its independence from A_k,

$$\begin{aligned}\mathbf{P}(\{S_n \ge a\} \cap A_k) &= \mathbf{P}(\{S_n - S_k \ge 0\} \cap A_k) \\ &= \mathbf{P}\{S_n - S_k \ge 0\}\mathbf{P}\{A_k\} \ \ge \ \frac{1}{2}\mathbf{P}(A_k). \end{aligned} \tag{12}$$

Sum over k. On the right we get $\frac{1}{2}\mathbf{P}\{\max\{S_0, \dots, S_n\} \ge a\}$ while on the left we get $\mathbf{P}\{S_n \ge a\}$ (since $\{S_n \ge a\} \subseteq A_1 \cup \ldots \cup A_n$). Hence the second inequality is proved. To prove the first inequality, using the same idea, write

$$\begin{aligned}\mathbf{P}(\{S_n > a\} \cap A_k) &= \mathbf{P}(\{S_n - S_k > 0\} \cap A_k) \\ &= \mathbf{P}\{S_n - S_k > 0\}\mathbf{P}\{A_k\} \ \le \ \frac{1}{2}\mathbf{P}(A_k). \end{aligned} \tag{13}$$

Add up over k to get $2\mathbf{P}\{S_n > a\} \le \mathbf{P}\{\max\{S_0, \dots, S_n\} \ge a\}$.

If n has the opposite parity, then $\mathbf{P}\{S_n = a\} = 0$, hence all three probabilities in the statement are equal. ■

Returning to the proof of (11), if D_k occurs, then there is some $n \leq n_{k+1}$ (in fact some $n \geq n_k$) such that $S_n \geq a_{n_k}(1+\delta)$. The reflection principle in Lemma 33 applies to give the bound

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

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

$$\frac{(1+\delta)^2 2\lambda^k \log\log\lambda^k}{2\lambda^{k+1}} = \frac{(1+\delta)^2}{\lambda}\log(k\log\lambda) \tag{14}$$

from which (11) immediately follows. ■

Proof of (11) via the modified Markov inequality (2): Let $X_k = \sum_{n=n_k}^{n_{k+1}-1} \mathbf{1}_{S_n > a_{n_k}(1+\delta)}$, so that D_k is the event that $X_k \geq 1$. Apply the strengthened form of Markov's inequality (2) to write

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

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

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

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 $\mathbf{E}[X_k \bigm| X_k \geq 1]$. The heuristic idea is that if $X_k \geq 1$, there is some (random) $N \in [n_k, n_{k+1})$ 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,

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

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,

$$\mathbf{P}(D_k) \leq \frac{(n_{k+1} - n_k)\exp\left\{-\frac{a_{n_k}^2(1+\delta)^2}{2n_{k+1}}\right\}}{\frac{1}{8}(n_{k+1} - n_k)} \leq 8e^{-\frac{a_{n_k}^2(1+\delta)^2}{2n_{k+1}}}.$$

By the computation shown in (14), this is of the form given in (11). ■

7.1. Proof of the lower bound (10). Again we choose a subsequence $n_k = \lfloor \lambda^k \rfloor$, the difference being that we shall choose λ to be a large constant in the end. It suffices to show for any $\delta > 0$ that

$$\mathbf{P}\{S_{n_k} \geq (1-2\delta)a_{n_k} \text{ i.o.}\} = 1 \tag{15}$$

where $a_n = \sqrt{2n \log\log n}$ as before. By the upper bound and the symmetry of S_n, we know that almost surely, $S_{n_k} \geq -2a_{n_k}$ for all but finitely many k. Also, $a_{n_k} \leq a_{n_{k+1}}/\sqrt{\lambda}$, hence

$$S_{n_{k+1}} \geq S_{n_{k+1}} - S_{n_k} - \frac{2}{\sqrt{\lambda}}a_{n_{k+1}}$$

for all but finitely many k, a.s. Therefore, (15) follows if we choose $\lambda > 4/\delta^2$ and show that

$$\mathbf{P}\{S_{n_{k+1}} - S_{n_k} \geq (1-\delta)a_{n_{k+1}} \text{ i.o.}\} = 1.$$

These events are independent across k, and hence a good lower bound on the individual probabilities is sufficient. The one given below in Claim 34 gives

$$\begin{aligned}\mathbf{P}\{S_{n_{k+1}} - S_{n_k} \geq (1-\delta)a_{n_{k+1}}\} &\geq \frac{\sqrt{2}}{\sqrt{\pi(n_{k+1} - n_k)}}\exp\left\{-\frac{(1-\delta)^2 a_{n_{k+1}}^2}{2(n_{k+1} - n_k)}\right\} \\ &= \frac{\sqrt{2}}{\sqrt{\pi n_{k+1}(1 - \frac{1}{\lambda})}}\exp\left\{-\frac{(1-\delta)^2 \log\log n_{k+1}}{1 - \frac{1}{\lambda}}\right\}\end{aligned}$$

Claim 34: An estimate for binomial coefficients

If $n, k \to \infty$ in such a way that $|k - \frac{1}{2}n| \leq n^{2/3}$, then

$$\binom{n}{\frac{n+k}{2}} \frac{1}{2^n} \sim \frac{\sqrt{2}}{\sqrt{\pi n}} e^{-\frac{k^2}{2n}}.$$

In particular, for such k, we have

$$\mathbf{P}\{S_n \geq k\} \geq e^{-\frac{1}{2}\frac{k^2}{2n}}$$

In a basic probability class you may have seen the de Moivre-Laplace theorem that compares binomial coefficients to the Gaussian density. This one is almost the same, except that in the de Moivre-Laplace theorem one only needs $k = \frac{1}{2}n + x\sqrt{n}$ with fixed x, while here we allow x to grow like $O(n^{1/6})$.

PROOF. The first one is just by Stirling's approximation. ■

8. 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 14

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 35: 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

$$(16) \qquad \mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } n \geq N\right) = \lim_{m\to\infty} \mathbf{P}\left(|S_n - S_N| > \delta \text{ for some } N \leq n \leq N+m\right).$$

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

Invoking 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 (16), 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 35 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 35 that $\sum Y_n$ converges a.s. Thus we have proved

Theorem 36: 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.

CHAPTER 6

Sums of independent random variables - II

1. 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 37: Central limit theorem for i.i.d. variables

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

Informally, letting Z denote a standard Normal variable, we may write $S_n \approx n\mu + \sigma\sqrt{n}Z$. More precisely, $\mathbf{P}\{S_n \le n\mu + \sigma\sqrt{n}t\} \to \mathbf{P}\{Z \le t\}$ for any $t \in \mathbb{R}$. 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. This feature that the behaviour of a large yet random system does not depend on the details of the microscopic parts that go into building it, is called *universality* and is a major theme of modern probability.

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 scale by $\sqrt{n}$? 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 mean is zero and 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 $(2n)^{-\frac{1}{2}}S_{2n} \overset{d}{\to} Y$ and further,

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

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

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

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

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

Thus we must have $Y_1 + Y_2 \overset{d}{=} \sqrt{2}Y$. If $Y_1 \sim N(0,\sigma^2)$, then certainly it is true that $Y_1 + Y_2 \overset{d}{=} \sqrt{2}Y$. 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. The claim itself is not trivial, we discuss it in the section on the Gaussian distribution.

Justification by examples: 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.

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 $\text{Var}(X_1) = 1$. Then $S_n \sim \text{Gamma}(n, 1)$ and hence $\frac{S_n - n}{\sqrt{n}}$ has density

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

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

Exercise 17

Prove the CLT for for the following distributions of X_is. (1) Ber(p). (2) Bin(k,p). (3) Poisson(λ). (4) Geometric(p).

The special feature of these cases is that we can explicitly work out the distribution of S_n. This is not the case in general, and in fact one of the uses of central limit theorem (for example, in statistics) goes the other way. We use the Normal distribution as an approximation to the distribution of S_n.

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

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

Fix $k = (k_1, \ldots, k_{2p})$ and consider the corresponding summand. The expectation factors as a product of $\mathbf{E}[X^{\ell_i}]$, $1 \le i \le n$, where ℓ_i is the number of j for which $k_j = i$. Unless each ℓ_i is even, the summand vanishes and if each $\ell_i = 1$. 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 $n(n-1)\ldots(n-p+1)$, and the number of ways that these indices may be chosen is $(2p-1)(2p-3)\ldots(3)(1)$.

Next those terms in which at least one ℓ_i is equal to 4. Then there are at most $p-1$ distinct indices, and they can be chosen in at most n^{p-1} ways. The number of ways of choosing ℓ_is is itself a number that depends only on p, say C_p.

2. Gaussian distribution

We collect some basic facts about the Gaussian distribution here. The standard Gaussian measure is denoted γ, its density is denoted φ and its distribution function is denote Φ. The density of $N(\mu, \sigma^2)$ is then $\sigma^{-1}\varphi((x-\mu)/\sigma)$. We also use the notation $p_t(\cdot)$ for the density of $N(0,t)$. We usually write $Z, Z_1, Z_2, \ldots$ for standard Gaussian random variables.

2.1. Heat equation. Consider $p_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$ for $t > 0$ and $x \in \mathbb{R}$. Differentiation gives

$$\left(\frac{\partial}{\partial t} - \frac{1}{2}\frac{\partial^2}{\partial x^2}\right) p_t(x) = 0.$$

In other words, $p_t(x)$ is a solution to the heat equation. This is the single most important fact about the Gaussian distribution.

2.2. Integration by parts formula. Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function such that $|x|^j f^{(k)}(x) \in L^1(\gamma)$ for any j, k (we need much less below). Then, as $\int f(x/\sqrt{t})p_t(x)dx = \mathbf{E}[f(Z)]$ for any t, differentiating w.r.t. t under the integral, we get

$$\begin{aligned} 0 &= \frac{d}{dt}\int_{\mathbb{R}} f(x/\sqrt{t})p_t(x)dx \\ &= -\frac{1}{2t^{3/2}}\int_{\mathbb{R}} f'(x/\sqrt{t})xp_t(x)dx + \frac{1}{2}\int_{\mathbb{R}} f(x/\sqrt{t})p_t''(x)dx \quad \text{(by heat equation)} \\ &= -\frac{1}{2t^{3/2}}\int_{\mathbb{R}} f'(x/\sqrt{t})xp_t(x)dx + \frac{1}{2t}\int_{\mathbb{R}} f''(x/\sqrt{t})p_t(x)dx \quad \text{(integration by parts)} \end{aligned}$$

from which, setting $t = 1$, we arrive at the *Gaussian integration by parts formula*

$$\mathbf{E}[Zf'(Z)] = \mathbf{E}[f''(Z)]. \tag{17}$$

We leave it as an exercise to justify the differentiation under integral and the integration by parts. If we set $h = f'$, then (17) transforms to

$$\mathbf{E}[Zh(Z)] = \mathbf{E}[h'(Z)], \tag{18}$$

which is often called *Stein's identity*[1]. With a bit more care, one can prove that (18) holds for any $h : \mathbb{R} \to \mathbb{R}$ that is absolutely continuous with $h' \in L^1(\gamma)$ (this means that $h(x) = \int_{-\infty}^{x} g(t)dt$ for some $g \in L^1(\gamma)$, which is then called the derivative of h and denoted as h').

2.3. Moments. The odd moments are zero by symmetry, while the even moments can be got by a direct integration. Alternately, use integration by parts formula (17) with $f(x) = x^{2p}$ we get $\mathbf{E}[Z^{2p}] = (2p-1)\mathbf{E}[Z^{2p-2}]$, from which it follows that

$$\mathbf{E}[Z^{2p}] = (2p-1) \times (2p-3) \times \ldots \times 3 \times 1.$$

2.4. Characteristic function. Formally one can see that $\mathbf{E}[e^{itZ}] = e^{-\frac{1}{2}t^2}$ by substituting it in the moment generating function. For an honest proof, apply the integration by parts formula to $f(x) = e^{itx}$ to get $\mathbf{E}[itZe^{itZ}] = -t^2\mathbf{E}[e^{itZ}]$. Setting $\varphi(t) = \mathbf{E}[e^{itZ}]$ we see (again, differentiating under the expectation) that $\varphi'(t) = -t\varphi(t)$, for which the unique solution satisfying $\varphi(0) = 1$ is

$$\varphi(t) = e^{-\frac{1}{2}t^2}.$$

[1] As Arka Das pointed out in class, (18) can be got directly by writing $\mathbf{E}[f'(Z)] = \int f'(x)\varphi(x)dx$ and integrating by parts. We gave a more roundabout derivation to emphasize its connection with the heat equation. In addition, the dynamical viewpoint of considering p_t, $t > 0$, is of great importance. The identity (17) is related to the Ornstein-Uhlenbeck process, a Markov process with stationary distribution $N(0,1)$.

2.5. Characterizations of Gaussian distribution. A feature of a probability distribution that is not shared by any other probability distribution is called a *characterization* of the said distribution. For example, the characteristic function determines the distribution, hence is always a characterization. Any distribution μ with finite moment generating function (i.e., $\int e^{tx}d\mu(x) < \infty$ for $|t| < \delta$ for some $\delta > 0$) is characterized by its moment sequence.

In particular, the Gaussian distribution is characterized by its moments, i.e., no other distribution has the same moments as the standard Gaussian distribution. The identities (17) and (18) are also characterizations of the standard Gaussian distribution. This means that if $\mathbf{E}[h'(W)] = \mathbf{E}[Wh(W)]$ for a large enough class of functions h, then $W \sim N(0,1)$. For instance, we saw that applying it to $h = e_t$ one can derive that the characteristic function of $N(0,1)$ is $e^{-t^2/2}$, but one can also consider other classes of functions (e.g., $C_c^1(\mathbb{R})$) that do not contain e_ts. Yet another characterization is the *stability* property that we used earlier: If W, W' are i.i.d. and $W + W' \overset{d}{=} \sqrt{2}W$, then $W \sim N(0, \sigma^2)$ for some $\sigma^2 \geq 0$. To see this, suppose $\psi(\cdot)$ denotes the characteristic function of W, then

$$\psi(t) = \mathbf{E}\left[e^{itW}\right] = \mathbf{E}\left[e^{\frac{it(W+W')}{\sqrt{2}}}\right]^2 = \psi\left(\frac{t}{\sqrt{2}}\right)^2.$$

From this, by standard methods (note that characteristic functions are necessarily continuous), one can deduce that $\psi(t) = e^{-at^2}$ for some $a > 0$. Therefore, $W \sim N(0, 2a)$.

3. Strategies of proof of central limit theorem

To show that a random variable $W \sim N(0,1)$, it suffices to show that it has any one of the characterizing properties of the standard Gaussian distribution. In the context of CLT, we have a sequence $W_n = S_n/\sqrt{n}$ that we must show converges to $N(0,1)$ in distribution. Hence we wish to know if W_n approximately has a characterizing property (and the approximation gets better as $n \to \infty$), does it mean that $W_n \xrightarrow{d} N(0,1)$? Here are the essential statements that give a positive answer, hence each of them provides a possible route to showing that $W_n \xrightarrow{d} N(0,1)$.

Theorem 38

Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$ and let $W_n \sim \mu_n$ and $W \sim \mu$. Each of the following is equivalent to $W_n \xrightarrow{d} W$.

(1) $\mathbf{E}[f(W_n)] \to \mathbf{E}[f(W)]$ for all $f \in C_b^{(\infty)}(\mathbb{R})$ (i.e., $f^{(j)} \in C_b(\mathbb{R})$ for all j).

(2) $\mathbf{E}[e_t(W_n)] \to \mathbf{E}[e_t(W)]$ for all $t \in \mathbb{R}$.

If $\mu = \gamma$, then the following statement also implies that $W_n \xrightarrow{d} N(0,1)$: $\mathbf{E}[|W_n|] < \infty$ and

$$\mathbf{E}[h'(W_n)] - \mathbf{E}[W_n h(W_n)] \to 0 \qquad \text{if } h \in C_b^1(\mathbb{R}).$$

The second statement is known as *Levy's continuity theorem* and is proved in the section on characteristic functions. Further, what we need is the conclusion that $W_n \overset{d}{\to} W$, so we prove the relevant one-way implications in the first and third statements.

PROOF. (1) Fix t and for $k \geq 1$ find $f_k \in C^\infty$ such that $\mathbf{1}_{(-\infty,t]} \leq f_k \leq \mathbf{1}_{(-\infty,t+\frac{1}{k}]}$. Taking expectations, we see that

$$\mathbf{P}\{W_n \leq t\} \leq \mathbf{E}[f_k(W_n)] \to \mathbf{E}[f_k(W)] \leq \mathbf{P}\{W \leq t + \frac{1}{k}\}.$$

Let $k \to \infty$ to get $\limsup F_{\mu_n}(t) \leq F_\mu(t)$. Similarly,

$$\mathbf{P}\{W_n \leq t + \frac{1}{k}\} \geq \mathbf{E}[f_k(W_n)] \to \mathbf{E}[f_k(W)] \geq \mathbf{P}\{W \leq t\}.$$

Replace t by $t - \frac{1}{k}$ and let $k \to \infty$ to get $\liminf F_{\mu_n}(t) \geq F_\mu(t-)$.

(2)

■

3.1. Outline of three proofs of CLT. We present three proofs of the central limit theorem.

(1) Using characteristic functions: In this proof we show that $\mathbf{E}[e_t(S_n/\sqrt{n})] \to e^{-t^2/2}$ for all $t \in \mathbb{R}$. The reason that the characteristic function is so effective is that for sums of independent random variables, the characteristic function will be a product of the individual characteristic functions. Additional ingredients are basic facts about characteristic functions, which imply that if $\mathbf{E}[e_t(X_1/\sqrt{n})] \approx 1 - \frac{t^2}{2n}$ if $\mathbf{E}[X_1] = 0$ and $\mathbf{E}[X_1^2] = 1$. Hence $\mathbf{E}[e_t(S_n/\sqrt{n})] \approx (1 - \frac{t^2}{2n})^n \approx e^{-t^2/2}$. A little work is needed to make the approximations precise.

(2) Using Lindeberg's replacement principle: In this proof, along with X_i, we construct independent standard Gaussians Z_is on the same probability space, and show that $\mathbf{E}[f(S_n^X/\sqrt{n})] \approx \mathbf{E}[f(S_n^Z/\sqrt{n})]$. As the latter is the same as $\mathbf{E}[f(Z)]$, CLT follows. To show the closeness of expectations, the idea is to go from S_n^X to S_n^Z in n steps, by replacing each X_i by Z_i, one after another. The heart of the proof is in showing that the difference in expectations in each step is $o(1/n)$.

(3) Using Stein's method: This proof works by showing that $W_n = S_n/\sqrt{n}$ satisfies the Stein identity approximately.

To not obfuscate the main ideas with less important technicalities, we present the first two proofs assuming that the third moment of X_is is finite. Then we shall in fact state the more general *Lindeberg-Feller central limit theorem* and prove it under minimal conditions, thereby also proving the standard CLT under second moment assumption. The proof by Stein's method is given thereafter.

4. Central limit theorem - two proofs assuming third moments

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

Theorem 39

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 37.

4.1. Proof via characteristic functions.

We shall need the following facts.

Exercise 18

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

PROOF OF THEOREM 39. By Lévy's continuity theorem (Lemma **??**), 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 18 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 $N(0,1)$. ■

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

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

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

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

By Taylor expansion, we see that

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

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

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

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

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

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 **??** we see that $\frac{1}{\sqrt{n}}S_n \xrightarrow{d} N(0,1)$. ■

5. 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 40: Lindeberg-Feller CLT

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

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

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

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

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

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

PROOF OF THEOREM 37 FROM THEOREM 40. Let $X_{n,k} = n^{-\frac{1}{2}} X_k$ for $k = 1, 2, \ldots, n$. Then, $\mathbf{E}[X_{n,k}] = 0$ while $\sum_{k=1}^{n} \mathbf{E}[X_{n,k}^2] = \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}[X_1^2] = \sigma^2$, for each n. Further, $\sum_{k=1}^{n} \mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] = \mathbf{E}[X_1^2 \mathbf{1}_{|X_1|>\delta\sqrt{n}}]$ which goes to zero as $n \to \infty$ by DCT, since $\mathbf{E}[X_1^2] < \infty$. Hence the conditions of Lindeberg Feller theorem are satisfied and we conclude that $\frac{S_n}{\sqrt{n}}$ converges in distribution to $N(0,1)$. ■

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

Example 15

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 40, 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 40 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 $1 \leq k \leq n$). 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 19

Suppose X_k are independent random variables and $\mathbf{E}[|X_k|^{2+\delta}] \le M$ 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 16

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 40.

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

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

Exercise 20

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 $T_n := \sum_{k=1}^n a_k X_k$.

6. 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.

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

Exercise 21

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

PROOF OF THEOREM 40. The characteristic function of $S_n = X_{n,1} + \ldots + X_{n,n}$ is given by $\psi_n(t) = \prod\limits_{k=1}^{n} \mathbf{E}\left[e^{itX_{n,k}}\right]$. Again, we shall use the Taylor expansion of e^{itx}, but we shall need both the second and first order expansions.

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

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

$$|R_{n,k}(t)| \le t^2\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|>\delta}X_{n,k}^2\right] + \frac{|t|^3\delta}{6}\mathbf{E}\left[X_{n,k}^2\right]. \tag{19}$$

We want to apply Exercise 21 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 $\max\limits_{k\le n}\sigma_{n,k}^2 \to 0$, then it will follow that $|w_k| \le 1$ and hence with $\theta = 1$ in Exercise 21, we get

$$\begin{aligned} \limsup_{n\to\infty}\left|\prod_{k=1}^{n}\mathbf{E}\left[e^{itX_{n,k}}\right] - \prod_{k=1}^{n}\left(1 - \frac{1}{2}\sigma_{n,k}^2t^2\right)\right| &\le \limsup_{n\to\infty}\sum_{k=1}^{n}|R_{n,k}(t)| \\ &\le \frac{1}{6}|t|^3\sigma^2\delta \quad (\text{by } 19) \end{aligned}$$

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

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

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

$$\lim_{n\to\infty}\prod_{k=1}^{n}\mathbf{E}\left[e^{itX_{n,k}}\right] = \lim_{n\to\infty}\prod_{k=1}^{n}\left(1 - \frac{1}{2}\sigma_{n,k}^2t^2\right). \tag{20}$$

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

6.2. 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,

$$\begin{aligned}\mathbf{E}\left[f\left(V_n\right)\right]-\mathbf{E}[f(V_0)] &= \sum_{k=1}^n \mathbf{E}\left[f\left(V_k\right)-f\left(V_{k-1}\right)\right] \\ &= \sum_{k=1}^n \mathbf{E}\left[f\left(V_k\right)-f\left(U_k\right)\right] - \sum_{k=1}^n \mathbf{E}\left[f\left(V_{k-1}\right)-f\left(U_k\right)\right].\end{aligned} \tag{21}$$

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

$$|\mathbf{E}[f(V_k)]-\mathbf{E}[f(V_{k-1})]| \le \frac{1}{2}\mathbf{E}[|f''(U_k)|]\left|\mathbf{E}[X_{n,k}^2\mathbf{1}_{|X_{n,k}|\le\delta}] - \mathbf{E}[Y_{n,k}^2\mathbf{1}_{|Y_{n,k}|\le\delta}]\right| \tag{22}$$

$$+\frac{1}{2}\left|\mathbf{E}[|f''(U_k^{\#})|X_{n,k}^2\mathbf{1}_{|X_{n,k}|>\delta}]\right| + \frac{1}{2}\left|\mathbf{E}[|f''(U_k^{\#\#})|Y_{n,k}^2\mathbf{1}_{|Y_{n,k}|>\delta}]\right| \tag{23}$$

$$+\frac{1}{6}\left|\mathbf{E}[|f'''(U_k^*)||X_{n,k}|^3\mathbf{1}_{|X_{n,k}|\le\delta}]\right| + \frac{1}{6}\left|\mathbf{E}[|f'''(U_k^{**})||Y_{n,k}|^3\mathbf{1}_{|Y_{n,k}|\le\delta}]\right| \tag{24}$$

Since $\mathbf{E}[X_{n,k}^2] = \mathbf{E}[Y_{n,k}^2]$, the term in the first line (22) is the same as $\frac{1}{2}\mathbf{E}[|f''(U_k)|] \, \big|\mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] - \mathbf{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\big|$ 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 (23) 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 (24), we show how to deal with the first.

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

The same bound holds for the second term in (24). Putting all this together, we arrive at

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

Add up over k and use (21) to get

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

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 $\limsup \big|\mathbf{E}[f(V_0)] - \mathbf{E}[f(V_n)]\big| \le 2\sigma^2\delta$. 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 $X_{n,1} + \ldots + X_{n,n} \xrightarrow{d} N(0, \sigma^2)$. ■

7. 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 17

Let X_i be i.i.d Cauchy(1), with density $\frac{1}{\pi(1+x^2)}$. Then, one can check that $\frac{S_n}{n}$ has exactly the same Cauchy distribution! Thus, to get distributional convergence, we just write $\frac{S_n}{n} \xrightarrow{d} C_1$.

If X_i were i.i.d with density $\frac{a}{\pi(a^2+(x-b)^2)}$ (which can be denoted $C_{a,b}$ with $a > 0$, $b \in \mathbb{R}$), then $\frac{X_i - b}{a}$ are i.i.d C_1, and hence, we get

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

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

This raises the following questions.

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

(2) What are the possible limiting distributions?

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

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

Symmetric α-stable distributions Fix $\alpha > 0$. Do there exist i.i.d. random variables X, Y such that $X + Y \overset{d}{=} 2^{\frac{1}{\alpha}} X$? When $\alpha = 2$, centered Gaussian distributions satisfy the distributional equation, and when $\alpha = 1$, the symmetric Cauchy distributions do. What about other α?

From the distributional identity, if $X, Y \sim \mu$ are i.i.d., then the characteristic function $\hat{\mu}$ satisfies $\hat{\mu}(2^{1/\alpha}t) = \hat{\mu}(t)^2$. As $\hat{\mu}$ is continuous, real-valued and symmetric, it is not hard to see that $\hat{\mu}(t) = e^{-c|t|^\alpha}$. Of course, we don't know if this is a valid characteristic function, i.e., if such a distribution μ exists. This is answered in the following theorem.

Theorem 41: Symmetric stable distributions

The symmetric α-stable distribution exists if and only if $0 < \alpha \leq 2$.

PROOF. First suppose $\alpha \geq 2$. Then $e^{-|t|^\alpha}$ is a C^2 function, with a maximum at 0. and hence if μ_α with characteristic function $e^{-|t|^\alpha}$ were to exist, it would have finite variance and zero mean. But taking variance of both sides in the identity $X + Y \overset{d}{=} 2^{1/\alpha} X$ where X, Y are i.i.d. μ_α, we see that $2\text{Var}(X) = 2^{2/\alpha}\text{Var}(X)$. Either $\text{Var}(X) = 0$, in which case $X = 0$ $a.s.$, or $\alpha = 2$, in which case $X \sim N(0, \sigma^2)$ for some $\sigma \geq 0$.

Next suppose $0 < \alpha < 2$. Recall that $X \sim \text{Pois}(\lambda)$ has characteristic function $\exp\{\lambda(e^{it} - 1)\}$, hence uX has characteristic function $\exp\{\lambda(e^{iut} - 1)\}$. Adding independent copies of such variables that $\exp\{\sum_{j=1}^N \lambda_j(e^{iu_jt} - 1)\}$ is also a characteristic function for $u_j \in \mathbb{R}$ and $\lambda_j > 0$. As a special case, take $\pm u_j$ with equal weight λ_j to get the characteristic function $\exp\{\sum_{j=1}^N \lambda_j(2\cos(u_jt) -$

$2)\}$. Taking Riemann sum approximations to the integral and Lévy's continuity theorem, we see that for any continuous function $\lambda(\cdot)$

$$\exp\left\{\int_0^\infty (\cos(ut)-1)\lambda(u)du\right\}$$

is a characteristic function. Of course, we need the integral inside the exponent to make sense and be the limit of its Riemann sums. One example is $\lambda(u) = u^{-\alpha-1}$. Integrability near ∞ forces $\alpha > 0$ and integrability near 0 forces $\alpha < 2$. On the other hand, if $I(t) = \int_0^\infty (\cos(ut)-1)u^{-\alpha-1}du$, then by a change of variables $I(t) = t^\alpha I(1)$. This proves that $\exp\{-|t|^\alpha\}$ is a characteristic function for $0 < \alpha < 2$. ■

Henceforth, we write μ_α for the symmetric α-stable distribution with characteristic function $\exp\left\{\int_0^\infty (\cos(ut)-1)\alpha u^{-\alpha-1}du\right\}$ (which is $e^{-c_\alpha |t|^\alpha}$ for some c_α that we don't care to evaluate). These distributions are heavy tailed. The proof above in fact shows that none of them (except $\alpha = 2$) can have finite variance.

Theorem 42: Moments of symmetric stable distributions

Let $0 < \alpha < 2$. Then $\int |x|^p d\mu_\alpha(x) < \infty$ if $p < \alpha$ and $\int |x|^p d\mu_\alpha(x) = \infty$ if $p > \alpha$.

PROOF. In the chapter on characteristic functions in the appendix, the following estimate is proved:

$$\mu([-2M, 2M]^c) \le M \int_{-1/M}^{1/M} (1-\hat{\mu}(t))dt.$$

Applying this to μ_α and using the fact that $1 - e^{-|t|^\alpha} \sim |t|^\alpha$ as $t \to 0$, we get $\mu_\alpha([-2M, 2M]^c) \le CM \times \frac{1}{M^{1+\alpha}} = CM^{-\alpha}$. Now,

$$\begin{aligned}\int |x|^p d\mu_\alpha(x) &= \int_0^\infty \mu_\alpha\{|x|^p > t\}dt \\ &\le C(1 + \int_1^\infty t^{-\alpha/p}dt)\end{aligned}$$

which is finite if $p < \alpha$.

To write: Proof that moments above α do not exist ■

Domains of attraction of symmetric stable distributions Let μ_α be the symmetric α-stable distribution with characteristic function $e^{-|t|^\alpha}$, where $0 < \alpha < 2$. If $X_i \sim \mu_\alpha$, then it is easy to see that $S_n/n^{1/\alpha}$ has the same distribution as X_1, in particular $n^{-\frac{1}{\alpha}} S_n \overset{d}{\to} \mu_\alpha$. The question is, what are the other distributions for which S_n (with the same scaling or different) have the same limit. For $\alpha = 2$, all we needed for the CLT was that X_i have zero mean and unit variance.

We stick to symmetric distributions here. Nevertheless, it is not sufficient to ask for X_i to have finite moments of order up to α and infinite moments beyond. A certain regularity in the tail behaviour of the distribution is needed. The regularity is stated in terms of the important concept of *slowly varying functions*. We say that $L : (0,\infty) \to (0,\infty)$ is slowly varying if $\frac{L(at)}{L(t)} \to 1$ as $t \to \infty$, for any $a > 0$. Examples are $\log t$, powers and iterates of logarithm. Observe that t^{ε} is not slowly varying if $\varepsilon \neq 0$.

Theorem 43: Convergence to symmetric stable distributions

Let X_i be i.i.d. with symmetric distribution μ. Assume that $t^{\alpha}\mu([-t,t]^c)$ is a slowly varying function. Define $b(u) = \inf\{t : \mu([-t,t]^c) = u\}$. Then

$$\frac{S_n}{b(1/n)} \xrightarrow{d} \mu_{\alpha}.$$

What is the scaling $b(1/n)$ here? If $\mu([-t,t]^c) \sim Ct^{-\alpha}$, then $b(1/n) \asymp n^{1/\alpha}$. But if $\mu([-t,t]^c) \sim Ct^{-\alpha}\log t$, then $b(1/n) \asymp n^{\frac{1}{\alpha}}(\log n)^{\frac{1}{\alpha}}$ and if $\mu([-t,t]^c) \sim Ct^{-\alpha}/\log t$, then $b(1/n) \asymp n^{\frac{1}{\alpha}}(\log n)^{-\frac{1}{\alpha}}$. Thus the exact scaling depends on the correction to $t^{-\alpha}$ in the tail of μ. The limit distribution does not.

The proof of the above theorem requires another limit theorem that is of fundamental importance in itself.

7.1. Poisson limit theorems. We know that $\text{Bin}(n,\lambda/n) \xrightarrow{d} \text{Pois}(\lambda)$ as $n \to \infty$. Like the de Moivre Laplace theorem, this is just a baby version of a rather widespread phenomenon. Here is one particular version of it.

Theorem 44: Poisson convergence of sums of independent Bernoullis

Let $\xi_{n,j} \sim \text{Ber}(p_{n,j})$, $1 \le j \le n$, be a triangular array of Bernoulli random variables such that (1) For each n, the variables $\xi_{n,1},\ldots,\xi_{n,n}$ are independent, (2) $p_{n,1} + \ldots + p_{n,n} \to \lambda$ as $n \to \infty$, (3) $p_n^* := \max_{j \le n} p_{n,j} \to 0$ as $n \to \infty$. Then $S_n := \xi_{n,1} + \ldots + \xi_{n,n}$ converges in distribution to $\text{Pois}(\lambda)$.

PROOF. By a direct calculation,

$$\begin{aligned}\mathbf{P}\{S_n = \ell\} &= \sum_{j_1 < \ldots < j_\ell \le n} \prod_{i=1}^{\ell} p_{n,j_i} \prod_{i \notin \{j_1,\ldots,j_\ell\}} (1 - p_{n,j}) \\ &= \prod_{i=1}^{n}(1 - p_{n,j}) \sum_{j_1 < \ldots < j_\ell} \prod_{r=1}^{\ell} \frac{p_{n,j_r}}{1 - p_{n,j_r}}.\end{aligned}$$

From the inequality $e^{-x} \geq 1-x \geq e^{-x-x^2}$ (valid when $|x| \leq \frac{1}{2}$), for large enough n,

$$e^{-\sum_{j=1}^n (p_{n,j}+p_{n,j}^2)} \leq \prod_{i=1}^n (1-p_{n,j}) \leq e^{-\sum_{j=1}^n p_{n,j}},$$

$$e^{p_n^*} \leq \frac{1}{1-p_{n,j_r}} \leq e^{p_n^*(1+p_n^*)}.$$

Thus,

$$e^{-\sum_{j=1}^n (p_{n,j}+p_{n,j}^2)} e^{p_n^*} \sum_{j_1<\ldots<j_\ell} \prod_{r=1}^{\ell} p_{n,j_r} \leq \mathbf{P}\{S_n = \ell\} \leq e^{-\sum_{j=1}^n p_{n,j}} e^{p_n^*(1+p_n^*)} \sum_{j_1<\ldots<j_\ell} \prod_{r=1}^{\ell} p_{n,j_r}$$

Now, $\sum_{j=1}^n p_{n,j} \to \lambda$ and $\sum_{j=1}^n p_{n,j}^2 \leq p_n^* \sum_{j=1}^n p_{n,j} \to 0$. Thus the exponential factors outside the sum on both left and right converge to $e^{-\lambda}$. Further,

$$\sum_{j_1<\ldots<j_\ell} \prod_{r=1}^{\ell} p_{n,j_r} = \frac{1}{\ell!}\left(\left(\sum_{j=1}^n p_{n,j}\right)^{\ell} - \sum_{j_1,\ldots,j_\ell}^{*} \prod_{r=1}^{\ell} p_{n,j_r}\right)$$

where the second sum is over tuples $(j_1, \ldots, j_\ell)$ of which at least two are equal. The first term inside the brackets converges to λ^ℓ. As

$$\sum_{j_1=j_2} \prod_{r=1}^{\ell} p_{n,j_r} \leq \left(\sum_j p_{n,j}^2\right)\left(\sum_j p_{n,j}\right)^{\ell-2} \to 0,$$

and the same is true of the other $\binom{\ell}{2}$ possible pairs of equal (j_r, j_s), we conclude that

$$\sum_{j_1<\ldots<j_\ell} \prod_{r=1}^{\ell} p_{n,j_r} \to \frac{1}{\ell!}\lambda^\ell.$$

In summary, $\mathbf{P}\{S_n = \ell\} \to e^{-\lambda}\frac{\lambda^\ell}{\ell!}$ for $\ell \in \mathbb{N}$, and thus $S_n \xrightarrow{d} \text{Pois}(\lambda)$. ■

ALTERNATE PROOF. For $t \in \mathbb{R}$,

$$\mathbf{E}[e^{itS_n}] = \prod_{k=1}^n (1 - p_{n,j} + p_{n,j}e^{it}).$$

By Exercise 21,

$$\begin{aligned}\left|\mathbf{E}[e^{itS_n}] - \prod_{j=1}^n e^{-p_{n,j}+p_{n,j}e^{it}}\right| &\leq \sum_{j=1}^n \left|e^{-p_{n,j}+p_{n,j}e^{it}} - (1-p_{n,j}+p_{n,j}e^{it})\right| \\ &\leq C\sum_{j=1}^n p_{n,j}^2\end{aligned}$$

which converges to 0. As $\prod_{j=1}^n e^{-p_{n,j}+p_{n,j}e^{it}} \to e^{-\lambda+\lambda e^{it}}$, which is the characteristic function of Pois(λ), we see that $S_n \xrightarrow{d} \text{Pois}(\lambda)$. ■

7.2. Proof of Theorem 43. The proof is very different from all the proofs of central limit theorem, because the underlying phenomena are themselves different. In CLT, all the variables contribute about the same, but for the heavy tailed variables under consideration, the sum S_n essentially comes from the largest few X_is.

For example, if $\mathbf{P}\{X_1 \geq x\} \sim Cx^{-\alpha}$, then the expected number of $X_1, \ldots, X_n$ that are above x is $Cnx^{-\alpha}$, which shows that the maximum $M_n = \max\{X_1, \ldots, X_n\}$ is not likely to be significantly more than $n^{1/\alpha}$. By the second moment method, one can show that M_n is of the order of $n^{1/\alpha}$, which is also the order of magnitude of S_n (as the statement of Theorem 43 asserts). Contrast this with the Gaussian case, where the maximum is of order $\sqrt{\log n}$ while the sum is of order $\sqrt{n}$.

First we prove a Theorem that is in the same spirit as Theorem 43, but technically much simpler.

Theorem 45: Poissonized version of convergence to symmetric stable distributions

Let X_i be i.i.d. with symmetric distribution μ and let $K_n \sim \text{Pois}(n)$ be independent of X_is. Assume that $t^\alpha \mu([-t,t]^c)$ is a slowly varying function. Define $b(u) = \inf\{t : \mu([-t,t]^c) = u\}$. Then

$$\frac{S_{K_n}}{b(1/n)} \overset{d}{\to} \mu_\alpha.$$

Proof. The advantage of considering S_{K_n} instead of S_n is that its characteristic function can be written in a form similar to that of μ_α. Define the measure μ_n by $\mu_n(J) = 2n\mu(a_n J)$ for $J \in \mathcal{B}_\mathbb{R}$ and let $a_n = b(1/n)$. We claim that

$$\mathbf{E}\left[e^{itS_{K_n}/a_n}\right] = \exp\left\{\int_0^\infty (\cos(tu) - 1) d\mu_n(u)\right\}. \tag{25}$$

To see this[2], let $M_n = \delta_{X_1/a_n} + \ldots + \delta_{X_{K_n}/a_n}$, a random measure, in terms of which $a_n^{-1}S_{K_n} = \int t dM_n(t)$. For $\delta > 0$, let $I_{j,\delta} = (j\delta, (j+1)\delta]$ and $\varphi_\delta = \sum_{j\geq 1} j\delta(\mathbf{1}_{I_{j,\delta}} - \mathbf{1}_{-I_{j,\delta}})$. Then $\varphi_\delta(t) \to t$ as $\delta \downarrow 0$, and $|\varphi_\delta(t)| \leq t$. Hence, by DCT,

$$\frac{S_{K_n}}{a_n} = \lim_{\delta \downarrow 0} \sum_{j=1}^\infty j\delta M_n(I_{j,\delta}) - \sum_{j=1}^\infty j\delta M_n(-I_{j,\delta}) \quad a.s.$$

If $J_1, \ldots, J_k$ are pairwise disjoint intervals, then $M_n(J_1), \ldots, M_n(J_k)$ are independent random variables with $M_n(J) \sim \text{Pois}(n\mu(a_n J))$. This is a well-known fact about thinning of Poissons. Thus, for fixed $\delta > 0$, the quantity on the right is a weighted sum of independent Poisson random

[2] If you are familiar with Poisson processes, it is possible to see this formula and nod "yes, it is obvious". The explanation given is for those who did not nod.

variables, hence it has characteristic function (using the symmetry $\mu(I_{j,\delta}) = \mu(-I_{j,\delta})$)

$$\exp\left\{\sum_{j=1}^{\infty} n\mu(I_{j,\delta})(e^{itj\delta} + e^{-itj\delta} - 1)\right\} = \exp\left\{\sum_{j=1}^{\infty} 2n\mu(I_{j,\delta})(\cos(j\delta) - 1)\right\}.$$

The exponent is $2\int_0^\infty (\cos(\varphi_\delta(t)) - 1)d\mu_n(t)$, hence it converges to $2\int_0^\infty(\cos t - 1)d\mu_n(t)$ by another application of DCT. This proves (25).

Now we need to let $n \to \infty$. For any $s > 0$,

$$\mu_n[s,\infty) = n\mu[a_n,\infty) \times \frac{\mu[sa_n,\infty)}{\mu[a_n,\infty)} \to \frac{1}{2s^\alpha}$$

as $n\mu[a_n,\infty) = 1/2$ by choice of a_n, and using the fact that $s^\alpha\mu[sa_n,\infty)$ is slowly varying. This is almost like saying that μ_n (restricted to $(0,\infty)$) converges in distribution to the measure $\frac{1}{2}\alpha s^{-\alpha-1}ds$. However the limiting measure here is infinite, and hence we need to justify that

$$2\int_0^\infty (\cos t - 1)d\mu_n(t) \to \int_0^\infty (\cos t - 1)\frac{\alpha}{t^{\alpha+1}}dt. \tag{26}$$

Once we justify (26), the proof is complete, as it shows that the characteristic function of $S_{K_n}/n^{1/\alpha}$ converges pointwise to the characteristic function of μ_α (refer back to the definition of μ_α). ■

To justify (26), we fix $\varepsilon > 0$ and divide the integral over $(0,\varepsilon)$, $[\varepsilon, 1/\varepsilon]$ and $(1/\varepsilon, \infty)$. Since the limiting integral is convergent, we can choose ε small enough to make the first and third integrals smaller than ε. On $[\varepsilon, 1/\varepsilon]$, the measures are finite, and we can scale and pretend that we are working with probability measures to conclude that (we leave the details as exercise)

$$2\int_\varepsilon^{1/\varepsilon} (\cos t - 1)d\mu_n(t) \to \int_\varepsilon^{1/\varepsilon} (\cos t - 1)\frac{\alpha}{t^{\alpha+1}}dt.$$

It only remains to show that the first and third integrals can be made arbitrarily small uniformly over n, by choosing ε small. As the integrand is bounded by 2, the third integral is bounded by

$$4\mu_n[1/\varepsilon,\infty) = 4n\mu[a_n,\infty)\frac{\mu[a_n/\varepsilon,\infty)}{\mu[a_n,\infty)} \sim 2\varepsilon^\alpha$$

by the same argument that we used above. This shows that the third integral can be made uniformly small by choosing ε small enough. The first integral is to complete

CHAPTER 7

Appendix: Characteristic functions and their properties

Definition 14

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, which we shall discuss and prove one by one[1].

(A) It transforms well under certain operations, such as shifting, scaling and under convolutions. The last of these makes it a tool of amazing power in studying sums of independent random variables.

(B) The characteristic function determines the measure. Further, the smoothness of the characteristic function encodes the tail decay of the measure, and vice versa. In general, c.f. encodes properties of the distribution in a not-so-direct but still tractable manner.

(C) $\hat{\mu}_n(t) \to \hat{\mu}(t)$ pointwise, if and only if $\mu_n \overset{d}{\to} \mu$. The forward implication is the key property that is 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. of a measure. Because of this and part (B), sometimes one defines a measure by its characteristic function.

0.1. Basic observations.

[1] 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 46

Let X, Y be random variables with distributions μ, ν respectively.

(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 47

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

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

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

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. ■

Lemma 48: Parseval's identity

If $\mu, \nu \in \mathcal{P}(\mathbb{R})$, then $\int \hat{\mu}\, d\nu = \int \hat{\nu}\, d\mu$.

PROOF. Integrate e^{ixy} against $\mu \otimes \nu$ in two ways, using Fubini's theorem. The two iterated integrals are $\int \hat{\mu} d\nu$ and $\int \hat{\nu} d\mu$. ■

0.2. Decay and smoothness. Smoothness of the characteristic function is related to the tail decay of the measure and smoothness of the measure is related to the tail decay of the characteristic function. We give some statements illustrating all four directions of implication.

Theorem 49: Decay of the measure to smoothness of Fourier transform

Let $\mu \in \mathcal{P}(\mathbb{R})$. If $\int |x|^k d\mu(x) < \infty$ for some $k \in \mathbb{N}$, then $\hat{\mu} \in C^{(k)}(\mathbb{R})$ and

$$\hat{\mu}^{(k)}(t) = \int_{\mathbb{R}} (ix)^k e^{itx} d\mu(x).$$

PROOF. It is a matter of justifying the differentiation w.r.t. t under the integral $\hat{\mu}(t) = \int e^{itx} d\mu(x)$. We show it for $k = 1$ and leave the rest as an exercise. As $h^{-1}(e^{i(t+h)x} - e^{itx}) \to ixe^{itx}$ as $h \to 0$

and $h^{-1}|e^{i(t+h)x} - e^{itx}| \le |x|$ by mean value theorem, if $\int |x| d\mu(x) < \infty$ then DCT justifies

$$\lim_{h\to 0} \frac{1}{h} \int (e^{i(t+h)x} - e^{itx}) d\mu(x) = \int ixe^{itx} d\mu(x)$$

which is the same as $\hat{\mu}'(t) = \int ixe^{itx} d\mu(x)$. ■

In fact, by expanding e^{itx} in finite order Taylor expansion and applying expectations, one can write the Taylor expansion for $\hat{\mu}$ with coefficient given by moments of μ.

Theorem 50: Smoothness of measure to decay of Fourier transform

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 $t \to \pm\infty$.

PROOF. First assume that f is smooth and that its derivatives are also integrable (and hence vanish at infinity). Then, integrating by parts, we get

$$\hat{\mu}(t) = -\int \frac{1}{it} e^{itx} f'(x) dx$$

which is bounded by $\frac{1}{|t|}\|f\|_{L^1(\mathbb{R})}$. This goes to 0 as $|t| \to \infty$. In general, we can approximate f by a smooth g whose derivatives are integrable so that $\|f - g\|_{L^1(\mathbb{R})} \le \varepsilon$. Then $\|\hat{f} - \hat{g}\|_{\sup} \le \varepsilon$ (we use $\hat{f}(t)$ for $\int f(x)e^{itx}dx$). Therefore,

$$\limsup_{t\to\pm\infty} |\hat{f}(t)| \le \limsup_{t\to\pm\infty} |\hat{g}(t)| + \varepsilon \; = \; \varepsilon$$

as $\hat{g}(t) \to 0$. This completes the proof of the first part.

Observe that the positivity of f was not used, only its integrability. Hence if f is k times differentiable and $f^{(k)} \in L^1(\mathbb{R})$, then $\widehat{f^{(k)}}(t) = o(1)$ as $t \to \pm\infty$. Now, integrating by parts we see that $\hat{f}(t) = (-i/t)^k \widehat{f^{(k)}}(t)$, which is $o(t^{-k})$. ■

Theorem 51: Smoothness of the characteristic function to the decay of the measure

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

$$\mu([-2M, 2M]^c) \le M \int_{-M}^{M} (1 - \hat{\mu}(t)) dt.$$

PROOF. Let $\delta = 1/M$ and write

$$\int_{-\delta}^{\delta} (1 - \hat{\mu}(t))\, dt = \int_{-\delta}^{\delta}\int_{\mathbb{R}} (1 - e^{itx})\, d\mu(x)\, dt \; = \; \int_{\mathbb{R}}\int_{-\delta}^{\delta} (1 - e^{itx})\, dt\, d\mu(x)$$

$$= \int_{\mathbb{R}} \left(2\delta - \frac{2\sin(x\delta)}{x}\right) d\mu(x) \; = \; 2\delta \int_{\mathbb{R}} \left(1 - \frac{\sin(x\delta)}{x\delta}\right) d\mu(x).$$

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

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

This is the claim. ■

Theorem 52: Decay of the Fourier transform to the smoothness of the measure

If $\hat{\mu} \in L^1(\mathbb{R})$, then μ has a bounded continuous density f given by

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

If further $t^k\hat{\mu}(t)$ is integrable over $\mathbb{R}$, then f is k times differentiable.

The first part is proved below under the heading of Fourier inversion formula. Once that is proved, we have essentially express f as the Fourier transform of $\hat{\mu}$ (except for the negative sign in the exponent and the factor of $1/2\pi$). Hence, the earlier proof, where we showed that if the kth moment is finite, then the characteristic function is k times differentiable, applies here with $\hat{\mu}(t)dt$ taking the place of the measure.

0.3. 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 $\hat{\mu}(t) = p_1e^{ita_1} + \ldots + p_ke^{ita_k}$.

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

(3) Let $X, X' \sim$ unif$[-1,1]$ be independent and let $Y = X + X'$. The density of X is $\frac{1}{2}$ on $[-1,1]$ while that of Y is $\frac{1}{2}(1-\frac{1}{2}|x|)$ for $|x| \le 2$. The characteristic function of X is easily computed to be $\sin t/t$ and hence the characteristic function of Y is $(\sin t/t)^2$.

(4) The characteristic function of Pois(λ) distribution is

$$\sum_{k\ge 0} e^{ikt}e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda+\lambda e^{it}}.$$

(5) 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 $Y \overset{d}{=} X_1 + \ldots + X_\nu$ where X_k are i.i.d Exp(λ). Therefore, $\psi_Y(t) = \frac{\lambda^\nu}{(\lambda-it)^\nu}$. This is true even if ν is not an integer, but the proof would have to be a direct computation.

(6) 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.

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

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

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

The final conclusion is that $N(\mu, \sigma^2)$ has characteristic function $e^{it\mu - \frac{\sigma^2 t^2}{2}}$. We gave an alternate rigorous proof using Stein's identity in the notes.

(8) 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 $\frac{e^{-t\,\text{Im}\,z}}{|1+z^2|} \le \frac{1}{R^2-1}$, 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 $\psi(t) = e^{-|t|}$.

0.4. Inversion formulas.

Theorem 53

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

Proof. Let θ_σ denote the $\text{N}(0, \sigma^2)$ 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,

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

[2] 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 $R \to \infty$.

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

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

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

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

Further, $\Phi_{\frac{1}{\sigma}}$ 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 54: Fourier inversion formula

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

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

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

(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 (1) is equal to $\lim\limits_{\sigma\to\infty} \int_a^b f_\sigma$ 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 53 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!!

0.5. 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 55: Lévy's continuity theorem

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

(1) If $\mu_n \overset{d}{\to} \mu$ 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 $\mu_n \overset{d}{\to} \mu$.

Observe that in the second statement, we did not a priori assume that ψ is a characteristic function. It of course implies that if $\hat{\mu}_n \to \hat{\mu}$ pointwise for some $\mu \in \mathcal{P}(\mathbb{R})$, then $\mu_n \overset{d}{\to} \mu$.

PROOF. (1) If $\mu_n \overset{d}{\to} \mu$, 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 $\mu_n \overset{d}{\to} \mu$.

It remains to show tightness[3]. From Lemma 56 below, as $n \to \infty$,

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

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. ■

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

Lemma 56

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

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

PROOF. We write

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

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

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

From the continuity theorem, it follows that if $\hat{\mu}_n$ converge to a continuous function, then the limit is a characteristic function too. Here is an application of this.

0.6. 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

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

This motivates the following definition.

Definition 15: 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 57: 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 $|\varphi| \le 1$. Indeed, for any t, the positive semi-definiteness of $M_\varphi[0,t]$ shows that $1 - |\varphi(t)|^2 \ge 0$ (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 $Z_i = X_i Y_i$, $1 \le i \le n$, is precisely C. Hence C is positive semi-definite.
- For any nice function $c : \mathbb{R} \mapsto \mathbb{C}$, we have

$$\iint c(t)\overline{c(s)}\varphi(t-s)dtds \ge 0. \tag{2}$$

  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 the natural one. 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 continuous, positive-definite and $\varphi(0) = 1$.

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

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

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[4]. Fix an interval $I_M = [-M, M]$ and observe that

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

The first integral is positive by (2) (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

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

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. We now claim that $\int f(x)dx = 1$. To start with, since $|f| \le \|\varphi\|_1$, for any $\sigma > 0$ we have

$$\begin{aligned} \int_{\mathbb{R}} f(x)e^{-\sigma^2x^2/2}dx &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \varphi(t)e^{ixt}e^{-\sigma^2x^2/2}dx\ dt \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \varphi(t)\int_{\mathbb{R}} e^{ixt}e^{-\sigma^2x^2/2}dt\ dx \end{aligned}$$

where the application of Fubini's theorem is justified because $|\varphi(t)|e^{-\sigma^2x^2/2} \in L^1(\mathbb{R}\times\mathbb{R})$. The inner integral is essentially the Fourier transform of the Gaussian and equal to $\sqrt{2\pi}e^{-\frac{t^2}{2\sigma^2}}$. Plugging this in, we see that

$$\int_{\mathbb{R}} f(x)e^{-\sigma^2x^2/2}dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} \varphi(t)e^{-\frac{t^2}{2\sigma^2}}dt$$

The right side is $\mathbf{E}[\varphi(\sigma Z)]$ where $Z \sim N(0,1)$. By continuity and boundedness of φ, DCT implies that it converges to $\varphi(0) = 1$ as $\sigma \downarrow 0$. The integrand on the left side increases (as $f \ge 0$) to $f(x)$.

[4] It may be easier to first see the following formal argument. Fix $x \in \mathbb{R}$ and use $c(t) = e^{ixt}$ in (2) to get

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

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.

hence by MCT, the limit as $\sigma \downarrow 0$ of the integral is $\int_{\mathbb{R}} f(x)dx$. This shows that f is a probability density.

As $f$is integrable, the Fourier inversion formula applies to show that $\int_{\mathbb{R}} f(x)e^{-itx}dx = \varphi(t)$ for all t. Thus, φ is the characteristic function of the probability measure $f(x)dx$.

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, we see that $\mu_\sigma \xrightarrow{d} \mu$ as $\sigma \downarrow 0$ for some $\mu \in \mathcal{P}(\mathbb{R})$ and that $\hat{\mu} = \varphi$. ■

0.7. Multivariate situation. Let $X \sim \mu \in \mathcal{P}(\mathbb{R}^d)$. Its Fourier transform or characteristic function is a function $\hat{\mu} : \mathbb{R}^d \to \mathbb{C}$ defined as $\hat{\mu}(t) = \int e^{i\langle t,x\rangle}d\mu(x) = \mathbf{E}[e^{i\langle t,X\rangle}]$. All the theorems proved in the univariate case go through with the most obvious modifications. In particular, we have

(1) Parseval relation: $\int_{\mathbb{R}^d} \hat{\mu}d\nu = \int_{\mathbb{R}^d} \hat{\nu}d\mu$.

(2) Fourier inversion formula: If $\hat{\mu} = \hat{\nu}$, then $\mu = \nu$. In particular, if $\hat{\mu}$ is integrable, then μ has bounded continuous density given by $f(x) = (2\pi)^{-d}\int_{\mathbb{R}^d} \hat{\mu}(t)e^{i\langle t,x\rangle}dt$.

(3) Lévy's continuity theorem: Identical to the one-dimensional case.

(4) Joint moments of X_is are related to partial derivatives of the characteristic function at the origin.

And these tools can be used to prove CLT just as before.

Remark 9

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.