Probability theory

Manjunath Krishnapur

To the children of Gaza

Tears for the killed

Forlorn hopes for the surviving

(Dec 2024)

Contents

CHAPTER 1

Probability measures and random variables

The goal of the course is to first understand the measure theoretic foundations of probability theory. Then we study the basic theorems and techniques of probability. Although we shall introduce many interesting probability situations, what we study most in depth are sums of independent random variables, essentially devoted to the following single question: Given independent random variables with known distributions, what can be said about the distribution of the sum? Here is a brief outline, some of which will not make sense till we go into them in detail.

▶ Measure theoretic foundations of probability: Borel and Lebesgue founded measure theory, in particular the Lebesgue measure and Lebesgue integral, mainly motivated by the question of understanding lengths, areas, volumes, in the greatest generality possible. Very soon it was realized that the same would also give a mathematical foundation to probability. However, beyond the basics, the key aspects of probability that make it richer than general measure theory are the notions of independence and conditional probability. Several analysts had given reasonably satisfactory measure theoretic foundation to independence (most importantly Daniell) but it was Kolmogorov who put the entire theory, including conditional probability, on a firm foundation. We shall see most of this in the first part of the course (though conditional probability will be largely postponed).

▶ 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 or Fourier transforms, a tool of great importance, as well as the less profound but very common and useful techniques of proofs such as truncation and approximation. It is these techniques that one must really get comfortable with, to be able to do anything further.

▶ Thirdly, we 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]. These are also important to introduce, as techniques cannot be learned in vacuum, but by seeing how they are used in these problems.

► Lastly, some of the fundamental results of probability theory. Laws of large numbers, Central limit theorems, Law of iterated logarithm, Sums of heavy tailed random variables, etc. Their importance cannot be overemphasized.

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.

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

Definition 1: Discrete probability space

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

The set Ω is called the *sample space* (the collection of all possible outcomes), $p(\omega)$ are called *elementary probabilities*, subsets of Ω are called *events*, and $\mathbb{P}(A)$ 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*, a notion to be introduced later) where $p \in [0,1]$ is a parameter describing the coin. An example of an event is that of getting exactly k heads, i.e., $A = \{\omega : \omega_1 + \ldots + \omega_n = k\}$, which has probability $\mathbb{P}(A) = \binom{n}{k} p^k (1-p)^{n-k}$.

[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.

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 $\mathbb{P}(A) = \frac{(n-1)^r}{n^r}$.

But when the number of possible outcomes is uncountable, this framework fails. 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 dart board.

Sample space. This is the set of all possible outcomes and is denoted[2] Ω. In the above cases it is easy to see that the sample space must be equal to

(1) $[0, 1]$, where we think of the glass rod as the line segment $[0, 1]$ and the outcome denotes 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) $\{(x, y) : x^2 + y^2 \leqslant 1\}$, 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 the events $[0.1, 0.35]$ and $\{\omega \in \{0, 1\}^{\mathbb{N}} : \omega_1 = 1, \omega_2 = 0\}$ and $\{(x, y) : x > 0 > y\}$ in the three examples must have probability $\frac{1}{4}$. But where does that come from? If any elementary probabilities are 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.

[2]This is universal among probabilists of the world. If you use a different letter for the sample space, you will be looked at with concern, but if you use Ω for anything else in probability context, no one will talk to you.

Definition 2: Probability space

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

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

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 $\mathbb{P}(A)$ 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 should the sigma-algebra be? Except for the trivial sigma-algebras 2^{Ω} and $\{\emptyset, \Omega\}$, there is no sigma-algebra of interest that can be defined by explicitly specifying a membership criterion for which subsets of Ω belong to it. They are almost always defined indirectly 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 $\mathcal{S}$.

Arbitrary intersection of sigma-algebras is a sigma-algebra, hence $\sigma(\mathcal{S})$ is a sigma-algebra. The most important example of a generated sigma-algebra is the Borel sigma algebra of a topological space. This is the sigma-algebra generated by the collection of all open sets.

An important point to keep in mind is that many different collections $\mathcal{S}$ generate the same sigma-algebra as the following exercise shows.

Exercise 1

On $\mathbb{R}$, show that the Borel sigma-algebra is generated by any of the following collections of sets: (1) open sets, (2) closed sets, (3) compact sets, (4) intervals, (5) open intervals, (6) closed intervals, (7) open intervals with rational end-points, (8) left-open, right-closed intervals, (9) the collection of intervals $(-\infty, x]$ with $x \in \mathbb{Q}$.

Sigma-algebra. Now let us decide the sigma-algebras in the three examples we considered above. As suggested above, the sigma-algebra is defined by giving a generating set $\mathcal{S}$. How to decide what $\mathcal{S}$ to take? Into $\mathcal{S}$ we put in all subsets for which we definitely wish to define probabilities. Then take $\mathcal{F} = \sigma(\mathcal{S})$ as our sigma-algebra.

In all three examples we take $\mathcal{S}$ to be the collection of open sets in Ω, and so $\mathcal{F} = \mathcal{B}_\Omega$ is the Borel sigma-algebra. The topology in the first example ($\Omega = [0,1]$) and third example ($\Omega = \{(x,y) : x^2 + y^2 \leqslant 1\}$) are the standard topologies coming from $\mathbb{R}^d$. In the second example, we take the product topology on $\Omega = \{0,1\}^{\mathbb{N}}$, which is in fact metrized by

$$d(\omega, \omega') = \sum_{n \geqslant 1} \frac{|\omega_n - \omega'_n|}{2^n} \qquad \text{for } \omega, \omega' \in \{0,1\}^{\mathbb{N}}.$$

This certainly defines a sigma-algebra in each case, but is it the right one for us? Let us discuss this point.

For example, in the stick-breaking example, you might either worry that we are asking for too little (don't we want closed sets in our sigma-algebra?) or that we are asking for too much (do we need all open sets?). The first is not a worry because of the earlier exercise that shows that we would get the same Borel sigma-algebra from many different collections of sets, including the collection of closed sets. We are not asking for too much either. Indeed, if we ask for open intervals to be in the sigma-algebra, then all open sets must also be there (as any open set in $\mathbb{R}^d$ is a countable union of open balls).

Identical considerations apply to the dart throwing example.

The stick-breaking example looks a bit different. Introducing a metric out of the blue and taking its Borel sigma-algebra looks unnatural. More natural would have been to take $\mathcal{S}$ to be the collection of sets of the form $\{\omega : \omega_{i_1} = \varepsilon_1, \ldots, \omega_{i_m} = \varepsilon_m\}$ where $1 \leqslant i_1 < \ldots < i_m$ and $\varepsilon_i \in \{0,1\}$. Observe that these sets specify the outcome of finitely many tosses. These are called (finite-dimensional) *cylinder sets*. If $\mathcal{S}$ denotes the collection of cylinders, then the generated sigma-algebra $\sigma(\mathcal{S})$ is the same as the Borel sigma algebra of the product topology on $\{0,1\}^{\mathbb{N}}$ (can you see why?). Thus, we did take the right sigma-algebra.

The probability measure. 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, let us say that we know what probability of events in $\mathcal{S}$ ought to be (or to be pendantically clear, we include in $\mathcal{S}$ those events for which we do know what the probabilities ought to be, and then define the sigma-algebra). 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 $\mathbb{P}$ on $\sigma(\mathcal{S})$ such that $\mathbb{P}(A) = P(A)$ for $A \in \mathcal{S}$. If so, is it unique?

1.1. The uniqueness question. The uniqueness part is easier. But it is not true in general!

Example 3

Let $\Omega = \{1,2,3,4\}$ and $\mathcal{S} = \{\{1,2\},\{2,3\},\{3,4\}\}$. Then $\sigma(\mathcal{S}) = 2^\Omega$. Let $p(i) = \frac{1}{4}$ for all i and let $q(i) = \frac{1}{2}$ for $i = 1,3$ and $q(i) = 0$ for $i = 2,4$. Use these as elementary probabilities to define probability measures $\mathbb{P}, \mathbb{Q}$ on 2^Ω. Then $\mathbb{P}(A) = \frac{1}{2} = \mathbb{Q}(A)$ for all$A \in \mathcal{S}$ but $\mathbb{P} \neq \mathbb{Q}$.

Uniqueness is true if we assume more structure on $\mathcal{S}$, for example if it is a π-system (closed under intersections[3]). The proof is a good illustration of one of the standard tricks of measure theory.

Proof for π-systems. Indeed, suppose $\mathcal{S}$ is a π-system and that $\mathbb{P}, \mathbb{Q}$ are two probability measures on $\sigma(\mathcal{S})$ such that $\mathbb{P}(A) = \mathbb{Q}(A)$ for $A \in \mathcal{S}$. Then the collection $\mathcal{G} := \{A \in \sigma(\mathcal{S}) : \mathbb{P}(A) = \mathbb{Q}(A)\}$ contains $\mathcal{S}$. If $A_n \in \mathcal{G}$ and $A_n \uparrow A$, then $A \in \mathcal{S}$ because

$$\mathbb{P}(A) = \lim \mathbb{P}(A_n) = \lim \mathbb{Q}(A_n) = \mathbb{Q}(A).$$

Further, if $A, B \in \mathcal{G}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{G}$ because

$$\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A) = \mathbb{Q}(B) - \mathbb{Q}(A) = \mathbb{Q}(B \setminus A).$$

As $\mathcal{G}$ contains the empty set and Ω, this shows that it is a λ-system. It contains the π-system $\mathcal{S}$. The π-λ theorem asserts that then $\mathcal{G}$ contains $\sigma(\mathcal{S})$, which means that $\mathbb{P} = \mathbb{Q}$ on $\sigma(\mathcal{S})$. ■

The standard trick referred to above is in the consideration of $\mathcal{G}$, the collection of all sets with the property that we wish to show for all sets of $\sigma(\mathcal{S})$. In fact, there is no other way, as the only definition of $\sigma(\mathcal{S})$ is as the smallest sigma-algebra containing $\mathcal{S}$ (there is no way to write elements of $\sigma(\mathcal{S})$ using countable operations on elements of $\mathcal{S}$).

1.2. The existence question. The existence of a measure is the harder question. But it has a clean and efficient answer!

[3]For measure theory terms and facts not explained in detail here, see my other notes.

Theorem 1: Carathéodory's extension theorem

Let $\mathcal{A}$ be an algebra of subsets of Ω and let $\mathcal{F} = \sigma(\mathcal{A})$. Let $P : \mathcal{A} \to [0,1]$ satisfy (i) $P(A \cup B) = P(A) + P(B)$ if $A, B \in \mathcal{A}$ are disjoint, (ii) if $A_n, A \in \mathcal{A}$ and $A_n \uparrow A$, then $P(A_n) \uparrow P(A)$, (iii) $P(\Omega) = 1$. Then, there exists a probability measure $\mathbb{P}$ on $\mathcal{F}$ such that $\mathbb{P} = P$ on $\mathcal{A}$.

Observe that the assumed conditions are obviously necessary. It is amazing that they are sufficient! In practise, the hardest part of the checking is the countable additivity on $\mathcal{A}$ (second condition). But it can be done in many situations of interest including that of Lebesgue measure. The proof of Carathéodory's extension theorem can be found in every book on measure theory. A one-paragraph summary: One defines the *outer measure* $P^* : 2^\Omega \to [0,1]$ by

$$P^*(A) := \inf \left\{ \sum_n P(A_n) : A_n \in \mathcal{A},\ \cup_n A_n \supseteq A \right\}.$$

It turns out that $P^* = P$ on $\mathcal{A}$ and P^* is countably sub-additive on 2^Ω (i.e., $P^*(\cup_n A_n) \leqslant \sum_n P^*(A_n)$). However, there is a sigma-algebra $\overline{\mathcal{F}}$ (defined by the "Carathéodory cut condition") that contains $\mathcal{F}$ and on which P_* is countably additive. The restriction of P to $\mathcal{F}$ is the $\mathbb{P}$ we want.

Remark 1

In the uniqueness part, we only needed the generating set to be a π-system whereas in the existence part we needed it to be an algebra. It would have been more convenient to just start with a π-system (much less structure than an algebra). This can be done occasionally by finding π-systems $\mathcal{S}$ with the property that the complement of any set in $\mathcal{S}$ is a finite disjoint union of elements of $\mathcal{S}$. The reason this helps is that for such $\mathcal{S}$, the collection $\mathcal{A}$ of finite disjoint unions of elements of $\mathcal{S}$ becomes an algebra. And given $P : \mathcal{S} \to [0,1]$, it is clear that for $A = S_1 \sqcup \ldots \sqcup S_k \in \mathcal{A}$, we must take $P(A)$ to be $P(S_1) + \ldots + P(S_k)$. Thus we first extend P to $\mathcal{A}$, and then check the conditions of Carathéodory's extension theorem.

The existence of Lebesgue measure is a special case of paramount importance.

Theorem 2: Existence of Lebesgue measure

There exists a unique probability measure λ on the Borel sigma-algebra $\mathcal{B}$ of $[0,1]$ such that $\lambda([a,b]) = b - a$ whenever $[a,b] \subseteq (0,1]$.

Proof. Observe that $\mathcal{S} = \{(a,b] : 0 \leqslant a < b \leqslant 1\}$ is a π-system. It has the special property mentioned in the above remark: $(a,b]^c = (0,a] \sqcup (b,1]$ is a disjoint union of two elements of $\mathcal{S}$. Hence $\mathcal{A} = \{I_1 \sqcup \ldots \sqcup I_k : k \geqslant 0,\ I_j = (a_j, b_j],\ b_j < a_{j+1} \text{ for all } j\}$ is an algebra. For $A = I_1 \sqcup \ldots \sqcup I_k$

with $I_j = (a_j, b_j]$ pairwise disjoint, we define

$$P(A) = \sum_{j=1}^{k} (b_j - a_j).$$

The finite additivity of P on $\mathcal{A}$ is easy to check and left as exercise. To apply Carathéodory's extension theorem, it only remains to check that if $A_n, A \in \mathcal{A}$ and $A_n \uparrow A$, then $P(A_n) \uparrow P(A)$, or equivalently that $P(A \backslash A_n) \downarrow 0$. Let $A = J_1 \sqcup \ldots \sqcup J_m$ where $J_i = (a_i, b_i]$. Then $P(A \backslash A_n) \leqslant \sum_{i=1}^{k} P(J_i \backslash A_{n,i})$ where $A_{n,i} = A_n \cap J_i \in \mathcal{A}$.

Thus, it suffices to show that $P(J \backslash B_n) \downarrow 0$ for any $J = (a, b]$ and $B_n \in \mathcal{A}$ such that $B_n \uparrow J$. Replace J by the smaller compact interval $J' = [a + \varepsilon, b]$ and replace the intervals in $B_n = I_{n,1} \sqcup \ldots \sqcup I_{n,k_n}$ by slightly larger open intervals (say the intervals are enlarged by ε/k_n each) to get an open set B_n'. Then $J' \setminus B_n'$ are compact sets that decrease to empty set, hence equal to empty set for some large n. But then

$$P(J) \leqslant P(J') + \varepsilon \leqslant P(B_n') + \varepsilon \leqslant P(B_n) + 2\varepsilon.$$

As ε is arbitrary, $\liminf P(B_n) \geqslant P(J)$. The other inequality $\limsup P(B_n) \leqslant P(J)$ is clear as $B_n \subseteq J$. ■

Remark 2

Do not forget to check the finite additivity of P on $\mathcal{A}$. In general, when you start with a special π-system, it is important to check that P satisfies finite and countable additivity on the generated algebra $\mathcal{A}$. For example, can we get measures on $([0,1], \mathcal{B})$ such that $\mathbb{P}(a, b] = (b - a)^2$ for all $a < b$ or $\mathbb{P}(a, b] = \sqrt{b - a}$ for all $a < b$?

2. Probability measures

One can imagine that by a similar method Carathéodory extension theorem, one can define probability spaces to capture other random experiments such as "throwing a dart" and "tossing a coin infinitely many times". For example, in the coin-tossing case, we can start with the π-system of cylinders for which we know what probabilities to assign, and proceed from there.

However, we emphasize a different point of view here. Once we have the stick-breaking probability space $([0,1], \mathcal{B}, \lambda)$, every other probability space of interest can be constructed from it! First we introduce a fundamental notion of probability theory.

Definition 4: Random variable or 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 $A \in \mathcal{G}$. We also say that T is a (Y-valued) *random variable*.

When we just say random variable, we usually mean an $\mathbb{R}$-valued random variable or preferably $\overline{\mathbb{R}}$-valued random variable, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ is the extended real line whose topology is got by identifying it with $[-1, 1]$ via the map $x \mapsto \frac{x}{1+|x|}$ (or if you prefer metric spaces, the metric on $\overline{\mathbb{R}}$ is $d(x, y) = |\frac{x}{1+|x|} - \frac{y}{1+|y|}|$). When the target space is $\mathbb{R}^d$, we talk of random vectors and depending on the target space, we may have random sets, random graphs, random measures, etc.

Push-forward to get new measures from old. The existence of Lebesgue measure λ on $([0, 1], \mathcal{B})$ solves our search for a mathematical framework for the "breaking a stick" random experiment.

Lemma 1: Push-forward measure

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sigma-algebra on Λ. Suppose $T : \Omega \to \mathcal{G}$ is a Λ-valued random variable. Then, $\mathbb{Q} : \mathcal{G} \to [0, 1]$ defined by $\mathbb{Q}(A) = \mathbb{P}(T^{-1}(A))$ is a probability measure on $(\Lambda, \mathcal{G})$. It is called the *distribution* of the random variable T.

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

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

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

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

2.1. Tossing a coin infinitely many times. Here $\Omega = \{0, 1\}^{\mathbb{N}}$ and $\mathcal{F}$ is the Borel sigma-algebra (generated by finite dimensional cyinder sets). For a cylinder set

$$A_{\varepsilon_1,\ldots,\varepsilon_n} = \{\omega = (\omega_1, \omega_2, \ldots) \in \Omega : \omega_1 = \varepsilon_1, \ldots, \omega_n = \varepsilon_n\},$$

we know that we want the probability to be 2^{-n}.

Define $T : [0, 1] \to \{0, 1\}^{\mathbb{N}}$ by $T(x) = (x_1, x_2, \ldots)$ where $x = \sum_{n \geqslant 1} x_n 2^{-n}$ is the binary expansion of x. To avoid ambiguity, for dyadic rational $x = k/2^n$ (these are the ones that have more than one binary expansion), we take the expansion that has infinitely many ones. We claim that T is measurable. Indeed, for the cylinder set above, $T^{-1}(A_{\varepsilon_1,\ldots,\varepsilon_n})$ is an interval of length 2^{-n} (with left end-point $a = \varepsilon_1 2^{-1} + \ldots + \varepsilon_n 2^{-n}$). Clearly, if B is a cylinder set specified by a subset of co-ordinates $i_1 < \ldots < i_k \leqslant n$, then B is a union of 2^{n-k} pairwise disjoint sets of the form $A_{\varepsilon_1,\ldots,\varepsilon_n}$. Therefore $T^{-1}B$ is a union of finitely many (pairwise disjoint) intervals, and hence a Borel subset of $[0, 1]$. Thus, T is measurable.

As T is measurable, we can define $\mathbb{P} = \lambda \circ T^{-1}$ as a probability measure on $\mathcal{F}$. Is this the probability measure we want? As we saw above, if $B = \{\omega : \omega_{i_1} = \varepsilon_1, \ldots, \omega_{i_k} = \varepsilon_k\}$, then $T^{-1}(B)$

is a union of 2^{n-k} intervals, each of length 2^{-n}, hence $\mathbb{P}(B) = \lambda(T^{-1}(A)) = 2^{n-k} \times 2^{-n} = 2^{-k}$. Thus, this measure agrees with the probabilities we wanted for cylinder sets.

Observe that all the hard work that was done in constructing Lebesgue measure did not have to be repeated here.

2.2. Picking a point at random from the Cantor set. Let K be the standard $\frac{1}{3}$-Cantor set. Recall that $K = \cap K_n$ where $K_0 = [0,1]$, $K_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, $K_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$ and so on. It is also the set of $x \in [0,1]$ whose base-3 expansion has no digit equal to 1. As a compact set, the Borel sigma-algebra of K is nothing but the collection of $A \cap K$, $A \in \mathcal{B}_{\mathbb{R}}$.

A natural map from $[0,1]$ to K is given by $T(x) = \sum_{n \geqslant 1} \frac{2x_n}{3^n}$ where $x = \sum_{n \geqslant 1} \frac{x_n}{2^n}$. Clearly T maps $[0,1]$ into K. Why is it measurable? Accepting that, we get a measure $\mu = \lambda \circ T^{-1}$ on K (with its Borel sigma-algebra). It is easy to see that each of the intervals comprising K_n get a measure of 2^{-n}. That justifies calling it "uniform measure on the Cantor set". It is also known as Cantor measure.

As an aside, one can think of μ as a measure on $(K, \mathcal{B}_K)$, but one can also think of it as a measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ by setting $\nu(A) = \mu(A \cap K)$ for $A \in \mathcal{B}_{\mathbb{R}}$. Then $\nu(K) = 1$.

2.3. All Borel probability measures on $\mathbb{R}$. What are all the probability measures on the Borel sigma-algebra of $\mathbb{R}$? In principle, the Carathéodory extension gives the way: Propose a candidate $P : \mathcal{S} \to [0,1]$, where $\mathcal{S}$ consists of all left open, right closed intervals. Extend it to the algebra of finite disjoint unions of such intervals, and check the conditions of finite and countable adiditivity. If the conditions are satisfied, you get a measure, otherwise not. This is not satisfactory as it is not explicit enough - how do know which function P work before hand?

To give the answer, recall that a *cumulative distribution function* (CDF) is any function $F : \mathbb{R} \to [0,]1$ that is increasing ($s \leqslant t \implies F(s) \leqslant F(t)$), right-continuous ($F(t+h) \downarrow F(t)$ as $h \downarrow 0$) and converges to 0 at $-\infty$ and to 1 at $+\infty$.

Let $\mathcal{P}(\mathbb{R})$ denote the set of all Borel probability measures on $\mathbb{R}$. For any $\mu \in \mathcal{P}(\mathbb{R})$, the function $F_\mu(t) := \mu(-\infty, t]$ is a CDF. This is easy to check. Interestingly, the converse is true.

Theorem 3

Let F be a CDF. Then, there is a unique $\mu \in \mathcal{P}(\mathbb{R})$ such that $F(t) = \mu(-\infty, t]$.

This gives a complete characterization of Borel probability measures on $\mathbb{R}$ in terms of much easier to understand objects, namely CDFs. One approach to the above theorem would be to define $\mu(a,b] = F(b) - F(a)$ for $a < b$, and extend it to the algebra of finite disjoint unions of left-open, right-closed intervals (including $(-\infty, a]$ and (b, ∞)) and check the conditions for the Carathéodory's extension theorem. A simpler way is below.

PROOF. Given a CDF F, define $T : (0,1) \to \mathbb{R}$ by $T(u) := \inf\{x \in \mathbb{R} : F(x) \geqslant u\}$ (well-defined as $F(x) \to 1$ as $x \to +\infty$). It is a kind of "generalized inverse" in the sense that $T(u) \leqslant x$ if and only if $F(x) \geqslant u$. Further, T is increasing, right-continuous and hence Borel measurable. Therefore $\mu := \lambda \circ T^{-1}$ is a probability measure. Further

$$\mu(-\infty, x] = \lambda\{u : T(u) \leqslant x\} = \lambda\{u : F(x) \geqslant u\} = \lambda(0, F(x)] = F(x).$$

Thus μ has CDF F. If ν was another probability measure on $\mathcal{B}_{\mathbb{R}}$ with the same CDF, then $\mu(a,b] = \nu(a,b]$ for all $a < b$. As they agree on a π-system that generates $\mathcal{B}_{\mathbb{R}}$, it follows that $\mu = \nu$. ■

As it is much easier to understand CDFs than to understand measures, this parameterization of $\mathcal{P}(\mathbb{R})$ by CDFs is very useful. At the very least, it allows us to write down many probability measures on $\mathbb{R}$. Two particular classes are useful to keep in mind.

(1) Measures with pmf (probability mass function): Give a real sequence $(x_1, x_2, \ldots)$ and $(p_1, p_2, \ldots)$ such that $p_i \geqslant 0$ and $\sum_i p_i = 1$. Then define $F(x) = \sum_{i:x_i \leqslant x} p_i$. This is a CDF that increases only by jumps, and the corresponding probability measure is said to have pmf given by (x_i) and (p_i).

Binomial, Poisson, Hypergeometric, Geometric, Negative-Binomial are important classes of examples of probability measures having pmf.

(2) Measures with pdf (probability density function): Give a Borel measurable $f : \mathbb{R} \to \mathbb{R}_+$ such that $\int_{\mathbb{R}} f(x)dx = 1$ (this integral is Lebesgue integral). Then $F(x) = \int_{-\infty}^{x} f(u)du = \int_{\mathbb{R}} f(u)\mathbf{1}_{u \leqslant x}du$ defines a CDF. The corresponding probability measure is said to have density F.

Normal, Exponential, Gamma, Uniform, Beta, Cauchy, are important classes of probability measures having pdf.

2.4. Higher dimensions. A CDF on $\mathbb{R}^d$ is a function $F : \mathbb{R}^d \to [0,1]$ that is increasing in each co-ordinate, is right continuous, and $F(t) \to 0$ if $\min\{t_1, \ldots, t_d\} \to -\infty$ and $F(t) \to 1$ if $\min\{t_1, \ldots, t_d\} \to +\infty$.

For a Borel probability measure μ on $\mathbb{R}^d$, one can associate a CDF by

$$F(t_1, \ldots, t_d) = \mu((-\infty, t_1] \times \ldots \times (-\infty, t_d]).$$

The converse is also true. For any CDF F, there is a unique probability measure μ on $\mathcal{B}_{\mathbb{R}^d}$ whose CDF is F.

Unlike in one dimension, it is not easy to prove this by giving a measurable function $T : [0,1] \to \mathbb{R}^d$ such that $\lambda \circ T^{-1} = \mu$. Instead it is better to take the way out using the Carathéodory extension theorem as follows.

For a left-open, right-closed rectangle $R = (a_1^-, a_1^+] \times \ldots \times (a_d^-, a_d^+]$ define

$$P(R) = \sum_{\varepsilon \in \{-,+\}^d} \varepsilon_1 \ldots \varepsilon_d F(a_1^{\varepsilon_1}, \ldots, a_d^{\varepsilon_d}).$$

The idea behind this definition is the inclusion-exclusion principle (if F was the CDF of a measure μ, then $\mu(R)$ would be precisely given by the above formula). For a finite, disjoint union of such rectangles, $A = R_1 \sqcup \ldots \sqcup R_m$, define $P(A) = P(R_1) + \ldots + P(R_m)$. Observe that the collection of all finite, disjoint unions of such rectangles forms an algebra $\mathcal{A}$. Further, $\sigma(\mathcal{A}) = \mathcal{B}_{\mathbb{R}^d}$.

Thus, to extend P to a probability measure on $\mathcal{B}_{\mathbb{R}^d}$, we must check the conditions of the Carathéodory extension theorem. The finite additivity is easy. Checking the second condition (continuity under increasing limits) can be reduced to the following (see the proof of existence of Lebesgue measure): If $A_n = R_{n,1} \sqcup \ldots \sqcup R_{n,k_n}$ increase to a rectangle R, then $P(R \setminus A_n) \downarrow 0$. This can be done exactly as in the case of Lebesgue measure (slightly decrease R to a compact set and slightly increase the $R_{n,j}$ to open sets and so on).

2.5. Polish spaces. More generally, every probabilty space of interest to probabilists can be got this way by pushing forward Lebesgue measure on $[0,1]$ by a measurable mapping. A *Polish space* is a complete, separable metric space (more precisely, a separable metric space whose topology can be induced by a complete metric, e.g., $(-\frac{\pi}{2}, \frac{\pi}{2})$ carries the complete metric $d(x,y) = |\tan^{-1} x - \tan^{-1} y|$).

Theorem 4: Borel isomorphism theorem

Let (X,d) be a Polish 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.

3. Random variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ 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. Recall that random variables take values in the extended real numbers $\overline{\mathbb{R}}$. Random variables are measurements in a random experiment (e.g., number of heads in a sequence of coin tosses, the distance from the

center at which a dart hits a dartboard, etc.). For every event $A \in \mathcal{F}$, one can associate the random variable $\mathbf{1}_A$ (indicator of A). Thus random variables are a generalization of events, from 2-valued measurements to multi-valued measurements.

3.1. Distributions of random variables. When one is interested in one single random variable X, all probability questions about it can be answered by finding its *distribution*, which is the push-forward measure $\mu := \mathbb{P} \circ X^{-1}$ on $\overline{\mathbb{R}}$. For example, $\mathbb{P}\{X \leqslant t\} = \mu(-\infty, t]$ is the CDF of μ. Here and in future, we just write $\{X \in A\}$ to mean $\{\omega \in \Omega : X(\omega) \in A\}$.

When considering several random variables, say $X_1, \ldots, X_n$, then one is interested in events such as $\{X_1 \leqslant t_1, \ldots, X_n \leqslant t_n\}$. The probability of such an event cannot be computed from the individual distributions of X_ks, but can be calculated from their *joint distribution*, which is just the probability measure $\nu := \mathbb{P} \circ X^{-1}$ on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$, where $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$. For example,

$$\mathbb{P}\{X_1 \leqslant t_1, \ldots, X_n \leqslant t_n\} = \nu((-\infty, t_1] \times \cdots \times (-\infty, t_n])$$

When considering a sequence of random variables $X_1, X_2, \ldots$, we can again form a single function $X = (X_1, X_2, \ldots) : \Omega \to \mathbb{R}^{\mathbb{N}}$. It is easy to check that X is measurable (when $\mathbb{R}^{\mathbb{N}}$ is endowed with the cylinder sigma-algebra) and hence we can define its distribution $\theta = \mathbb{P} \circ X^{-1}$, a probability measure on $\mathbb{R}^{\mathbb{N}}$. But a probability measure on the cylinder sigma-algebra is determined by its values on finite dimensional cylinders. In other words, the distribution of the sequence is completely determined by the collection of finite dimensional joint distributions of $(X_1, \ldots, X_n)$ for each n.

One can recast some of what we have discussed in the language of random variables. For example, given a CDF $F : \mathbb{R} \to [0, 1]$, instead of asking for a probability measure μ with CDF F, we could ask for a random variable X (on a probability space of your choice) such that $\mathbb{P}\{X \leqslant t\} = F(t)$ for all t. Indeed, if the measure μ exists, then we can take the probability space $(\mathbb{R}, \mathcal{B}, \mu)$ and define $X(t) = t$. Then X has distribution μ. Conversely, if there is some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ satisfying $\mathbb{P}\{X \leqslant t\} = F(t)$, we can construct the measure $\mu = \mathbb{P} \circ X^{-1}$.

Remark 3: A matter of language

Probabilists universally use the language of random variables and leave the probability space unidentified in the background. Statisticians and engineers do the same, but this tends to confuse other mathematicians who prefer the language of measures.

For example, a probabilist will say, "If X has $\text{Exp}(1)$ distribution, then $\mathbb{P}\{X \geqslant 2\} = 1/e^2$" while an analyst might say "In the measure space $(\mathbb{R}, \mathcal{B}, \mu = \text{Exp}(1))$, we have $\mu[2,\infty) = 1/e^2$". Observe that the probabilist did not specify the probability space (and the underlying probability measure is almost always denoted $\mathbb{P}$!) and the analyst did not specify a random variable (as in this example, if the probability space is chosen minimally, the random variable will be the identity function).

The convenience of the probabilists' way becomes more apparent when we have many random variables and start doing operations on random variables (adding/multiplying and later, conditioning). It also has the advantage of being closer to the way we think when applying probability to real-world situations[a].

[a] See this blog post by Timothy Gowers, paragraphs 3-11, for a discussion akin to this. In it is a though-provoking remark of David Aldous that a random variable is like a cake whereas a measure is like a recipe for a cake.

4. User's guide to expectation

For an indicator random variable $\mathbf{1}_A$, its distribution is completely described by giving one number, $\mathbb{P}(A)$. For a general random variable, the distribution is a complicated object, but if one wants a single-number summary, we give its *expectation* (but it does not always exist). Here are the fundamental facts about expectation:

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

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

(2) *Positivity:* $\mathbb{E}[X] \geqslant 0$ 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 $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$.

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

We did not say how $\mathbb{E}[\cdot]$ is defined. But accepting the above fact, one has the following explicit form: For any $X \in \text{RV}_+$,

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

This is got by observing that $X_n = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\frac{k}{2^n} \leqslant X < \frac{k+1}{2^n}}$ increase to X pointwise, and hence $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$ by the MCT. And $\mathbb{E}[X_n]$ can be got from linearity and the fact that $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$, leading to (1).

One may also take (1) as the definition of $\mathbb{E}[X]$ for $X \in \mathrm{RV}_+$. It is not hard to see that the limit exists, and one must then prove that it satisfies the four properties stated above. But the point we emphasize is that how expectation is defined rarely needs to be used, it is how it behaves (the fours properties listed above) that matters.

For general $X \in \mathrm{RV}$, we write it as $X = X_+ - X_-$ where $X_+ = X \vee 0$ and $X_- = (-X)_+ = -(X \wedge 0)$. If $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ are both finite, then we say that X has expectation (or that X is integrable) and define $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$. Observe that $X_+ + X_- = |X|$, hence integrability is equivalent to $\mathbb{E}[|X|] < \infty$. We also write $X \in L^1$ if X is integrable. More generally, if $|X|^p$ is integrable, we write $X \in L^p$ (or $L^p(\mathbb{P})$ or $L^p(\Omega, \mathcal{F}, \mathbb{P})$ if we must). We did not actually say what is L^1 or L^p, usually they are defined as a collection of equivalence classes got by identifying random variables that are equal a.s., i.e. $X \sim Y$ if $\mathbb{P}\{X = Y\} = 1$.

In conclusion, on the space of integrable random variables L^1, expectation is a positive linear functional that maps $\mathbf{1}_A$ to $\mathbb{P}(A)$. The notation $\int_\Omega X(\omega) d\mathbb{P}(\omega)$ or just $\int X d\mathbb{P}$ is also used for $\mathbb{E}[X]$, and it is also called *Lebesgue integral*.

Remark 4

For a random vector $X = (X_1, \ldots, X_n)$, we define $\mathbb{E}[X]$ to be $(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])$, if each X_i is integrable. Similarly, for a complex valued random variable $X = X_1 + iX_2$, we write $\mathbb{E}[X] = \mathbb{E}[X_1] + i\mathbb{E}[X_2]$, if X_1, X_2 have expectation. We cannot in general talk of Expectation of X if X is a measurable function into some arbitrary space Λ. The least we need is that Λ has a vector space structure (or at least Λ should be a convex set in a vector space). Indeed, whatever be the general notion of expectation, for the random variable X taking 2 values $a, b \in \Lambda$ (assume singletons are measurable) with equal probability, we would want the expectation to be $(a + b)/2$.

4.1. Lebesgue spaces. Fix $(\Omega, \mathcal{F}, \mathbb{P})$ and for $p > 0$ let $L^p(\Omega, \mathcal{F}, \mathbb{P})$ (or $L^p(\mathbb{P})$ or L^p in short) denote the set of all $X \in \mathrm{RV}$ such that $\mathbb{E}[|X|^p] < \infty$. For $p_1 < p_2$ we have $|x|^{p_1} \leqslant |x|^{p_2} + 1$ for all $x \in \mathbb{R}$, hence it is clear that the spaces L^p are decreasing in p (i.e., if $X \in L^{p_2}$ then $X \in L^{p_1}$). For any p, the space L^p is a vector space because

$$|X + Y|^p \leqslant \begin{cases} |X|^p + |Y|^p & \text{if } 0 < p \leqslant 1, \\ 2^{p-1}(|X|^p + |Y|^p) & \text{for } p \geqslant 1. \end{cases}$$

The case $p < 1$ is obvious and the case $p \geqslant 1$ follows by convexity of $x \to x^p$ for $p \geqslant 1$. In many ways the latter case is special, and one defines the p-norm $\|X\|_p = \mathbb{E}[|X|^p]^{\frac{1}{p}}$. While the homogeneity $\|\alpha X\|_p = |\alpha|\|X\|_p$ holds for any $p > 0$, the triangle inequality holds only for $p \geqslant 1$ (this is Minkowski's inequality, discussed later). For $p \geqslant 1$, the space L^p is a normed linear space. Often it is extended to $p = \infty$ by defining L^∞ as the set of all bounded random variables (we say that $X \in RV$ is bounded if $\mathbb{P}\{|X| \leqslant M\} = 1$ for some $M < \infty$). The most important of the L^p spaces are L^1, L^2 and L^∞.

Some remarks.

(1) Lebesgue showed that L^p space endowed with the L^p norm is complete (all Cauchy sequences converge). Surprisingly, this fundamental result will not play a role in this course, and we shall not discuss it.

(2) Another fact is that the p-norm is not quite a norm because $\|X\|_p = 0$ if and only if $X = 0$ a.s.$[\mathbb{P}]$. One can get a genuine norm by quotienting the space by the equivalence relation identifying X and Y if $X = Y$ a.s.$[\mathbb{P}]$. But we don't need this irritating language of equivalence classes and avoid it. For us, elements of L^p are in fact random variables.

(3) Although $X \in L^p$ means that $\mathbb{E}[|X|^p] < \infty$, observe that unless $X \in RV_+$ or p is a positive integer, we cannot talk of X^p (and hence $\mathbb{E}[X^p]$ does not make sense). For positive integers p, we can talk of $\mathbb{E}[X^p]$ (if it exists) and we call it the pth *moment* of X.

4.2. Inequalities. Cauchy-Schwarz, Hölder's and Minkowski's and Jensen's inequalities are important and repeatedly used. Fundamental to these is the notion of convexity.

Definition 5: Convex functions

A function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be convex if $\varphi(tx + (1-t)y) \leqslant t\varphi(x) + (1-t)\varphi(y)$ for all $x, y \in \mathbb{R}^d$ and all $t \in (0,1)$ and $\mathrm{Dom}(\varphi) := \{\varphi < \infty\}$ is not empty.

Why did we allow $+\infty$ as a value? Observe that $\mathrm{Dom}(\varphi)$ is a convex set (if $\varphi(x) < \infty$ and $\varphi(y) < \infty$ then $\varphi(tx + (1-t)y) < \infty$ for $t \in (0,1)$). Conversely, if $K \subseteq \mathbb{R}^d$ is a convex set and $\varphi : K \to \mathbb{R}$ is convex, then so is the extended function $\Phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined by $\Phi = \varphi$ in K and $\Phi = +\infty$ on K^c. Thus, by allowing the value $+\infty$, we can assume that the domain is all of $\mathbb{R}^d$. The following is a fundamental fact.

Supporting hyperplane theorem: Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex. Assume that $x_0 \in \mathrm{Dom}(\varphi)$. Then there exists $b \in \mathbb{R}^d$ such that $\varphi(x_0) + \langle b, x - x_0\rangle \leqslant \varphi(x)$ for all $x \in \mathbb{R}^d$.

The reason for the name is that the graph of $x \mapsto \varphi(x_0) + \langle b, x - x_0\rangle$ is an affine hyperplane that lies below the graph of φ, but touches it at $(x_0, \varphi(x_0))$.

Lemma 2: Jensen's inequality

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Let X be a random variable such that $\mathbb{P}\{X \in \mathrm{Dom}(\varphi)\} = 1$. Assume that $\mathbb{E}[X]$ exists. Then $\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)]$.

Proof. Let $x_0 = \mathbb{E}[X]$ and find $b \in \mathbb{R}^d$ such that $\varphi(x_0) + \langle b, x - x_0\rangle \leqslant \varphi(x)$ for all $x \in \mathbb{R}^d$. Then $\varphi(x_0) + \langle b, X - x_0\rangle \leqslant \varphi(X)$ a.s., and taking expectations we see that $\varphi(x_0) \leqslant \mathbb{E}[\varphi(X)]$, since $\mathbb{E}[X - x_0] = 0$. ■

Next we prove the triangle inequality for p-norms, $p \geqslant 1$.

Lemma 3: Minkowski's inequality

For any $p \geqslant 1$, we have $\|X + Y\|_p \leqslant \|X\|_p + \|Y\|_p$.

The important special cases of $p = 1, 2, \infty$ can be checked easily. The general case is non-trivial!

Proof. Take $1 \leqslant p < \infty$ and assume that $\|X\|_p > 0$ and $\|Y\|_p > 0$. Let $X' = X/\|X\|_p$ and $Y' = Y/\|Y\|_p$. Convexity of $x \mapsto x^p$ yields $|aX' + bY'|^p \leqslant a|X'|^p + b|Y'|^p$ where $a = \frac{\|X\|_p}{\|X\|_p + \|Y\|_p}$ and $b = \frac{\|Y\|_p}{\|X\|_p + \|Y\|_p}$. Take expectations and observe that $\mathbb{E}[|aX' + bY'|^p] = \frac{\mathbb{E}[|X+Y|^p]}{(\|X\|_p + \|Y\|_p)^p}$ while $\mathbb{E}[a|X'|^p + b|Y'|^p] = 1$ since $\mathbb{E}[|X'|^p] = \mathbb{E}[|Y'|^p] = 1$. Thus we get

$$\frac{\mathbb{E}[|X + Y|^p]}{(\|X\|_p + \|Y\|_p)^p} \leqslant 1,$$

which is precisely Minkowski's inequality. ■

Lastly, we prove Hölder's inequality of which the most important special case is the Cauchy-Schwarz inequality.

Lemma 4: Cauchy-Schwarz and Hölder inequalities

(1) If X, Y are L^2 random variables on a probability space, then XY is integrable and $\mathbb{E}[XY]^2 \leqslant \mathbb{E}[X^2]\mathbb{E}[Y^2]$.

(2) If X, Y are L^p r.v.s on a probability space, then for any $p, q \geqslant 1$ satisfying $p^{-1} + q^{-1} = 1$, we have $XY \in L^1$ and $\|XY\|_1 \leqslant \|X\|_p \|Y\|_q$.

Proof. Cauchy-Schwarz is a special case of Hölder with $p = q = 2$, but one can also give a direct proof. First observe that $2|XY| \leqslant X^2 + Y^2$ showing the integrability of XY. For any $t \in \mathbb{R}$

$$0 \leqslant \mathbb{E}[|X + Y|^2] = \mathbb{E}[X]^2 + 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2]$$

hence the discriminant of this quadratic expression must be negative, i.e., $\mathbb{E}[XY]^2 \leqslant \mathbb{E}[X^2]\mathbb{E}[Y^2]$.

Hölder's inequality follows by applying the inequality $a^p/p + b^q/q \geqslant ab$ valid for $a, b \geqslant 0$, to $a = |X|/\|X\|_p$ and $b = |Y|/\|Y\|_q$ and taking expectations.

The inequality $a^p/p + b^q/q \geqslant ab$ is evident by noticing that the rectangle $[0, a] \times [0, b]$ (with area ab) is contained in the union of the region$\{(x, y) : 0 \leqslant x \leqslant a,\ 0 \leqslant y \leqslant x^{p-1}\}$ (with area a^p/p) and the region $\{(x, y) : 0 \leqslant y \leqslant b,\ 0 \leqslant x \leqslant y^{q-1}\}$ (with area b^q/q). This is because the latter regions are the regions between the x and y axes (resp.) and curve $y = x^{p-1}$ which is also the curve $x = y^{q-1}$ since $(p-1)(q-1) = 1$. ■

Remark 5

To see the role of convexity, here is another way to prove that $a^p/p + b^q/q \geqslant ab$. Set $a' = p \log a$ and $b' = q \log b$ and observe that the desired inequality is equivalent to $\frac{1}{p}e^{a'} + \frac{1}{q}e^{b'} \geqslant e^{\frac{1}{p}a' + \frac{1}{q}b'}$, which follows from the convexity of $x \to e^x$.

In the study of L^p spaces, there is a close relationship between L^p and L^q where $\frac{1}{p} + \frac{1}{q} = 1$. In the proof of Hölder's inequality, we see one elementary way in which it arises (the inverse of $y = x^{p-1}$ is $x = y^{q-1}$).

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

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

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

Fatou's lemma follows directly from MCT by observing that $Y_n := \inf_{k \geqslant n} X_k$ increase to $Y := \liminf X_n$ and that $0 \leqslant Y_n \leqslant X_n$.

DCT follows by applying Fatou's lemma to $Y - X_n$ and to $Y + X_n$, both of which are sequences of positive random variables converging respectively to $Y - X$ and $Y + X$ a.s. Then, Fatou's lemma then gives

$$\mathbb{E}[Y] + \mathbb{E}[X] = \mathbb{E}[Y + X] \leqslant \liminf \mathbb{E}[Y + X_n] = \liminf \mathbb{E}[Y] + \mathbb{E}[X_n] = \mathbb{E}[Y] + \liminf \mathbb{E}[X_n],$$

$$\mathbb{E}[Y] - \mathbb{E}[X] = \mathbb{E}[Y - X] \leqslant \liminf \mathbb{E}[Y - X_n] = \liminf \mathbb{E}[Y] - \mathbb{E}[X_n] = \mathbb{E}[Y] - \limsup \mathbb{E}[X_n].$$

Thus, $\mathbb{E}[X] \leqslant \liminf \mathbb{E}[X_n] \leqslant \limsup \mathbb{E}[X_n] \leqslant \mathbb{E}[X]$ showing that $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Apply this conclusion to the sequence $|X_n - X|$ that is dominated by $2Y$ and converges almost surely to 0 to get $\mathbb{E}[|X_n - X|] \to 0$.

4.4. Change of variables. Suppose $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2, 3$ are probability spaces. Assume that $\mathbb{P}_2 = \mathbb{P}_1 \circ T^{-1}$ for some measurable function $T : \Omega_1 \to \Omega_2$ and that $\mathbb{P}_3 = \mathbb{P}_2 \circ S^{-1}$ for some measurable function $S : \Omega_2 \to \Omega_3$. It is trivial to check that $U = S \circ T : \Omega_1 \to \Omega_3$ is measurable and that $\mathbb{P}_3 = \mathbb{P}_1 \circ U^{-1}$.

This easy observation will be used throughout. Here are some ways.

► Let X be a real-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ having distribution μ. Then for any Borel measurable $\varphi : \mathbb{R} \to \mathbb{R}$, the distribution of $f(X)$ is $\mu_\varphi := \mu \circ f^{-1}$. This is got by taking $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1) = (\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2) = (\mathbb{R}, \mathcal{B}_\mathbb{R}, \mu)$ and $(\Omega_3, \mathcal{F}_3, \mathbb{P}_3) = (\mathbb{R}, \mathcal{B}_\mathbb{R}, \mu_f)$ and $T = X$ and $S = \varphi$.

E.g., if X has CDF F, then X^3 has CDF $x \mapsto F(x^{1/3})$ and e^X has CDF $x \mapsto F(\log x)$.

► The same is true if $X = (X_1, \ldots, X_d)$ is $\mathbb{R}^d$-valued (or even $\mathbb{R}^\mathbb{N}$-valued) and $\varphi : \mathbb{R}^d \to \mathbb{R}$ (or $f : \mathbb{R}^\mathbb{N} \to \mathbb{R}$). The distribution of $\varphi(X)$ is $\mu \circ \varphi^{-1}$ where μ is the distribution of X (a probability measure on $\mathbb{R}^d$ or $\mathbb{R}^\mathbb{N}$). In particular, as $X_k = \Pi_k \circ X$, where $\Pi_k(x_1, \ldots, x_d) = x_k$ is the kth projection, the marginal distribution of X_k is determined by the distribution of X.

E.g., if (X_1, X_2) has density 1 on $[0,1]^2$ and zero outside, then $X_1 - X_2$ has the density $(1 - |x|)$ on $[-1, 1]$.

► It is useful to remember the change of variables formula for densities. Let X be an $\mathbb{R}^d$-valued random variable and assume that it has density g that is positive on an open set U and zero outside. Let $T : U \to V$ be a bijection to another open set $V \subseteq \mathbb{R}^d$ (same dimension) such that T^{-1} is differentiable. Then the density of $Y := g(X)$ is $h(y) = g(T^{-1}y)|JT^{-1}(y)|$ on V.

The point above is that if we know the distribution of X, to compute the distribution of $f(X)$, we need no further information (in particular the original probability space is irrelevant). Then the same must be true for expectation of $f(X)$. First we state the general point.

Suppose $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2$ be probability spaces. Assume that $\mathbb{P}_2 = \mathbb{P}_1 \circ T^{-1}$ for some measurable function $T : \Omega_1 \to \Omega_2$. If $Y : \Omega_2 \to \mathbb{R}_+$ is a random variable on Ω_2, then $Y \circ T$ is a random variable on Ω_1 and $\mathbb{E}_{\mathbb{P}_2}[Y] = \mathbb{E}_{P_1}[Y \circ T]$. In other notation,

$$\int_{\Omega_2} Y(\omega') d\mathbb{P}_2(\omega') = \int_{\Omega_1} Y(T(\omega)) d\mathbb{P}_1(\omega).$$

For general random variable, $(Y \circ T)_\pm = (Y_\pm) \circ T$, hence the same conclusion holds, except that we must make the more cautious statement: "Y has expectation w.r.t. $\mathbb{P}_2$ if and only if $Y \circ T$ has expectation w.r.t. $\mathbb{P}_1$, and in that case the two quantities are equal".

Proof. If $Y = \mathbf{1}_B$ for some $B \in \mathcal{F}_2$, the identity follows from the definition of push-forward measure. By linearity, it holds for simple random variables (linear combinations of indicators). For $Y : \Omega_2 \to \mathbb{R}_+$, we can find $Y_n : \Omega_2 \to \mathbb{R}_+$ that are simple and increase to Y pointwise. Then $Y_n \circ T \uparrow Y \circ T$ pointwise too. By applying MCT to both sequences, we get the conclusion for positive random variables. The reduction from general (integrable) random variables to positive ones is straightforward. ■

► As a particular case, if X has distribution μ on $\mathbb{R}^d$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ is Borel measurable and bounded, then $\mathbb{E}[\varphi(X)]$ which is $\int_\Omega \varphi(X(\omega))d\mathbb{P}(\omega)$ can also be written as $\int_{\mathbb{R}^d} \varphi(x)d\mu(x)$. Boundedness was assumed only to ensure that the expectations exist.

► In particular, if X has density g on $\mathbb{R}^d$, then $\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x)g(x)dx$ (which is easier than trying to find the distribution of $\varphi(X)$ and then integrating x w.r.t. that).

Example 4

Let $X \sim N(0,1)$, i.e., X has density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ on $\mathbb{R}$. Then $\mathbb{E}[X^n] = 0$ for odd n and

$$\mathbb{E}[X^{2n}] = (2n-1)(2n-3)\dots(3)(1).$$

To see this, we don't need to know what the original probability space is. Just compute

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x^{2n}e^{-x^2/2}dx &= \frac{2}{\sqrt{2\pi}}\int_0^\infty (2u)^{n-\frac{1}{2}}e^{-u}du \\ &= \frac{2^n}{\sqrt{2\pi}}\Gamma(n+\frac{1}{2}) \\ &= \frac{2^n}{\sqrt{2\pi}}\Gamma(1/2)\frac{1}{2}\times\frac{3}{2}\times\ldots\times\frac{2n-1}{2}\end{aligned}$$

which is the claim (recall that $\Gamma(1/2) = \sqrt{\pi}$).

4.5. Reweighting a measure to get new measures. On $(\Omega, \mathcal{F}, \mathbb{P})$, let X be a positive random variable with $\mathbb{E}[X] = 1$. Define $\mathbb{Q} : \mathcal{F} \to \mathbb{R}_+$ by $\mathbb{Q}(A) = \mathbb{E}[X\mathbf{1}_A]$. Then, $\mathbb{Q}$ is a probability measure.

Proof. Finite additivity of $\mathbb{Q}$ follows from the linearity of expectation (if $A, B \in \mathcal{F}$ are disjoint, $\mathbf{1}_{A\sqcup B} = \mathbf{1}_A + \mathbf{1}_B$). Further, if $A_n \in \mathcal{F}$ and $A_n \uparrow A$, then $X\mathbf{1}_{A_n} \uparrow X\mathbf{1}_A$, hence MCT shows that $\mathbb{Q}(A_n) \uparrow \mathbb{Q}(A)$. ■

One can think of $\mathbb{Q}$ as got by reweighting points of Ω according to the value of X. If we use the integral notation for expectation, then $\int \mathbf{1}_A(\omega)d\mathbb{Q}(\omega) = \int_A \mathbf{1}_A(\omega)X(\omega)d\mathbb{P}(\omega)$. Hence we also write this relationship as $d\mathbb{Q} = Xd\mathbb{P}$. We also say that say that X is the *Radon-Nikodym derivative* or the *density* of $\mathbb{Q}$ w.r.t. $\mathbb{P}$.

The reason for this name is in the *Radon-Nikodym theorem* to be discussed elsewhere. That theorem answers the converse question: Given $\mathbb{P}$ and $\mathbb{Q}$ (or even infinite measures), how can we tell if $\mathbb{Q}$ can be got from $\mathbb{P}$ by reweighting by some $X \in RV_+$?

5. Independence

Definition 6: Independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

- Let $\mathcal{G}_1, \ldots, \mathcal{G}_k$ be sub-sigma algebras of $\mathcal{F}$. We say that $\mathcal{G}_i$ are *independent* if for every $A_1 \in \mathcal{G}_1, \ldots, A_k \in \mathcal{G}_k$, we have $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbb{P}(A_1) \ldots \mathbb{P}(A_k)$.
- Random variables $X_1, \ldots, X_n$ on $\mathcal{F}$ are said to be independent if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.
- An arbitrary collection of σ-algebras $\mathcal{G}_i$, $i \in I$, (each $\mathcal{G}_i$ contained in $\mathcal{F}$) is said to be independent if every finite sub-collection of them is independent. Same applies for random variables.

How does this compare with the definitions we have seen in basic probability class?

- Since $\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}_{\mathbb{R}}\}$ for a real-valued random variable X, the definition above is equivalent to saying that $\mathbb{P}(X_i \in A_i \; i \leqslant k) = \prod_{i=1}^{k} \mathbb{P}(X_i \in A_i)$ for any $A_i \in \mathcal{B}(\mathbb{R})$. The same definition can be made for random variables X_i taking values in some metric space (Λ_i, d_i), but then A_i must be a Borel subset of Λ_i.
- Events $A_1, \ldots, A_k$ are said to be independent if $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}$ are independent. This is equivalent to *either* of the following sets of 2^n conditions:
  (1) $\mathbb{P}(A_{j_1} \cap \ldots \cap A_{j_\ell}) = \mathbb{P}(A_{j_1}) \ldots \mathbb{P}(A_{j_\ell})$ for any $1 \leqslant j_1 < j_2 < \ldots < j_\ell \leqslant k$.
  (2) $\mathbb{P}(A_1^{\pm} \cap A_2^{\pm} \cap \ldots \cap A_n^{\pm}) = \prod_{k=1}^{n} \mathbb{P}(A_k^{\pm})$ where we use the notation $A^+ = A$ and $A^- = A^c$.
  The second is clear, since $\sigma(A_k) = \{\emptyset, \Omega, A_k, A_k^c\}$. The equivalence of the first and second is an exercise.

Some remarks are in order.

(1) Independence is defined with respect to a fixed probability measure $\mathbb{P}$.

(2) It would be convenient if we need check the condition in the definition only for a sufficiently large class of sets. However, if $\mathcal{G}_i = \sigma(\mathcal{S}_i)$, and for every $A_1 \in \mathcal{S}_1, \ldots, A_k \in \mathcal{S}_k$ if we have $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbb{P}(A_1) \ldots \mathbb{P}(A_k)$, we *cannot* conclude that $\mathcal{G}_i$ are independent! If $\mathcal{S}_i$ are π-systems, then it is indeed true that $\mathcal{G}_i$ are independent (proof below).

(3) Checking pairwise independence is insufficient to guarantee independence. For example, suppose X_1, X_2, X_3 are independent and $\mathbb{P}(X_i = +1) = \mathbb{P}(X_i = -1) = 1/2$. Let $Y_1 = X_2X_3$, $Y_2 = X_1X_3$ and $Y_3 = X_1X_2$. Then, Y_i are pairwise independent but not independent.

Lemma 5

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Assume that $\mathcal{G}_i = \sigma(S_i) \subseteq \mathcal{F}$, that S_i is a π-system and that $\Omega \in S_i$ for each $i \leqslant k$. If for every $A_1 \in S_1, \ldots, A_k \in S_k$ if we have $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbb{P}(A_1) \ldots \mathbb{P}(A_k)$, then $\mathcal{G}_i$ are independent.

PROOF. Fix $A_2 \in S_2, \ldots, A_k \in S_k$ and set

$$\mathcal{F}_1 := \{B \in \mathcal{G}_1 : \mathbb{P}(B \cap A_2 \cap \ldots \cap A_k) = \mathbb{P}(B)\mathbb{P}(A_2) \ldots \mathbb{P}(A_k)\}.$$

Then $\mathcal{F}_1 \supseteq S_1$ by assumption. We claim that $\mathcal{F}_1$ is a λ-system. Assuming that, by the π-λ theorem, it follows that $\mathcal{F}_1 = \mathcal{G}_1$ and we get the assumptions of the lemma for $\mathcal{G}_1, S_2, \ldots, S_k$. Repeating the argument for S_2, S_3 etc., we get independence of $\mathcal{G}_1, \ldots, \mathcal{G}_k$.

To prove that $\mathcal{F}_1$ is a λ system is straightforward. If $B_n \uparrow B$ and $B_n \in \mathcal{F}_1$, then $B \in \mathcal{F}$ and $\mathbb{P}(B_n \cap A_2 \cap \ldots \cap A_k) \uparrow \mathbb{P}(B \cap A_2 \cap \ldots \cap A_k)$ and $\mathbb{P}(B_n) \prod_{j=2}^{k} \mathbb{P}(A_j) \uparrow \mathbb{P}(B) \prod_{j=2}^{k} \mathbb{P}(A_j)$. Hence $B \in \mathcal{F}_1$. Similarly, check that if $B_1 \subseteq B_2$ and both are in $\mathcal{F}_1$, then $B_2 \setminus B_1 \in \mathcal{F}_1$. Lastly, $\Omega \in S_1 \subseteq \mathcal{F}_1$ by assumption. Thus, $\mathcal{F}_1$ is a λ-system. ■

Remark 6

If $A_1, \ldots, A_k$ are events, then $\mathcal{G}_i = \{\emptyset, A_i, A_i^c, \Omega\}$ is generated by the π-system $S_i = \{A_i\}$. However, checking the independence condition for the generating set (which is just one equation $\mathbb{P}(A_1 \cap \ldots \cap A_k) = \prod_{j=1}^{k} \mathbb{P}(A_j)$) does not imply independence of $A_1, \ldots, A_k$. This shows that the condition that S_i should contain Ω is not redundant in the above Lemma!

Corollary 1

(1) Random variables $X_1, \ldots, X_k$ are independent if and only if for every $t_1, \ldots t_k \in \mathbb{R}$ we have $\mathbb{P}(X_1 \leqslant t_1, \ldots, X_k \leqslant t_k) = \prod_{j=1}^{k} \mathbb{P}(X_j \leqslant t_j)$.

(2) Suppose $\mathcal{G}_\alpha$, $\alpha \in I$ are independent. Let $I_1, \ldots, I_k$ be pairwise disjoint subsets of I. Then, the σ-algebras $\mathcal{F}_j = \sigma\left(\cup_{\alpha \in I_j} \mathcal{G}_\alpha\right)$ are independent.

(3) If $X_{i,j}$, $i \leqslant n$, $j \leqslant n_i$, are independent, then for any Borel measurable $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$, the r.v.s $f_i(X_{i,1}, \ldots, X_{i,n_i})$ are also independent.

PROOF. (1) Pulling back the familiar π-system of left-closed, right-open intervals on the line, we get the π-system $S_i := \{X_i^{-1}(-\infty, t] : t \in \mathbb{R}\}$ on Ω. Further S_i generates $\sigma(X_i)$.

(2) For $j \leqslant k$, let S_j be the collection of finite intersections of sets A_i, $i \in I_j$. Then S_j are π-systems and $\sigma(S_j) = \mathcal{F}_j$.

(3) Infer (3) from (2) by considering $\mathcal{G}_{i,j} := \sigma(X_{i,j})$ and observing that $f_i(X_{i,1}, \ldots, X_{i,k}) \in \sigma(\mathcal{G}_{i,1} \cup \ldots \cup \mathcal{G}_{i,n_i})$. ■

So far, we stated conditions for independence in terms of probabilities of events. As usual, they generalize to conditions in terms of expectations of random variables.

Lemma 6

(1) Sigma algebras $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are independent if and only if for every $\mathcal{G}_i$-measurable, bounded random variable X_i, for $1 \leqslant i \leqslant k$, we have $\mathbb{E}[X_1 \ldots X_k] = \prod_{i=1}^k \mathbb{E}[X_i]$.

(2) In particular, random variables $Z_1, \ldots, Z_k$ (Z_i is an n_i dimensional random vector) are independent if and only if $\mathbb{E}[\prod_{i=1}^k f_i(Z_i)] = \prod_{i=1}^k \mathbb{E}[f_i(Z_i)]$ for any bounded Borel measurable functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$.

We say 'bounded measurable' just to ensure that expectations exist. The proof goes inductively by fixing $X_2, \ldots, X_k$ and then letting X_1 be a simple r.v., a non-negative r.v. and a general bounded measurable r.v.

PROOF. (1) Suppose $\mathcal{G}_i$ are independent. By the linearity of Expectation, we see that $(X_1, \ldots, X_k) \mapsto \mathbb{E}[X_1 \ldots X_k]$ is linear in each co-ordinate if the others are fixed. The same is true of $(X_1, \ldots, X_k) \mapsto \prod_{i=1}^k \mathbb{E}[X_i]$.

If $X_i = \mathbf{1}_{A_i}$ for some $A_i \in \mathcal{G}_i$, then the claimed equality holds by definition of independence. By the multi-linearity observed above, the claim also holds for simple random variables X_i. Further, if $0 \leqslant X_{k,n} \uparrow X_k$ and $X_{k,n}$ are simple, then applying MCT on both sides, we get the equality for positive random variables. For general X_k, write it is as the difference of its positive and negative parts and expand the products on both sides. We get 2^k summands and the claimed equality easily.

Conversely, if $\mathbb{E}[X_1 \ldots X_k] = \prod_{i=1}^k \mathbb{E}[X_i]$ for all $\mathcal{G}_i$-measurable functions X_is, then applying to indicators of events $A_i \in \mathcal{G}_i$ we see the independence of the σ-algebras $\mathcal{G}_i$.

(2) The second claim follows from the first by setting $\mathcal{G}_i := \sigma(Z_i)$ and observing that a random variable X_i is $\sigma(Z_i)$-measurable if and only if (see remark following the proof) $X = f \circ Z_i$ for some Borel measurable $f : \mathbb{R}^{n_i} \to \mathbb{R}$. ■

Remark 7

We stated a fact that if X is a real-valued random variable and $Y \in \sigma(X)$, then $Y = f(X)$ for some $f : \mathbb{R} \to \mathbb{R}$ that is Borel measurable. Why is that so?

If $X(\omega) = X(\omega')$, then it is clear that any set $A \in \sigma(X)$ either contains both ω, ω' or excludes both (this was an exercise). Consequently, we must have $Y(\omega) = Y(\omega')$ (otherwise, if $Y(\omega) < a < Y(\omega')$ for some $a \in \mathbb{R}$, then the set $Y < a$ could not be in $\sigma(X)$, as it contains ω but not ω'). This shows that $Y = f(X)$ for some function $f : \mathbb{R} \to \mathbb{R}$. But why is f measurable? Indeed, one should worry a little, because the correct statement is not that f is measurable, but that f may be chosen to be measurable. For example, if X is the constant 0 and Y is the constant 1, then all we know is $f(0) = 1$. We shall have $Y = f(X)$ however we define f on $\mathbb{R} \setminus \{0\}$ (in particular, we may make f non-measurable!).

One way out is to use the fact that the claim is true for simple random variables and that every random variable can be written as a pointwise limit of simple random variables (see exercise below). Consequently, $Y = \lim Y_n$, where Y_n is a $\sigma(X)$-measurable simple random variable and hence $Y_n = f_n(X)$ for some Borel measurable $f_n : \mathbb{R} \to \mathbb{R}$. Let $f = \limsup f_n$, also Borel measurable. But $Y = f(X)$.

5.1. Existence of independent random variables. Now we come to the question of existence of independent random variables with given distributions. The following is the starting point of probability theory.

Proposition 1: [Daniell, Kolmogorov]

Let $\mu_i \in \mathcal{P}(\mathbb{R})$, $i \geqslant 1$, be Borel p.m on $\mathbb{R}$. Then, there exist a probability space with *independent* random variables $X_1, X_2, \ldots$ such that $X_i \sim \mu_i$.

Proof. We arrive at the construction in three stages.

(1) **Independent Bernoullis:** On the probability space $((0,1), \mathcal{B}, \lambda)$, consider the random variables $X_k : (0,1) \to \mathbb{R}$, where $X_k(\omega)$ is defined to be the k^{th} digit in the binary expansion of ω (see Section **??** for convention regarding binary expansion). We have seen that $X_1, X_2, \ldots$ are independent Bernoulli$(1/2)$ random variables.

(2) **Independent uniforms:** Note that as a consequence[4], on any probability space, if Y_i are i.i.d. Ber(1/2) variables, then $U := \sum_{n=1}^{\infty} 2^{-n} Y_n$ has uniform distribution on $[0,1]$. Consider again the canonical probability space and the r.v. X_i, and set $U_1 := X_1/2 + X_3/2^2 + X_5/2^3 + \ldots$, $U_2 := X_2/2 + X_6/2^2 + \ldots$, $U_3 = X_4/2 + X_{12}/2^2 + \ldots$ etc. (in short, let $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be an injection and define $Y_k = \sum_{j=1}^{\infty} X_{g(k,j)} 2^{-j}$). Clearly, U_i are i.i.d. Unif$[0,1]$.

(3) **Arbitrary distributions:** For a p.m. μ, recall the left-continuous inverse G_μ that had the property that $G_\mu(U) \sim \mu$ if $U \sim U[0,1]$. Suppose we are given p.m.s $\mu_1, \mu_2, \ldots$. On the canonical probability space, let U_i be i.i.d uniforms constructed as before. Define $X_i := G_{\mu_i}(U_i)$. Then, X_i are independent and $X_i \sim \mu_i$. Thus we have constructed an independent sequence of random variables having the specified distributions. ■

The same proof works for a countable product of $(\Omega_i, \mathcal{F}_i, \mu_i)$, provided each μ_i is a pushforward of Lebesgue measure, that is, $\mu_i = \mathbb{P} \circ T_i^{-1}$ for some $T_i : [0,1] \to \Omega_i$. The only change needed is to set $X_i = T_i(U_i)$ (instead of $G_{\mu_i}(U_i)$) in the last step. As we know, all Borel probability measures on $\mathbb{R}^d$ are push-forwards of Lebesgue measure and hence, the above proof works if $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$, $i \geqslant 1$, and gives a sequence of independent random vectors X_k such that $X_k \sim \mu_k$.

One may ask whether one can construct uncountably many independent random variables with specified distributions. It is possible, but entirely useless. There is no situation in probability that requires or can benefit from the existence of uncountably many independent random variables. Hence we do not concern ourselves with that.

6. Product measures

Suppose $X_1, X_2, \ldots$ are real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ be defined by $X = (X_1, X_2, \ldots)$. Then X is measurable (on $\mathbb{R}^{\mathbb{N}}$ we have the Borel sigma-algebra which is the same as the cylinder sigma-algebra). Therefore, $\mu = \mathbb{P} \circ X^{-1}$ is a probability measure on $\mathbb{R}^{\mathbb{N}}$, and $\mu_k := \mathbb{P} \circ X_k^{-1}$ is a Borel probability measure on $\mathbb{R}$. If $\Pi_k : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is the projection on the kth co-ordinate, then $X_k = \Pi_k \circ X$, hence $\mu_k = \mu \circ \Pi_k^{-1}$ (change of variables). We say that μ is the joint

[4]Let us be pedantic and show this: Suppose Y_i are independent Bernoullis on $(\Omega, \mathcal{F}, \mathbb{P})$ and $V = \sum_{k \geqslant 1} Y_k/2^k$. For any dyadic interval $I = [p2^{-n}, (p+1)2^{-n}]$ with $p + 1 \leqslant 2^n$, we see that $V \in I$ if and only if $Y_1, \ldots, Y_n$ take on specific values, hence $\mathbb{P}\{V \in I\} = 2^{-n}$. From this, we see that $F_V(t) = t$ for any dyadic rational $t \in [0,1]$, and by right-continuity that $F_V(t) = t$ for all $t \in [0,1]$. Thus $V \sim$ Unif$[0,1]$.

Again, we emphasize the unimportance of the original probability space, what matters is the joint distribution of the random variables that we are interested in. In other words, the mapping $Y = (Y_1, Y_2, \ldots) : \Omega \to \{0,1\}^{\mathbb{N}}$ pushes forward $\mathbb{P}$ to the fair-coin-tossing measure that we had constructed earlier, and hence the distribution of any function of Y, such as V, is the same regardless of the original probability space. Since the claim is true for the binary digits X_k on $[0,1]$, it is true for any independent Bernoullis.

distribution of $X_1, X_2, \ldots$ (or simply the distribution of X) and that μ_k is the marginal distribution of X_k. We also say that μ_k is the kth marginal of μ.

Now suppose X_k are independent. Then one can recover μ from $(\mu_k)_k$, because for any finite dimensional cylinder set $A = A_1 \times \ldots \times A_n \times \mathbb{R} \times \mathbb{R} \ldots$ with $A_k \in \mathcal{B}_{\mathbb{R}^k}$, we have

$$\begin{aligned}\mu(A) &= \mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} \\ &= \mathbb{P}\{X_1 \in A_1\} \ldots \mathbb{P}\{X_n \in A_n\} \;=\; \mu_1(A_1) \ldots \mu_n(A_n).\end{aligned}$$

Thus, if we know the marginal distributions μ_k, then we can recover the joint distribution μ on the π-system of finite dimensional cylinders, and hence on the Borel sigma-algebra of $\mathbb{R}^{\mathbb{N}}$. Conversely, if for some $X = (X_1, X_2, \ldots)$, the above relationship between μ and the $(\mu_k)_k$ on cylinder sets holds, then the random variables X_k are independent. This is easy to see and left as exercise.

In other words, we have found a formulation of independence in terms of measures. Let us make a definition in greater generality.

Definition 7: Product measure

Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i \in I$, be probability spaces indexed by an arbitrary set I. Let $\Omega = \times_{i \in I} \Omega_i$ and let $\mathcal{F}$ (usually denoted $\otimes_{i \in I} \mathcal{F}_i$) be the sigma-algebra generated by all finite dimensional cylinders (equivalently, the smallest sigma-algebra on Ω for which all the projections $\Pi_i : \Omega \to \Omega_i$ are measurable). If μ is a probability measure on $(\Omega, \mathcal{F})$ such that for any cylinder set $A = \Pi_{i_1}^{-1}(A_{i_1}) \cap \ldots \cap \Pi_{i_k}^{-1}(A_{i_k})$ for some $A_{i_r} \in \mathcal{F}_{i_r}$,

$$\mu(A) = \prod_{r=1}^{k} \mu_{i_r}(A_{i_r}),$$

then we say that μ is the product of μ_i, $i \in I$, and write $\mu = \otimes_{i \in I} \mu_i$.

The existence of independent random variables and the discussion at the beginning of this subsection show that if $\mu_i \in \mathcal{P}(\mathbb{R})$ (or even $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$ for $i \in \mathbb{N}$, then the product measure $\mu = \mu_1 \otimes \mu_2 \otimes \ldots$ on $\mathbb{R}^{\mathbb{N}}$ exists.

For arbitrary probability measures on arbitrary spaces and arbitrary (even uncountable) index sets, does product measure exist? Yes, irrespective of the cardinality of I, one can use the Carathéodory construction, starting from the desired probabilities for finite dimensional cylinders. The algebra generated by cylinder sets consists of finite disjoin unions of cylinder sets, and the measure is naturally defined on that. The key point is to check that μ is finitely and countably additive on the algebra. Then it extends to a measure on the sigma-algebra. We do not discuss any further as uncountable products are not needed[5].

[5] If interested, consult Dudley's *Real analysis and probability* for example.

6.1. Fubini's theorem. One of the important and useful facts about product measures is that the integral w.r.t the product measure can be computed by integrating over each variable one after another.

Theorem 5: Fubini-Tonelli theorem

Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2$, be probability spaces and let $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ and $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$. Let $Y : \Omega \to \mathbb{R}$ be a random variable that is either positive or integrable w.r.t. $\mathbb{P}$. Then, $Y(\omega_1, \cdot) : \Omega_2 \to \mathbb{R}$ is a random variable on Ω_2 for each $\omega_1 \in \Omega_1$, and is either positive or integrable (w.r.t. $\mathbb{P}_2$) for a.e. ω_1 $[\mathbb{P}_1]$. Further, the function $\omega_1 \mapsto \int_{\Omega_2} Y(\omega_1, \omega_2) d\mathbb{P}_2(\omega_2)$ is a random variable on Ω_1, and is either positive or integrable. Finally,

$$\int_{\Omega_1} \left[\int_{\Omega_2} Y(\omega_1, \omega_2) d\mathbb{P}_2(\omega_2) \right] d\mathbb{P}_1(\omega_1) = \int_{\Omega} Y d\mathbb{P}.$$

Two remarks:

(1) The order of iterated integrals can be interchanged, so we also have

$$\int_{\Omega_2} \left[\int_{\Omega_1} Y(\omega_1, \omega_2) d\mathbb{P}_1(\omega_1) \right] d\mathbb{P}_2(\omega_2) = \int_{\Omega} Y d\mathbb{P}.$$

In particular, the two iterated integrals are equal.

(2) Both iterated integrals may exist but still not be equal! This is a common mistake in applying Fubini's theorem, forgetting to check that Y is integrable w.r.t. the product measure.

The essential idea is to prove the statements for indicator random variables, and hence for simple random variables Y by linearity. From there use MCT to prove it for positive random variables and take differences to prove it for integrable random variables. The key step is the first one, proving it for indicator random variables. That step is obvious for rectangles $A = A_1 \times A_2$, but not so obvious for general $A \in \mathcal{F}$. We just sketch how to go about this part and leave the rest as exercise.

Proof of Fubini-Tonelli theorem for indicators. Let $A \in \mathcal{F}$. We must show that (a) for any $\omega_1 \in \Omega_1$, the *section* $A_{\omega_1} = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in A\}$ is in $\mathcal{F}_2$, (b) $\omega_1 \to \mathbb{P}_2(A_{\omega_1})$ is $\mathcal{F}_1$-measurable, (c) $\int_{\Omega_1} \mathbb{P}_2(A_{\omega_1}) d\mathbb{P}_1(\omega_1) = \mathbb{P}(A)$.

Let $\mathcal{G} = \{A \in \mathcal{F} : \text{ (a), (b), (c) hold}\}$. Then $\mathcal{G}$ contains the π-system of rectangles. If we show that $\mathcal{G}$ is a λ-system, the $\pi - \lambda$ theorem then implies that $\mathcal{G} = \mathcal{F}$.

Suppose $A, B \in \mathcal{G}$ and $A \subseteq B$. Then $(B \setminus A)_{\omega_1} = B_{\omega_1} \setminus A_{\omega_1}$ (also a proper difference) and hence in $\mathcal{F}_2$, for any $\omega_1 \in \Omega_1$. Hence, $\mu_2((B \setminus A)_{\omega_1}) = \mu_2(B_{\omega_1}) - \mu_2(A_{\omega_1})$, a difference of two $\mathcal{F}_1$-measurable functions, hence $\mathcal{F}_1$-measurable. By the linearity of expectations,

$$\begin{aligned} \int_{\Omega_1} \mu_2((B \setminus A)_{\omega_1}) d\mathbb{P}_1(\omega_1) &= \int_{\Omega_1} \mu_2(B_{\omega_1}) d\mathbb{P}_1(\omega_1) - \int_{\Omega_1} \mu_2(A_{\omega_1}) d\mathbb{P}_1(\omega_1) \\ &= \mu(B) - \mu(A) \;=\; \mu(B \setminus A). \end{aligned}$$

Thus $\mathcal{G}$ is closed under proper differences.

Suppose $A_n \in \mathcal{G}$ and $A_n \uparrow A$. Of course $A \in \mathcal{F}$ and $(A_n)_{\omega_1} \uparrow A_{\omega_1}$ for each $\omega_1 \in \Omega_1$. Therefore, $\mu_2((A_n)_{\omega_1}) \uparrow \mu_2(A_{\omega_1})$ pointwise on Ω_1. As a limit of measurable functions, $\omega_1 \mapsto \mu_2(A_{\omega_1})$ is measurable, and MCT tells us that

$$\begin{aligned}\int_{\Omega_1} \mu_2(A_{\omega_1}) d\mathbb{P}_1(\omega_1) &= \lim_{n\to\infty} \int_{\Omega_1} \mu_2((A_n)_{\omega_1}) d\mathbb{P}_1(\omega_1) \\ &= \lim_{n\to\infty} \mu(A_n) \\ &= \mu(A).\end{aligned}$$

Thus $\mathcal{G}$ is closed under increasing limits, completing the proof that $\mathcal{G}$ is a λ-system. ■

6.2. Probability measures are special. Infinite products of measures only makes sense for probability measures. For example, suppose λ_a denotes Lebesgue measure on $[0, a]$ with total mass a. Can we construct the product measure $\lambda_a \otimes \lambda_a \otimes \ldots$? What does it even mean? If we ask for a μ on $[0, a]^{\mathbb{N}}$ such that

$$\mu(A_1 \times A_2 \times \ldots) = \mu(A_1)\mu(A_2)\ldots,$$

then all finite dimensional cylinders get infinite measure if $a \neq 1$ (as we $A_n = [0, a]$ for all large n, we get a product of a infinitely many times). One might try to salvage the situation by asking for the cylinder set $A = A_1 \times \ldots \times A_n \times [0, a] \times [0, a] \times \ldots$ to have measure equal measure μ such that $\mu(A) = \mu(A_1) \ldots \mu(A_n)$. But we can also write A as $A_1 \times \ldots \times A_{n+1} \times [0, a] \times \ldots$ with $A_{n+1} = [0, a]$, and the requirement then would be that $\mu(A) = \mu(A_1) \ldots \mu(A_{n+1}) = a\mu(A_1) \ldots \mu(A_n)$. The two requirements are inconsistent if $a \neq 1$. The case $a = 1$ is fine, as we have already constructed infinite product of probability measures. Thus, infinite products are a special feature of probability measures, and it is at this point that probability theory diverges from general measure theory and becomes a much richer subject!

However, finite products do make sense for sigma-finite measures. It is usually done in measure theory class (by the Caratheodory construction, what else?), but one can easily deduce it from the existence of products of probability measures. Indeed, if μ_i are finite nonzero measures on $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2$, then we can write $\mu_i = a_i \mathbb{P}_i$, where $a_i = 1/\mu_i(\Omega_i)$ are positive numbers and $\mathbb{P}_i(\cdot) = \mu_i(\cdot)/a_i$ are probability measures. Then we may simply define $\mu_1 \otimes \mu_2$ as $a_1 a_2(\mathbb{P}_1 \otimes \mathbb{P}_2)$.

For sigma-finite measures μ_i, we partition $\Omega_i = \sqcup_{k\geqslant 1}\Omega_{i,k}$ where $\mu_i(\Omega_{i,k}) < \infty$. Then we can define the finite measures $\mu_{i,k}(\cdot) = \mu_i(\cdot \cap \Omega_{i,k})$ (note that it is supported on $\Omega_{i,k}$) so that $\mu_i = \sum_k \mu_{i,k}$. We can then define

$$\mu = \sum_{k\geqslant 1} \sum_{\ell \geqslant 1} \mu_{1,k} \otimes \mu_{2,\ell}.$$

To check that this has the defining property of product measure, let $A_i \in \mathcal{F}_i$, and write $A_i = \sqcup A_{i,k}$, where $A_{i,k} = A_i \cap \Omega_{i,k}$. Then $A = \sqcup_{k,\ell} A_{1,k} \times A_{2,\ell}$, and $\mu_{1,k'} \otimes \mu_{2,\ell'}(A_{1,k} \times A_{2,\ell}) = \mu_1(A_{1,k})\mu_2(A_{2,\ell})$ if $k = k'$ and $\ell = \ell'$ and zero otherwise. From this it follows that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$.

Finally, the Fubini-Tonelli theorem continues to hold, and in fact follows from the corresponding theorem for probability measures.

7. Kolmogorov's consistency theorem

A generalization of the theorem on the existence of product measures is to go beyond independence. To motivate it, consider the following question. Given three Borel probability measures $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathbb{R})$, $i \leqslant 3$, does there exist a probability space and three random variables X_1, X_2, X_3 such that $X_i \sim \mu_i$? The answer is trivially yes, for example we can take three independent random variables X_1, X_2, X_3 such that $X_i \sim \mu_i$. There are other ways, for example, take one uniform random variable and set $X_i = G_{\mu_i}(U)$ (then X_i won't be independent).

Having disposed of that easy question, what if we specify three Borel probability measures $\nu_1, \nu_2, \nu_3 \in \mathcal{P}(\mathbb{R}^2)$ and want X_1, X_2, X_3 such that $(X_1, X_2) \sim \nu_1$, $(X_2, X_3) \sim \nu_2$ and $(X_1, X_3) \sim \nu_3$? Is it possible to find such random variables? If the first marginal of ν_1 and the first marginal of ν_3 do not agree, then it is not possible (because then we have two distinct specifications for the distribution of X_1!). This is because our specifications were internally inconsistent. The following theorem of Kolmogorov asserts that this is the only obstacle in constructing random variables with specified finite dimensional distributions.

Theorem 6: Consistency theorem (Daniell, Kolmogorov)

Let $\Omega_i = \mathbb{R}^{d_i}$ for some $d_i \geqslant 1$. For each $n \geqslant 1$ and each $1 \leqslant i_1 < i_2 < \ldots < i_n$, let $\mu_{i_1,\ldots,i_n}$ be a Borel p.m on $\Omega_{i_1} \times \ldots \times \Omega_{i_n}$. Then the following are equivalent.

(1) There exists a unique Borel probability measure μ on $\times_i \Omega_i$ such that $\mu \circ \Pi^{-1}_{i_1,\ldots,i_n} = \mu_{i_1,\ldots,i_n}$ for any $i_1 < i_2 < \ldots < i_n$ and any $n \geqslant 1$.

(2) The given family of probability measures satisfy the consistency condition

$$\mu_{i_1,\ldots,i_n}(B \times \Omega_{i_n}) = \mu_{i_1,\ldots,i_{n-1}}(B)$$

for any $B \in \mathcal{B}(\Omega_{i_1} \times \ldots \times \Omega_{i_{n-1}})$ and for any $n \geqslant 1$ and any $i_1 < i_2 < \ldots < i_n$.

We have stated the consistency theorem for Ω_i that are Euclidean spaces. It can be generalized, but some metric structure on Ω_is is needed. This is in contrast to the situation of product measures, which exist even if Ω_i have no structure.

Alternate form of the consistency condition: Suppose for each $n \geqslant 1$, we have a probability measure ν_n on $\Omega_1 \times \ldots \times \Omega_n$. Assume that $\nu_{n+1}(A_1 \times \ldots \times A_n \times \Omega_{n+1}) = \nu_n(A_1 \times \ldots \times A_n)$ for all $n \geqslant 1$ and all $A_i \in \mathcal{F}_i$. Then, for any $1 \leqslant i_1 < \ldots < i_k$ and any $n \geqslant i_k$, the probability measure $\nu_n \circ \Pi^{-1}_{i_1,\ldots,i_k}$ on $\Omega_{i_1} \times \ldots \times \Omega_{i_k}$ is the same. If we define this to be $\mu_{i_1,\ldots,i_k}$, then we get a consistent family of probability measures as required in the theorem.

The importance of the consistency theorem comes from having to construct dependent random variables such as Markov chains with given transition probabilities (see the next section). It also serves as a starting point for even more subtle questions such as constructing stochastic processes such as Brownian motion.

7.1. A more general consistency question. It clears things up if we take a more abstract viewpoint.

Question 2: A general consistency question

Let $\mathcal{F}_i$, $i \in I$ be sigma-algebras on a set Ω and let $\mathcal{F} = \sigma(\cup_{i \in I} \mathcal{F}_i)$. Suppose μ_i are probability measures on $(\Omega, \mathcal{F}_i)$. Does there exist a probability measure μ on $(\Omega, \mathcal{F})$ such that $\mu\big|_{\mathcal{F}_i} = \mu_i$? If so, is it unique?

Some remarks.

(1) It does not make sense to take $\mathcal{F}$ to be larger than $\sigma(\cup_{i \in I} \mathcal{F}_i)$. In general, a measure cannot be extended from a smaller sigma-algebra to a larger one (otherwise we would extend all measures to the power set!).

(2) An obvious necessary condition for the existence of μ is that μ_i and μ_j agree on $\mathcal{G}_i \cap \mathcal{G}_j$ for $i, j \in I$.

(3) If $\Omega = \prod_k \Omega_k$ and $\mathcal{F}_{i_1,\ldots,i_n} = \sigma(\Pi_{i_1}, \ldots, \Pi_{i_n})$ gives the setting of the Kolmogorov consistency theorem.

How would we try to prove the existence of such a μ? We make one extra assumption (which is clearly satisfied in the setting of the Kolmogorov consistency theorem) in addition to the consistency conditioned mentioned earlier.

Assumptions:

(1) If $A \in \mathcal{G}_i \cap \mathcal{G}_j$, then $\mu_i(A) = \mu_j(A)$.

(2) For any $i, j \in I$, there is some $k \in I$ such that $\mathcal{G}_i \cup \mathcal{G}_j \subseteq \mathcal{G}_k$.

Under the second assumption, $\mathcal{A} := \cup_i \mathcal{G}_i$ is an algebra that generates the sigma-algebra $\mathcal{F}$. We must define $\mu : \mathcal{G} \to [0, 1]$ by $\mu(A) = \mu_i(A)$ if $A \in \mathcal{G}_i$ for any $i \in I$, or what is the same, $\mu_{j_n}(B_n) \downarrow 0$. Because of the first assumption, this is a valid definition.

In view of the Caratheodory extension theorem, there is a unique extension of μ to $\mathcal{F}$ if and only if μ is countably additive on $\mathcal{A}$. As finite additivity is clear, this means checking that if $A_n, A \in \mathcal{A}$ and $A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$. If $A_n \in \mathcal{F}_{i_n}$ and $A \in \mathcal{F}_i$, we can find j_n such that $\mathcal{F}_{j_n} \supseteq \mathcal{F}_i \cup \mathcal{F}_{i_n}$ so that $B_n := A \setminus A_n \in \mathcal{F}_{j_n}$. What we need to check is that $\mu(B_n) \downarrow 0$. We write this as a conclusion.

Conclusion: Under the above assumptions, there is a unique probability measure μ on $\mathcal{F}$ that extends each μ_i if and only if whenever $B_n \in \mathcal{F}_{i_n}$ and $B_n \downarrow \emptyset$ we have $\mu_{i_n}(B_n) \downarrow 0$.

Proof of the Daniell-Kolmogorov consistency theorem. By the conclusion reached above in the general consistency theorem, the only point to check (a little reindexing may be needed first) is that if $B_n = A_{n,1} \times \ldots \times A_{n,n} \times \Omega_{n+1} \times \Omega_{n+2} \times \ldots$ for some $A_{n,i} \in \mathcal{B}(\Omega_i)$ and $B_n \downarrow \emptyset$, then $\nu_n(A_{n,1} \times \ldots \times A_{n,n}) \downarrow 0$, where $\nu_n = \mu_{1,\ldots,n}$.

Case-1: Assume that each μ_n is supported on a compact subset $K_n \subseteq \Omega_n$.

To check the condition above, assume to the contrary that $\nu_n(A_{n,1} \times \ldots \times A_{n,n}) \geqslant p$ for some $p > 0$, for all n, where $A_{n,j} \subseteq K_j$ for all j, n. By the regularity of ν_n, we can find compact $C_n \subseteq A_{n,1} \times \ldots \times A_{n,n}$ such that $\nu_n(C_n) \geqslant p/2$. As continuous images of compact sets are compact, it follows that $\Pi_j(C_n) \subseteq K_j$ is compact for each j. By a diagonal argument, we can get a subsequence n_r such that

Set $D_n = \cap_{k=1}^n (C_k \times \Omega_{k+1} \times \ldots \times \Omega_n)$. Then $D_n \subseteq C_n$ is also compact and

$$\nu_n(D_n) \geqslant \nu_n(A_{n,1} \times \ldots \times A_{n,n}) - \sum_{k=1}^{n} \nu_k(A_{k,1} \times \ldots \times A_{k,k})$$

Observe that $\mu_1(\Pi_1(C_n)) \geqslant \nu_n(C_n) \geqslant (1 - \frac{1}{2^n})p$. Thus, $\Pi_1(C_n)$ is a sequence of compact subsets of $\mathbb{R}$, and the

compact $C_{n,i} \subseteq A_{n,i}$ for $i \leqslant n$ such that

$$\nu_n(C_{n,1} \times \ldots \times C_{n,n}) \geqslant 0.5\nu_n(A_{n,1} \times \ldots \times A_{n,n}).$$

As $A_{n,j} \downarrow \emptyset$ for each j (because $B_n \downarrow \emptyset$), it follows that $C_{n,j} \downarrow \emptyset$ for each j. By compactness, there is a n_j such that $C_{n,j} = \emptyset$ for $n \geqslant n_j$. This is not good enough. We want an N such that $C_{n,j} = \emptyset$ for $n \geqslant N$. This follows by a diagonal argument. Complete the details ■

Remark 8

The proof of the consistency theorem does require some topology. The existence of product measure does not.

8. Applications of the consistency theorem

8.1. Markov chains. Consider $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and let $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ and let $\kappa : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \mapsto \mathbb{R}_+$ be a transition kernel. This means that $y \mapsto \kappa(x, \cdot)$ is a Borel probability measure function for each $x \in \mathbb{R}^d$ and $x \mapsto \kappa(x, A)$ is Borel measurable for each $A \in \mathcal{B}(\mathbb{R}^d)$. Then, define for each $n \geqslant 1$, a probability measure on $(\mathbb{R}^d)^n$ by

$$\nu_n(A_0 \times A_1 \times \ldots \times A_{n-1}) = \int_{A_0} \int_{A_1} \ldots \int_{A_{n-1}} \kappa(x_{n-2}, dx_{n-1}) \kappa(x_{n-3}, dx_{n-2}) \ldots \kappa(x_0, dx_1) d\mu(x_0).$$

for any $A_i \in \mathcal{B}(\mathbb{R}^d)$. It may be easier to parse this expression if we assume that all the measures μ_0 and $\kappa(x, \cdot)$ are absolutely continuous to one measure θ. In this case, write $d\mu_0(x) = \rho(x) d\theta(x)$ and $\kappa(x, dy) = p(x, y) d\theta(y)$ and then

$$\begin{aligned} &\nu_n(A_0 \times A_1 \times \ldots \times A_{n-1}) \\ &= \int_{A_0} \int_{A_1} \ldots \int_{A_{n-1}} p(x_{n-2}, x_{n-1}) p(x_{n-3}, x_{n-2}) \ldots p(x_0, x_1) \rho(x_0) \, d\theta(x_{n-1}) \ldots d\theta(x_0). \end{aligned}$$

That is, ν_n has density $\rho(x_0) p(x_0, x_1) \ldots p(x_{n-2}, x_{n-1})$ with respect to $\theta^{\otimes n}$.

It is easy to check that ν_n defines a probability measure on $(\mathbb{R}^d)^n$ and also that $\nu_{n+1}(A_0 \times \ldots \times A_{n-1} \times \mathbb{R}^d) = \nu_n(A_0 \times \ldots \times A_{n-1})$. Consequently, by the alternate form of the consistency condition stated above, we see that there is a probability measure μ on $(\mathbb{R}^d)^{\mathbb{N}}$ (endowed with the Borel/cylinder sigma algebra) such that $\mu \circ \Pi_{0,1,\ldots,n-1}^{-1} = \nu_n$. This measure μ on $\mathbb{R}^{\mathbb{N}}$ is what is called a *Markov chain* with state space $\mathbb{R}^d$, transition kernel p and initial distribution μ_0.

8.2. Gaussian processes. Suppose $m : \mathbb{Z} \to \mathbb{R}$ and $\sigma : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$. A Gaussian process with mean $\mu(\cdot)$ and covariance $\sigma = (\sigma_{i,j})_{i,j \in \mathbb{Z}}$ is a collection of jointly Gaussian random variables $(X_n)_{n \in \mathbb{Z}}$ such that $\mathbb{E}[X_n] = \mu(n)$ and $\mathrm{Cov}(X_n, X_m) = \sigma(m, n)$.

Question: Does it exist?

First let us note some necessary conditions. If we could construct a Gaussian process Y with mean 0 and we set $X = m + Y$ (i.e., $X_n = m(n) + Y_n$) has mean $m(\cdot)$ and the same covariance as Y. Hence the mean poses no challenge and we assume that it is zero henceforth.

The covariance is more subtle. For example, $\sigma(n, n) = \mathbb{E}[X_n^2]$ cannot be negative. More generally, for any $n \geqslant 0$ and $i_1 < \ldots < i_n$ and any $c_1, \ldots, c_n \in \mathbb{R}$, we must have

$$0 \leqslant \mathbb{E}[(c_1 X_{i_1} + \ldots + c_n X_{i_n})^2] \;=\; \sum_{p,q=1}^{n} c_p c_q \mathbb{E}[X_{i_p} X_{i_q}] \;=\; \sum_{p,q=1}^{n} c_p c_q \sigma(i_p, i_q).$$

Thus, every principal finite sub-matrix of σ must be positive semi-definite. We now claim that this is also sufficient.

Assume that σ is positive definite in the above sense. Then for any $n \geqslant 1$ and any $i_1 < \ldots < i_n$, the measure $\mu_{i_1,\ldots,i_n} = N_n(0, (\sigma(i_p, i_q))_{p,q,\leqslant n})$ is well-defined. This is because positive definiteness allows us to write

$$(\sigma(i_p, i_q))_{p,q\leqslant n} = BB^t$$

for a $n \times n$ matrix B. Taking $Z_1, \ldots, Z_n$ i.i.d. $N(0,1)$, the distribution of the random vector BZ, where $Z = (Z_1, \ldots, Z_n)^t$ is the desired Gaussian distribution.

From basic properties of Gaussian distributions (marginals of Gaussians are Gaussian) it follows that the family of distributions $\{\mu_{i_1,\ldots,i_n}\}$ is consistent. Hence by the consistency theorem, the Gaussian process with covariance σ exists.

8.3. Did we really need the consistency theorem? Actually no! We could have constructed Markov chains and Gaussian processes from the simpler fact that i.i.d. uniform random variables $V_0, V_1, V_2, \ldots$ exist. For Markov chains, to take the kth step, we can use V_k to generate a random variable from the required step distribution (depending on the current location). For Gaussian process, one can first convert V_k to $Z_k \sim N(0,1)$. Then the Gaussian process can be generated in the form $X = BZ$, where $Z = (Z_1, Z_2, \ldots)^t$ and B is an infinite, lower triangular matrix such that $BB^t = \sigma$ (here the indexing set is $\mathbb{N}$ instead of $\mathbb{Z}$ which of course makes no difference). As B is lower triangular, observe that in defining any entry of BZ or BB^t, only finite sums and products are needed, so there is no convergence issue.

In fact, every situation of interest to probabilists can be generated from a sequence of independent random variables, and hence on the probability space $([0,1], \mathcal{B}, \lambda)$. The idea is that we construct i.i.d. uniforms $U_1, U_2, \ldots$ and then set $X_n = f_n(U_n, X_1, \ldots, X_{n-1})$ (for $n = 1$ this means $X_1 = f_1(U_1)$) where f_n is the inverse of the cumulative distribution function of the conditional distribution of X_{n+1} given $\sigma\{X_1, \ldots, X_n\}$. We have not yet defined what conditional distribution means, but in the situations where you know what it means, it should be clear that the above procedure works.

Exercise 2

Let $S = [n]$ and let $P_{n\times n} = (p_{i,j})_{1\leqslant i,j\leqslant n}$ be a stochastic matrix (i.e., all entries are positive and row sums are 1). Show the existence of a Markov chain with transition matrix P by completing the following steps.

(1) Construct independent random variables $\xi_{i,t}$, $1 \leqslant i \leqslant n$, $t \geqslant 0$ (here t is also an integer) such that $\xi_{i,t} \sim p_{i,1}\delta_1 + \ldots + p_{i,n}\delta_n$ (the probability vector defined by the ith row of P).

(2) Define the "random mappings" $F_t : S \to S$ by $F_t(i) = \xi_{i,t}$. Then define $X_0 = i_0$ and $X_{t+1} = F_t \circ F_{t-1} \circ \ldots \circ F_0(X_0)$ for $t \geqslant 0$. Show that $(X_0, X_1, \ldots)$ is a Markov chain with transition matrix P and initial state i_0.

9. The Radon-Nikodym theorem and conditional probability

9.1. Absolute continuity and singularity. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X : \Omega \to \mathbb{R}$ be a non-negative random variable with $\mathbb{E}[X] = 1$. Define $\mathbf{Q}(A) = \mathbb{E}[X\mathbf{1}_A]$ for $A \in \mathcal{F}$. Then, $\mathbf{Q}$ is a probability measure on $(\Omega, \mathcal{F})$. Finite additivity is clear, by linearity of expectation. MCT shows that if $A_n, A \in \mathcal{F}$ and $A_n \uparrow A$ then $\mathbf{Q}(A_n) \uparrow \mathbf{Q}(A)$.

All this clearly remains valid even if $\mathbb{P}$ was an infinite measure and X was a general non-negative measurable function, except that $\mathbf{Q}$ is possibly an infinite measure too. One can think of $\mathbf{Q}$ as got from $\mathbb{P}$ by re-weighting the space according to the values of X. We say that $\mathbf{Q}$ has *density* X with respect to $\mathbb{P}$.

Question: Given two measures μ, ν on $(\Omega, \mathcal{F})$, does ν have a density with respect to μ and is it unique?

The uniqueness part is easy.

Proof of uniqueness. If f and g are two densities, then $\nu(A) = \int_A f d\mu = \int_A g d\mu$ for some f, g, then $h := f - g$ satisfies $\int_A h d\mu = 0$ for all $A \in \mathcal{F}$. Take $A = \{h > 0\}$ to get $\int h\mathbf{1}_{h>0} d\mu = 0$. But $h\mathbf{1}_{h>0}$ is a non-negative measurable function, hence it must be that $h\mathbf{1}_{h>0} = 0$ a.s.$[\mu]$. This implies that $\mu\{h > 0\} = 0$. Similarly $\mu\{h < 0\} = 0$ and we see that $h = 0$ a.s.$[\mu]$ or equivalently $f = g$ a.s$[\mu]$. The density is unique up to sets of μ-measure zero. More than that cannot be asked because, if f is a density and $g = f$ a.s.$[\mu]$, then it follows that $\int_A g d\mu = \int_A f d\mu$ and hence g is also a density of ν with respect to μ. ■

Existence of density is a more subtle question. First let us see some examples.

Example 5

On $([0,1], \mathcal{B}, \lambda)$ let ν be the measure with distribution $F_\nu(x) = x^2$. Then ν has density $f(x) = 2x\mathbf{1}_{x\in[0,1]}$ with respect to λ. Indeed, if we set $\theta(A) = \int_A f d\lambda$, then θ and ν are two measures on $[0,1]$ that agree on all intervals, since $\int_{[a,b]} f d\lambda = b^2 - a^2$ for any $[a,b] \subseteq [0,1]$. By the $\pi - \lambda$ theorem, $\theta = \nu$.

Note that the same logic works whenever $\nu \in \mathcal{P}(\mathbb{R})$ and F_ν has a continuous (or piecewise continuous) derivative. If $f = F_\nu'$, by the fundamental theorem of Calculus, $\int_{[a,b]} f d\lambda = F_\nu(b) - F_\nu(a)$ and hence by the same reasoning as above, ν has density f with respect to Lebesgue measure.

Example 6

Let Ω be some set and let $a_1, \ldots, a_n$ be distinct elements in Ω. Let $\nu = \sum_{k=1}^n p_k \delta_{a_k}$ and let $\mu = \sum_{k=1} q_k \delta_{a_k}$ where p_i, q_i are non-negative numbers such that $\sum_i p_i = \sum_i q_i = 1$.

Assume that $q_i > 0$ for all $i \leqslant n$. Then define $f(x) = \frac{p_i}{q_i}$ for $x = a_i$ and in an arbitrary fashion for all other $x \in \Omega$. Then, f is the density of ν with respect to μ. The key point is that $\int f\mathbf{1}_{\{a_i\}} d\mu = f(a_i)\mu\{a_i\} = p_i = \nu\{a_i\}$.

On the other hand, if $q_i = 0 < p_i$ for some i, then ν cannot have a density with respect to μ (why?).

Let us return to the general question of existence of density of a measure ν with respect to a measure μ (both measures are defined on $(\Omega, \mathcal{F})$). As in the last example, there is one necessary condition for the existence of density. If $\nu(A) = \int f\mathbf{1}_A d\mu$ for all A, then if $\mu(A) = 0$ we must have $\nu(A) = 0$ (since $f\mathbf{1}_A = 0$ a.s$[\mu]$). In other words, if there is even one set $A \in \mathcal{F}$ such that $\nu(A) > 0 = \mu(A)$, then ν cannot have a density with respect to μ. Let us make a definition.

Definition 8

Two measures μ and ν on the same $(\Omega, \mathcal{F})$ are said to be *mutually singular* and write $\mu \perp \nu$ if there is a set $A \in \mathcal{F}$ such that $\mu(A) = 0$ and $\nu(A^c) = 0$. We say that μ is *absolutely continuous to* ν and write $\mu \ll \nu$ if $\mu(A) = 0$ whenever $\nu(A) = 0$.

Remark 9

(1) Singularity is a symmetric relation, absolute continuity is not. If $\mu \ll \nu$ and $\nu \ll \mu$, then we say that μ and ν are *mutually absolutely continuous*. (2) If $\mu \perp \nu$, then we cannot also have $\mu \ll \nu$ (unless $\mu = 0$). (3) Given μ and ν, it is not necessary that they be singular or absolutely continuous to one another. (4) Singularity is not reflexive but absolute continuity is. That is, $\mu \ll \mu$ but μ is never singular to itself (unless μ is the zero measure).

Example 7

Uniform$([0,1])$ and Uniform$([1,2])$ are singular. Uniform$([1,3])$ is neither absolutely continuous nor singular to Uniform$([2,4])$. Uniform$([1,2])$ is absolutely continuous with respect to Uniform$([0,4])$ but not conversely. All these uniforms are absolutely continuous to Lebesgue measure. Any measure on the line that has an atom (eg., δ_0) is not absolutely continuous to Lebesgue measure. A measure that is purely discrete is singular with respect to Lebesgue measure. A probability measure on the line with density (eg., $N(0,1)$) is absolutely continuous to λ. In fact $N(0,1)$ and λ are mutually absolutely continuous. However, the exponential distribution is absolutely continuous to Lebesgue measure, but not conversely (since $(-\infty, 0)$, has zero probability under the exponential distribution but has positive Lebesgue measure).

Returning to the existence of density, we saw that for ν to have a density with respect to μ, it is necessary that $\nu \ll \mu$. This condition is also sufficient!

Theorem 7: Radon Nikodym theorem

Suppose μ and ν are two finite measures on $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, then $d\nu = f d\mu$ for some $f \in L^1(\mu)$.

The function f in the statement is called the *Radon-Nikodym derivative* of ν w.r.t. μ. When both ν is a probability measure, we also call it the *density of* ν w.r.t. μ. Of particular importance is the case when ν is a probability measure on $\mathbb{R}^d$ and μ is the Lebesgue measure on $\mathbb{R}^d$.

A first attempt at proof: Let $H = L^2(\mu)$ and define $L : H \mapsto \mathbb{R}$ by $Lf = \int f d\nu$. Suppose we could show that L is well-defined (then it is clearly linear) and bounded, i.e., $|Lf| \leqslant C\|f\|_H$ for all $f \in H$. Then, by the Riesz representation theorem for linear functionals on a Hilbert space, it follows that $Lf = \langle f, \psi \rangle$ for some $\psi \in H$. Take $f = \mathbf{1}_A$ with $A \in \mathcal{F}$ to see that $\nu(A) = \int_A \psi \, d\mu$. This is what we want to show.

The problem is that L need not be bounded. Indeed, it it were true, the above argument would have shown that the Radon -Nikodym derivative of ν w.r.t. μ is in $L^2(\mu)$, which is false in general! For example, let $\nu(A) = \int_A \frac{1}{\sqrt{x}} d\lambda(x)$, where λ is the Lebesgue measure on $[0,1]$. Then the Radon-Nikodym derivative is $1/\sqrt{x}$, whose square is not integrable w.r.t. μ. The proof below overcomes this issue by a small trick.

PROOF OF THE RADON NIKODYM THEOREM. Let $\theta = \mu + \nu$ and let $H = L^2(\Omega, \mathcal{F}, \theta)$. Define $L : H \mapsto \mathbb{R}$ by $Lf = \int f d\nu$. Since (note that $\int g d\nu \leqslant \int g d\theta$ for any $g \geqslant 0$)

$$\Big|\int f d\nu\Big| \leqslant \int |f| d\nu \leqslant \int |f| d\theta \leqslant \sqrt{\theta(\Omega)}\left(\int |f|^2 d\theta\right)^{\frac{1}{2}},$$

it follows that L is well-defined and $|Lf| \leqslant C\|f\|_H$ with $C = \sqrt{\theta(\Omega)}$.Therefore, L is bounded and $Lf = \int f\varphi \, d\theta$ for some $\varphi \in H$. Rewrite this as

$$\int f(1-\varphi) d\nu = \int f\varphi d\mu \quad \text{for all } f \in H. \tag{2}$$

From this identity, it is clear that $0 \leqslant \varphi \leqslant 1$ a.s.[μ] (hence also a.s.[ν]). Further, setting $f = \mathbf{1}_{\varphi=1}$, we see that the left hand side is zero while the right hand side is $\mu\{\varphi = 1\}$. Thus, $\varphi < 1$ a.s.[μ] (hence also a.s.[ν]).

Now for any $A \in \mathcal{F}$ and $\delta > 0$, setting $f = \frac{1}{1-\varphi}\mathbf{1}_A\mathbf{1}_{\varphi\leqslant 1-\delta}$ (which is bounded above by $1/(1-\delta)$ and hence in H), we get that $\nu(A \cap \{\varphi \leqslant 1-\delta\}) = \int_A \psi\mathbf{1}_{\varphi\leqslant 1-\delta} d\mu$, where $\psi = \varphi/(1-\varphi)$. Set $\delta = 1/n$ and let $n \uparrow \infty$. We get $\nu(A \cap \{\varphi < 1\}) = \int \psi\mathbf{1}_{\varphi<1} d\mu$. Since $\varphi < 1$ almost surely with respect to both measures, it is redundant to write that, and we get $\nu(A) = \int_A \psi d\mu$. ■

Exercise 3: Lebesgue decomposition

Let μ, ν be two finite measures on $(\Omega, \mathcal{F})$. Show that we can write $\nu = \nu_1 + \nu_2$, where ν_1, ν_2 are measures on $\mathcal{F}$ and $\nu_1 \ll \mu$ and $\nu_2 \perp \mu$. This decomposition is unique. [*Hint:* Follow the steps in the proof of Radon-Nikodym theorem and consider theset $\{\varphi = 1\}$ carefully!]

9.2. Some singular probability measures. This section is not directly needed for what comes next in the course. But these are some natural directions suggested by the previous discussion of absolute continuity and singularity of measures.

Is there any $\mu \in \mathcal{P}(\mathbb{R})$ that is singular to Lebesgue measure on $\mathbb{R}$? Of course, any discrete probability measure is singular, since it gives probability one to a countable set while Lebesgue measure gives probability zero to that set. The interesting question is whether there is a singular μ that has no atoms. For this, we must spread our set on some uncountable set of zero Lebesgue measure. The first example that comes to mind is the standard Cantor set.

Recall that the middle-thirds Cantor set is defined as the decreasing intersection K of K_ns where $K_0 = [0,1]$, $K_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, $K_3 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{3}{9}] \cup [\frac{6}{9}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, and so on. In general, K_n is a union of 2^n intervals each of length 3^{-n}, and K_{n+1} is got from K_n by deleting the middle third open

subinterval of each of these intervals. An alternate description of the Cantor set is

$$K = \left\{ x \in [0,1] : x = \sum_{n=1}^{\infty} \frac{x_n}{3^n} \text{ for some } x_n \in \{0,2\} \right\}.$$

In other words, it consists of those numbers that have a ternary (base-3) expansion without using the digit 1.

Example 8: Cantor measure

Let K be the middle-thirds Cantor set. Consider the canonical probability space $([0,1], \mathcal{B}, \lambda)$ and the random variable $X(\omega) = \sum_{k=1}^{\infty} \frac{2B_k(\omega)}{3^k}$, where $B_k(\omega)$ is the kth binary digit of ω (i.e., $\omega = \sum_{k=1}^{\infty} \frac{B_k(\omega)}{2^k}$). Then X is measurable (we saw this before). Let $\mu := \lambda \circ X^{-1}$ be the pushforward measure.

Then, $\mu(K) = 1$, because X takes values in numbers whose ternary expansion has no ones. Further, for any $t \in K$, $X^{-1}\{t\}$ is a set with atmost two points and hence $\mu\{t\} = 0$. Thus μ has no atoms and must have a continuous CDF. Since $\mu(K) = 1$ but $\lambda(K) = 0$, we also see that $\mu \perp \lambda$.

Exercise 4: Alternate construction of Cantor measure

Write $K = \cap K_n$ as in the definition of the Cantor set. Let μ_n be the uniform probability measure on K_n, i.e., $\mu_n(A) = (3/2)^n \lambda(A \cap K_n)$ for all $A \in \mathcal{B}_{\mathbb{R}}$. Show that F_{μ_n}s converge uniformly to a CDF F and that the measure having this CDF is the Cantor measure constructed above.

Example 9: Bernoulli convolutions - a fun digression (omit if unclear!)

We generalize the previous example. For any $\alpha > 1$, define $X_\alpha : [0,1] \to \mathbb{R}$ by $X_\alpha(\omega) = \sum_{k=1}^{\infty} \alpha^{-k} B_k(\omega)$. Let $\mu_\alpha = \lambda \circ X_\lambda^{-1}$ (did you check that X_α is measurable?). These measures are called Bernoulli convolutions. For $\alpha = 3$, this is almost the same as 1/3-Cantor measure, except that we have left out the irrelevant factor of 2 (so μ_3 is a probability measure on $\frac{1}{2}K := \{x/2 : x \in K\}$) and hence is singular. For $\alpha = 2$, the map X_α is identity, and hence μ_2 is the Lebesgue measure on $[0,1]$, certainly absolutely continuous to Lebsegue measure. What about the singularity and absolute continuity of μ_α for other values of α?

Exercise 5

For any $\alpha > 2$, show that μ_α is singular w.r.t. Lebesgue measure.

Hence, one might expect that μ_α is absolutely continuous to Lebesgue measure for $1 < \alpha < 2$. This is false! Paul Erdős showed that μ_α is singular to Lebesgue measure whenever α is a Pisot-Vijayaraghavan number, i.e., if α is an algebraic number all of whose conjugates have modulus less than one!! It is an open question as to whether these are the only exceptions.

9.3. Hausdorff measures. Consider two Cantor type sets: A consisting of those numbers who decimal expansion does not have the disgit 5 and B consisting of those numbers who decimal expansion does not have any odd digit. Both have Lebesgue measure zero. Is there another measure that can measure the sizes of these sets (one might feel that B is somehow smaller than A, but in what sense?).

Let (X, d) be a compact metric space. Fix $\alpha > 0$ and define for any $A \subseteq X$,

$$H_\alpha^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \mathrm{dia}(B_n)^\alpha : B_n \text{ are open balls whose union covers } A \right\}.$$

It is easy to check that $H_\alpha^*(A) \leqslant H_\alpha^*(B)$ if $A \subseteq B$ and $H_\alpha^*(\cup_n A_n) \leqslant \sum_n H_\alpha^*(A_n)$. Thus H_α^* is an outer measure H_α and can be used to construct a measure on $(X, \mathcal{B}_X)$ (one must check many things, for example that the Caratheodary construction gives a sigma algebra containing all Borel sets). As it happens, for most α, the measure H_α turns out to be trivial. For example, if $X = [0,1]$, then for any interval I, one can check that $H_\alpha(I) = 0$ if $\alpha > 1$ and $H_\alpha(I) = \infty$ if $\alpha < 1$. For $\alpha = 1$, we get the Lebesgue measure.

For a general X, again there is always a value α_0 such that for any open ball B we have $H_\alpha(B) = 0$ if $\alpha > \alpha_0$ and $H_\alpha(B) = \infty$ if $\alpha < \alpha_0$. At $\alpha = \alpha_0$, we may or may not get a meaningful measure. If we do, then H_{α_0} is called the *Hausdorff measure* on X. Whether H_{α_0} is trivial or not, the number α_0 is called the *Hausdorff dimension* of X.

Example 10

Let $X = K$, the middle-thirds Cantor set. Then $\alpha_0 = \log 2/\log 3$ and H_{α_0} is precisely the Cantor measure that we constructed earlier.

9.4. Conditional probability and expectation - a first view. So far (and for a few lectures next), we have seen how a rigorous framework for probability theory is provided by measure theory. We have not yet touched the two most important concepts in probability, *independence* and *conditional probability*. We shall see independence very shortly but may not have time to study conditional probability in detail in this course. But one of the important aspects of Kolmogorov's axiomatization of probability using measure theory was to define conditional probability using the Radon-Nikodym theorem. Here is a teaser for that story.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let X be a random variable that takes finitely many values $a_1, \ldots, a_n$ with $\mathbb{P}\{X = a_k\} > 0$ for each k. Then, the law of total probability says that for any $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(A \,\big|\, X = a_k)\mathbb{P}\{X = a_k\}$$

where $\mathbb{P}(A \,\big|\, X = a_k) = \frac{\mathbb{P}\{A \cap \{X = a_k\}\}}{\mathbb{P}\{X = k\}}$. Now suppose X takes uncountably many values, for eg., X has density f_X. Then, we would like to write

$$\mathbb{P}(A) = \int \mathbb{P}(A \,\big|\, X = t) f_X(t)\, dt$$

where f_X is the density of X and perhaps even generalize it to the case when X does not have density as $\mathbb{P}(A) = \int \mathbb{P}(A \,\big|\, X = t)\, d\mu_X(t)$. The question is, what is $\mathbb{P}(A \,\big|\, X = t)$? The usual definition makes no sense since $\mathbb{P}\{X = t\} = 0$.

The way around is this. Fix $A \in \mathcal{F}$ and set $\nu_A(I) = \mathbb{P}\{A \cap \{X \in I\}\}$ for $I \in \mathcal{B}_{\mathbb{R}}$. Then ν is a Borel probability measure on as a measure on $\mathbb{R}$. If μ_X is the distribution of X, then clearly $\nu_A \ll \mu_X$ (if $\mu_X(I) = 0$ then $\mathbb{P}\{X \in I\} = 0$ which clearly implies that $\nu_A(I) = 0$). Hence, by the Radon-Nikodym theorem, ν_A has a density $f_A(t)$ with respect to μ_X. In other words,

$$\mathbb{P}(A \cap \{X \in I\}) = \int_I f_A(t)\, d\mu_X(t)$$

and in particular, $\mathbb{P}(A) = \int_{\mathbb{R}} f_A(t)\, d\mu_X(t)$. Then, we may *define* $f_A(t)$ as the conditional probability of A given $X = t$! Note that f_A is defined only almost everywhere, hence $\mathbb{P}(A \,\big|\, X = t)$ should also be interpreted as being defined for almost every t (w.r.t. μ_X). This way, the intuitive notion of conditional probability is brought into the ambit of measure theoretical probability. We now elaborate on this a bit.

Let $\mathbb{P}, \mathbf{Q}$ be probability measures on $(\Omega, \mathcal{F})$. Assume that $\mathbf{Q} \ll \mathbb{P}$. Then there is a $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\mathbf{Q}(A) = \int_A X d\mathbb{P} \quad \text{for all } A \in \mathcal{F}.$$

Now suppose $\mathcal{G} \subseteq \mathcal{F}$ is a sub-sigma algebra. Let $\mathbb{P}', \mathbf{Q}'$ be the restrictions of $\mathbb{P}, \mathbf{Q}$ to $\mathcal{G}$. It is trivially the case that $\mathbf{Q}' \ll \mathbb{P}'$. Hence, again by the Radon-Nikodym theorem, there is some $X' \in L^1(\Omega, \mathcal{G}, \mathbb{P}')$ such that $\mathbf{Q}'(A) = \int_A X' d\mathbb{P}'$ for all $A \in \mathcal{G}$. The last statement can also be written as

$$\mathbf{Q}(A) = \int_A X' d\mathbb{P} \quad \text{for all } A \in \mathcal{G}.$$

This X' is not the same as X, because the latter need not be $\mathcal{G}$-measurable.

Now start with any integrable random variable Y on $(\Omega, \mathcal{F}, \mathbb{P})$. Writing as $Y_+ - Y_-$ and applying the above steps to find Y'_+, Y'_- (these are $\mathcal{G}$-measurable and give the same integrals as Y_+, Y_- over sets in $\mathcal{G}$). Writing $Y' = Y'_+ - Y'_-$, we have shown that there is a $\mathcal{G}$-measurable random variable Y' such that

$$\int_A Y \, d\mathbb{P} = \int_A Y' \, d\mathbb{P} \quad \text{for all } A \in \mathcal{G}.$$

This Y' is called the conditional expectation of Y w.r.t. $\mathcal{G}$ and denoted $\mathbb{E}[Y \mid \mathcal{G}]$.

Example 11

Again consider $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable partition $\{A_1, \ldots, A_k\}$ with $\mathbb{P}(A_i) > 0$ for all i. Let $\mathcal{G} = \sigma\{A_1, \ldots, A_k\}$. If Y is an integrable random variable ($\mathcal{F}$-measurable), we compute $Y' = \mathbb{E}[Y \mid \mathcal{G}]$. Since Y' is $\mathcal{G}$-measurable, we can write $Y' = \alpha_1 \mathbf{1}_{A_1} + \ldots + \alpha_k \mathbf{1}_{A_k}$. Equating its integral over A_i with that of Y, we arrive at $\alpha_i \mathbb{P}(A_i) = \int_{A_i} Y d\mathbb{P}$. Thus,

$$Y' = \sum_{i=1}^{k} \left(\frac{1}{\mathbb{P}(A_i)} \int_{A_i} Y d\mathbb{P} \right) \mathbf{1}_{A_i}.$$

The value of α_i is what you would have seen in basic probability class as the expected value of Y given A_i (just restrict the probability measure to A_i and renormalize by dividing by $\mathbb{P}(A_i)$. Then take expectation of Y w.r.t this new measure).

Example 12

Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, having a joint density $f(x, y)$ on $\mathbb{R}$. We want to talk of $\mathbb{E}[Y \mid X = x]$. For this, we take $\mathcal{G} = \sigma(X)$, the sigma-algebra generated by X and compute $\mathbb{E}[Y \mid \mathcal{G}]$. What are $\mathcal{G}$-measurable random variables? They are precisely those of the form $\varphi(X)$ for some Borel measurable $\varphi : \mathbb{R} \mapsto \mathbb{R}$ (why?). Let us simply write down the formula and check that it works: $Y' = \varphi(X)$ where

$$\varphi(x) := \begin{cases} \frac{1}{\int_{\mathbb{R}} f(x,y)\,dy} \int_{\mathbb{R}} y f(x, y)\,dy & \text{if } \int_{\mathbb{R}} f(x, y)\,dy > 0 \\ 0 & \text{if } \int_{\mathbb{R}} f(x, y)\,dy = 0. \end{cases}$$

Clearly Y' is $\mathcal{G}$-measurable (since it is a function of X). Check that $\mathbb{E}[Y'\mathbf{1}_A] = \mathbb{E}[Y\mathbf{1}_A]$ if $A = \{Z \in B\}$ for some $B \in \mathcal{B}_{\mathbb{R}}$. That shows that $Y' = \mathbb{E}[Y \mid \mathcal{G}]$.

It may be confusing for the first time that what we call conditional expectation is a random variable and not a number. But that is indeed the point. First we conceptualize an experiment which tells us for each element of $\mathcal{G}$, whether or not it has occurred. Then depending on the outcome of the experiment, we update our probabilities of event or expectations of random variables. In other words, the update is a function of the outcome of the experiment, hence a random variable.

CHAPTER 2

Convergence of probability measures and random variables

So far, we have looked at individual probability measures and random variables. Now we look at what properties they have as a collection, in particular, in particular the sense in which they can be close to one other. Some of the main theorems we shall prove later, the weak and strong laws of large numbers and the central limit theorem are statements about such closeness. This language is helpful for all future discussions.

First we discuss the important notion of convergence of probability measures on Euclidean spaces. Then we discuss multiple modes of convergence of a sequence of random variables to another random variable.

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 $(\Omega, \mathcal{F})$). We want to measure closeness of two probability distributions. Two natural 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 $\mathcal{P}(\mathbb{R}^d)$ (check the triangle inequality).

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 $\{j/n : 1 \leqslant j \leqslant n\}$, 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. In fact, if we take any measure-determining class[1] of sets $\mathcal{C} \subseteq \mathcal{B}_{\mathbb{R}^d}$, then $D_{\mathcal{C}}(\mu, \nu) := \sup_{A \in \mathcal{C}} |\mu(A) - \nu(A)|$ is a metric on $\mathcal{P}(\mathbb{R}^d)$.

For instance, taking the class of all semi-infinite rectangles $R_x := (-\infty, x_1] \times \ldots \times (-\infty, x_d]$ with $x \in \mathbb{R}^d$ gives us 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 $D_2(\mu, \nu) \leqslant D_1(\mu, \nu)$, 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. For example, if if μ is uniform distribution on $[0, 1]$ and μ_n is uniform distribution on the finite set $\{\frac{j}{n} : 1 \leqslant j \leqslant n\}$, then $D_2(\mu, \mu_n) \leqslant \frac{1}{n}$. 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 9

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

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

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

[1] We say that $\mathcal{C}$ is measure-determining if $\mu(A) = \nu(A)$ for all $A \in \mathcal{C}$ implies that $\mu = \nu$. We have seen that any π-system that generates the Borel sigma-algebra is measure-determining.

Remark 10

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 $\mu, \nu \in \mathcal{P}(X)$

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

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 $d(\delta_a, \delta_b) \leqslant (\max_i |b_i - a_i|) \wedge 1$. 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 2

Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R}^d)$. Then, $\mu_n \xrightarrow{d} \mu$ if and only if $F_{\mu_n}(x) \to F_\mu(x)$ for all continuity points x of F_μ.

Proof. Suppose $\mu_n \xrightarrow{d} \mu$. Let $x \in \mathbb{R}^d$ and fix $u > 0$. Then for large enough n, we have $F_\mu(x + u\mathbf{1}) + u \geqslant F_{\mu_n}(x)$, hence $\limsup F_{\mu_n}(x) \leqslant F_\mu(x + u\mathbf{1}) + u$ for all $u > 0$. By right continuity of F_μ, we get $\limsup F_{\mu_n}(x) \leqslant F_\mu(x)$. Further, $F_{\mu_n}(x) + u \geqslant F_\mu(x - u\mathbf{1})$ for large n, hence $\liminf F_{\mu_n}(x) \geqslant F_\mu(x - u)$ for all u. If x is a continuity point of F_μ, we can let $u \to 0$ and get $\liminf F_{\mu_n}(x) \geqslant F_\mu(x)$. 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 $x_{i+1} \leqslant x_i + u$ 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 $i \leqslant m$ and for $n \geqslant N$. Henceforth, let $n \geqslant N$.

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

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

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

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

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

Example 13

Again, let $a_n \to a$ in $\mathbb{R}$. Then $F_{\delta_{a_n}}(t) = 1$ if $t \geqslant a_n$ and 0 otherwise while $F_{\delta_a}(t) = 1$ if $t \geqslant a$ 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*. This is helpful, for example we can check continuity of a function on the space or compactness of a subset, using sequential criteria.

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)

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

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 [2].

In the exercise below, we give other ways we could 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 of probability measures converge to which probability measure. In fact, the Kolmogorov-Smirnov and total variation distances are more meaningful (and actually used!) when one really wants to measure distances, but in restricted settings.

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

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

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

[2] 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

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

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 $F_\mu(0.5+t) \geqslant F_{\mu_n}(0.5)$, then we must have $t \geqslant 0.5$. 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 $\mu = \delta_0$.

Theorem 8

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

(1) $\mu_n \xrightarrow{d} \mu$.

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

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

(4) $\limsup_{n\to\infty} \mu_n(C) \leqslant \mu(C)$ 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 $\mu(a_k', b_k') \geqslant \mu(a_k, b_k) - \varepsilon 2^{-k}$ (possible as there are at most countably many discontinuity points). Then

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

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

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

The left side is $\liminf \mu_n(G)$ and $\varepsilon > 0$ is arbitrary, hence $\liminf \mu_n(G) \geqslant \mu(G)$. 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 $\liminf \mu_n\{f > t\} \geqslant \mu\{f > t\}$ by assumption. By Fatou's lemma,

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

If $f \geqslant 0$, then this is the same as saying $\liminf \int f d\mu_n \geqslant \int f d\mu$. For general bounded continuous f with $M = \|f\|_{\sup}$, 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 $u \leqslant x$ and $\varphi_{x,y}(u) = 0$ for $u \geqslant y$. Then

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

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

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

This is true for all $x < y$. Let $y \downarrow x$ in the first inequality to get $\limsup F_{\mu_n}(x) \leqslant F_\mu(x)$ for all x. Let $x \uparrow y$ in the second inequality to get $\liminf F_{\mu_n}(y) \geqslant F_\mu(y-)$ 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, $\mu_n \xrightarrow{d} \mu$ 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 $\mu_n \xrightarrow{d} \mu$ and $\mu(\partial A) = 0$, then show that $\mu_n(A) \to \mu(A)$.

All these conditions may be thought of as convergence of certain integrals (as $\mu(A) = \int \mathbf{1}_A \, d\mu$). When the objective is to show that $\mu_n \xrightarrow{d} \mu$, then we would like the collection of integrals to check to be as small as possible. From this point of view, in condition 5 of Theorem 8, can we replace $C_b(\mathbb{R}^d)$ by $C_c(\mathbb{R}^d)$ (compactly supported continuous functions) or even $C_c^\infty(\mathbb{R}^d)$ (smooth ones)?

If μ is not assumed to be a probability measure, then it need not be true, as the example of $\mu_n = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_n$ and $\mu = \frac{1}{2}\delta_0$ shows. On the other hand, if we already assume that μ is a probability measure, then the statement is true. This is because the sequence is tight and we can find a compact set $K = [-M, M]^d$ such that $\mu_n(K) \geqslant 1 - \varepsilon$ for all n and $\mu(K) \geqslant 1 - \varepsilon$. Given any $f \in C_b(\mathbb{R}^d)$, replace it by $g \in C_c(\mathbb{R})$ such that $f = g$ on K. Then

$$\left| \int f \, d\mu_n - \int g \, d\mu_n \right| \leqslant \|f\|_{\sup} \mu_n(K^c) \leqslant \varepsilon \|f\|_{\sup}$$

and a similar inequality for μ. As $\int g \, d\mu_n \to \int g \, d\mu$ (by assumption as $g \in C_c(\mathbb{R}^d)$) and as ε is arbitrary, we get $\int f \, d\mu_n \to \int f \, d\mu$ for all $f \in C_b(\mathbb{R}^d)$. As $C_c(\mathbb{R}^d)$ functions can be approximated uniformly by $C_c^\infty(\mathbb{R}^d)$, it also suffices to check the convergence for smooth compactly supported functions (the details are left as exercise).

Remark 11

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 finite 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 space 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. 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 $\mu_n \xrightarrow{d} \mu$.

(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 $X_n \xrightarrow{a.s.} X$. 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 8. 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 $\mu_n \xrightarrow{d} \mu$. 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 $L(\mu_n) - \frac{1}{n} \leqslant \inf_\nu L(\nu)$. 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 7: 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. **Step-1**: Getting the subsequence $\{n_\ell\}$. 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 $j \leqslant \ell$.

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 $n_{\ell,\ell}$.

Step-2: Getting the function F. Define

$$F(x) := \inf\{\alpha_j : j \text{ for which } x_j > x\}.$$

F is well defined, takes values in $[0,1]$ and is increasing. 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 $\lim_{n\to\infty} F(y_n) \leqslant \alpha_j$. Take infimum over all j such that $x_j > y$ to get $\lim_{n\to\infty} F(y_n) \leqslant F(y)$. Of course $F(y) \leqslant \lim F(y_n)$ as F is increasing. This shows that $\lim F(y_n) = F(y)$ and hence F is right continuous.

Step-3: Proving the convergence. 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

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

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

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

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

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 14

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 10

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 $\mu(K_\varepsilon) \geqslant 1 - \varepsilon$ for all $\mu \in \mathcal{A}$.

Example 15

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 $\mu([-n, n]^d) \geqslant 1 - \varepsilon$. 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 9

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 7, 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 $\mu_{n_\ell}[-A, A] \geqslant 1-\varepsilon$ and such that A is a continuity point of F. Then, $F_{n_\ell}(-A) \leqslant \varepsilon$ and $F_{n_\ell}(A) \geqslant 1-\varepsilon$ for every n and by taking limits we see that $F(-A) \leqslant \varepsilon$ and $F(A) \geqslant 1-\varepsilon$. 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 2 it also follows that $\mu_{n_\ell} \xrightarrow{d} \mu$.

(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 $F_{\mu_k}(k) \leqslant 1-\varepsilon$ or/and $F_{\mu_k}(-k) \geqslant \varepsilon$. 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 $F_{\mu_k}(-k) \geqslant \varepsilon$ or there are infinitely many k for which $F_{\mu_k}(k) \geqslant 1-\varepsilon$. Suppose the former is true. Then, for any $x \in \mathbb{R}$, since $-k < x$ for large enough k, we see that $F_{\mu_k}(x) \geqslant F_{\mu_k}(-k) \geqslant \varepsilon$ for large enough k. This means that if F_{μ_k} converge to some F (at continuity points of F), then $F(x) \geqslant \varepsilon$ 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

One of the primary objects of study will be the sample averages $(X_1 + \ldots + X_n)/n$, where X_k are i.i.d. random variables. Laws of large numbers state that these sample averages are close to the mean of X_1, but there are multiple ways this could be made precise. Here we try to understand the different senses in which random variables can converge to other random variables.

Definition 11

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

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

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

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

Example 16

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

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

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

By Fatou's lemma, $\mathbb{P}(\limsup A_n) \geqslant \limsup \mathbb{P}(A_n)$, 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 $\mathbb{P}(A_n) = n^{-1}$, then $\mathbb{P}(A_n)$ goes to zero but, by the second Borel-Cantelli lemma $\mathbb{P}(\limsup A_n) = 1$. This example has all the ingredients for the following two implications.

Lemma 8

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

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

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

Proof. Note that analogous to the example, in general

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

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

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 $\sum_n \mathbb{P}(|X_n - X| > \delta) < \infty$, then $\mathbb{P}(|X_n - X| > \delta \text{ i.o}) = 0$ and hence $\limsup |X_n - X| \leqslant \delta$ 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 $X_n \xrightarrow{a.s.} 0$ is a statement about the joint distribution of the entire sequence $(X_1, X_2, \ldots)$ while convergence in probability $X_n \xrightarrow{P} 0$ 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 $X_n \xrightarrow{a.s.} 0$ but $\mathbb{P}(|X_n| \geqslant \delta)$ 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 $X_n \xrightarrow{P} X$, show that $X_{n_k} \xrightarrow{a.s.} X$ for some subsequence.

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

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

4.2. In distribution and in probability. We say that $X_n \xrightarrow{d} X$ 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 9

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

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

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

Proof.

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

$$\mathbb{P}(X_n \leqslant t) \leqslant \mathbb{P}(X \leqslant t + \delta) + \mathbb{P}(X - X_n > \delta)$$

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

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

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

■

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

Exercise 12

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

(2) If $X_n \xrightarrow{P} c$ (a constant) and $Y_n \xrightarrow{d} Y$ (all on the same probability space), then show that $X_n Y_n \xrightarrow{d} cY$.

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

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

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

Example 17

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

As always, the fruitful question is to ask for additional conditions to convergence in probability that would ensure convergence in L^p. Let us stick to $p = 1$. Is there a reason to expect a (weaker) converse? Indeed, suppose $X_n \xrightarrow{P} X$. Write

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

For each t the integrand goes to zero because $X_n \xrightarrow{P} X$. 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 12: 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 $\mathbb{E}[|X_i|\mathbf{1}_{|X_i|>A}] < \varepsilon$ for all $i \in I$.

Two remarks on the definition.

(1) If X is integrable and $\mathbb{P}\{|X| \geqslant M\} = \delta$, then for any set $A \in \mathcal{F}$ with $\mathbb{P}(A) \leqslant \delta$, we have (exercise!) $\mathbb{E}[|X|\mathbf{1}_A] \leqslant \mathbb{E}[|X|\mathbf{1}_{|X|\geqslant M}]$.

Therefore, the uniform integrability of $\{X_i\}_{i\in I}$ may be rephrased as: Given $\varepsilon > 0$, there is a $\delta > 0$ such that $\mathbb{E}[|X_i|\mathbf{1}_A] < \varepsilon$ for all $i \in I$ and for all $A \in \mathcal{F}$ with $\mathbb{P}(A) \leqslant \delta$.

(2) If μ is the distribution of X, then $\mathbb{E}[|X|\mathbf{1}_{|X|>M}] = \int_{[-M,M]^c} |x|d\mu(x) = \nu([-M,M]^c)$ where $d\nu(x) = |x|d\mu(x)$ (observe that ν is also the push-forward of $|X(\omega)|d\mathbb{P}(\omega)$ by the mapping X).

Therefore, the uniform integrability of $\{X_i\}_{i\in I}$ is equivalent to the tightness[3] of the family $\{\nu_i\}_{i\in I}$, where $\nu_i = \mathbb{P}_i \circ X_i^{-1}$ and $d\mathbb{P}_i(\omega) = |X_i(\omega)|d\mathbb{P}(\omega)$.

Next we discuss conditions that ensure uniform integrability. This also gives us many examples.

- If X is integrable, then by DCT, $\mathbb{E}[|X|\mathbf{1}_{|X|\geqslant M}] \to 0$ as $M \to \infty$. Therefore, any finite set of random variables is uniformly integrable. It is when we have an infinite family that the uniformity constraint starts to be felt.
- A family dominated by one integrable random variable (the condition in DCT) is uniformly integrable. Indeed, if $|X_i| \leqslant |Y|$ a.s. for each $i \in I$, then find M such that $\mathbb{E}[|Y|\mathbf{1}_{|Y|>M}] <$

[3] We defined tightness for probability measures but here ν_i are general (but finite) measures. By tightness, we naturally mean that given $\varepsilon > 0$ there is some M such that $\nu_i([-M,M]^c) < \varepsilon$ for all $i \in I$.

ε and observe that $|X_i|\mathbf{1}_{|X_i|>M} \leqslant |Y|\mathbf{1}_{|Y|>M}$. Hence the same M works for the whole family.

- An L^p-bounded family for $p > 1$ is u.i. For, if $\sup_{i\in I} \mathbb{E}[|X_i|^p] \leqslant B$ for some $B < \infty$, then

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

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

- The previous conclusion fails for $p = 1$. For example, let $X_n = n\mathbf{1}_{[0,\frac{1}{n}]}$ on $([0,1], \mathcal{B}, \lambda)$. Then $\mathbb{E}[|X_n|] = 1$, so $\{X_n\}$ is L^1-bounded. However, for any M, if $n > M$, then $\mathbb{E}[|X_n|\mathbf{1}_{|X_n|>M}] = 1$, hence the family is not uniformly integrable.
- But L^1 boundedness is necessary for uniform integrability. To see this find $M > 0$ so that $\mathbb{E}[|X_i|\mathbf{1}_{|X_i|>M}] < 1$ for all i. Then, for any $i \in I$, we get

$$\mathbb{E}[|X_i|] = \mathbb{E}[|X_i|\mathbf{1}_{|X_i|\leqslant M}] + \mathbb{E}[|X_i|\mathbf{1}_{|X_i|>M}] \leqslant M + 1.$$

Domination by an integrable random variable and L^p boundedness (for some $p > 1$) are sufficient for uniform integrability, not necessary.

Exercise 13

Produce examples of uniformly integrable families that are neither dominated by an integrable random variable nor bounded in L^p for some $p > 1$.

The union of two uniformly integrable families is obviously uniformly integrable. The following is less obvious.

Claim 1

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.

Proof. For any $x, y \in \mathbb{R}$ and $M > 0$, observe that $|x+y|\mathbf{1}_{|x+y|>M} \leqslant 2|x|\mathbf{1}_{|x|>M/2} + |y|\mathbf{1}_{|y|>M/2}$. Substitute X_i and Y_j for x and y and take expectations to get

$$\mathbb{E}[|X_i + Y_j|\mathbf{1}_{|X_i+Y_j|>M}] \leqslant 2\mathbb{E}[|X_i|\mathbf{1}_{|X_i|>M}] + 2\mathbb{E}[|Y_j|\mathbf{1}_{|Y_j|>M}].$$

By the uniformly integrability of $\{X_i\}_{i\in I}$ and $\{Y_j\}_{j\in J}$, this can be made arbitrarily small by choosing M sufficiently large. Thus $\{X_i + Y_j : i \in I,\ j \in J\}$ is uniformly integrable. ■

Now we come to the main reason why we started discussing uniform integrability.

Lemma 10

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

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

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

PROOF. If $Y_n = X_n - X$, then $X_n \xrightarrow{L^1} X$ iff $Y_n \xrightarrow{L^1} 0$, while $X_n \xrightarrow{P} X$ iff $Y_n \xrightarrow{P} 0$ and by Claim 10, $\{X_n\}$ is u.i if and only if $\{Y_n\}$ is. Hence we may work with Y_n instead (i.e., we may assume that the limiting r.v. is 0 a.s).

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

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

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

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

Corollary 2

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

To deduce convergence in mean from a.s convergence, we have so far always invoked DCT. The domination condition is sufficient. But as Lemma 10 and corollary 2 show, uniform integrability is the sharp condition, both necessary and sufficient. This is consistent with what we saw earlier, that a dominated family is u.i., while the converse is false. However, it is worth keeping in mind that uniform integrability is difficult to check from the definition. One does it by verifying either the domination condition or boundedness in L^2.

5.1. Relationship to compactness*. Uniform integrability is reminiscent of tightness, and in fact we rephrased it terms of tightness. Recall that tightness is the necessary and sufficient condition for a subset of $\mathcal{P}(\mathbb{R})$ to be precompact. Similarly, uniform integrability is also a criterion for

precompactness of a subset of $L^1(\Omega, \mathcal{F}, \mathbb{P})$, but not in the usual topology, but what is called the *weak topology*.

Weak and Weak-* topologies: Let $(X, \|\cdot\|)$ be a Banach space over $\mathbb{R}$ and let X^* be its dual, i.e., the space of all continuous linear functionals from X to $\mathbb{R}$. It is well-known that X^* is itself a Banach space when endowed with the norm $\|L\|_* = \sup\{|L(x)| : \|x\| \leqslant 1\}$.

Weak topology on X is the smallest topology on X that makes all elements of X^* continuous functions on X. Of course, the weak topology is weaker than the norm topology. For all infinite dimensional X, it is strictly weaker.

The weak-* topology on X^* is the smallest topology for which $L \mapsto L(x)$ is continuous for each $x \in X$.

If it so happens that X is reflexive, i.e., $(X^*)^* = X$, i.e., the only continuous linear functionals on X^* are the evaluations at elements of X (i.e., $L \mapsto L(x)$ for some $x \in X$), then the weak topology on X^* is identical to the weak-* topology on X^*.

It is a celebrated theorem of Riesz that for $1 \leqslant p < \infty$, the dual $(L^p(\mathbb{P}))^*$ is equal to L^q, where $\frac{1}{p} + \frac{1}{q} = 1$. What does that mean? Any $g \in L^q(\mathbb{P})$ defines a linear functional L_g on $L^p(\mathbb{P})$ by $L_g(f) = \int f g d\mathbb{P}$ and every continuous linear functional on $L^p(\mathbb{P})$ is of this form. As $1 \leqslant p < \infty$, we get $\infty \geqslant q > 1$. In particular, all L^p for $1 < p < \infty$ are reflexive. The two odd cases are L^1 and L^∞. While $L^\infty = (L^1)^*$, the dual of L^∞ is generally much larger than L^1.

One of the famous theorems of functional analysis is that of Banach and Alaoglu that asserts that in X^* with its weak topology, precompact sets are precisely bounded sets . But if $X = X^*$, as for L^p with $1 < p < \infty$, this tells us that for weak topology on L^p, the precompact sets are precisely bounded sets. In particular, any L^p bounded sequence (for $1 < p < \infty$) is precompact (in fact has a convergent subsequence).

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 its weak topology are precisely uniformly integrable subsets of $L^1(\mu)$!

CHAPTER 3

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 takes the values 0 and 10^6 with probabilities 0.999 and 0.001 respectively. Then $\mathbb{E}[X] = 1000$ although with a probability 0.999 its value is zero. Thus the typical value of 0 and the average value of 1000 are far from each other.

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). Therefore, inequalities that bound probabilities in terms of moments may be expected to be somewhat useful. In fact, they are extremely useful!

Lemma 11: First moment method or Markov's inequality

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

Proof. For any $t > 0$, clearly $t\mathbf{1}_{X \geqslant t} \leqslant X$. 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

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

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

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

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geqslant t\} = \mathbb{P}\{|X - \mathbb{E}[X]|^2 \geqslant t^2\} \leqslant t^{-2}\mathrm{Var}(X),$$

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

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

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 $\mathbb{E}[Y_n] \to \infty$ but $\mathbb{P}\{Y_n > 0\} = \frac{1}{n^2} \to 0$.

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

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

Thus, if the variance is smaller than $c\mathbb{E}[X]^2$ for some $c < \frac{1}{4}$, we get a non-trivial lower bound of $1 - \frac{c}{4}$ for the probability. More generally, if $\mathrm{Var}(X) < (1 - \delta)^2\mathbb{E}[X]^2$, then we get a lower bound for the probability that $X \geqslant \delta\mathbb{E}[X]$. Observe that in the example given above, $\mathrm{Var}(Y_n) \asymp n^3$ is way larger than $\mathbb{E}[Y_n]^2 \asymp n^2$, 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 $\mathrm{Var}(X) \leqslant C\mathbb{E}[X]^2$ for a large C.

Lemma 12: Second moment method or Paley-Zygmund inequality

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

$$\mathbb{P}\{X > \alpha\mathbb{E}[X]\} \geqslant (1 - \alpha)^2\frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} = \frac{(1 - \alpha)^2}{1 + \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2}}.$$

In particular, $\mathbb{P}\{X > 0\} \geqslant \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}$.

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

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

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

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

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

The remaining conclusions follow easily. ■

Remark 12

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

3. Borel-Cantelli lemmas

If A_n is a sequence of events in a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the event $\limsup A_n$ consists of all $\omega \in \Omega$ 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 $\limsup A_n$. One can write it as $\{A_n \text{ i.o.}\} = \bigcap_{N\geqslant 1}\bigcup_{n\geqslant N} A_n$. Observe that here the inner union decreases as N increases, hence

$$\mathbb{P}(\cup_{n\geqslant N} A_n) \downarrow \mathbb{P}\{A_n \text{ i.o.}\} \text{ as } N \uparrow \infty.$$

However, the probability on the left depends in a complicated way ("inclusion-exclusion") on intersections of the sets A_n, $n \geqslant N$. That is why the following lemma is extraordinarily useful, as it allows (in some cases) to compute the probability of $\{A_n$ i.o.$\}$ knowing only the probabilities of A_n individually, and not of their intersections.

Lemma 13: Borel Cantelli lemmas

Let A_n be events on a common probability space.

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

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

PROOF. (1) For any N, $\mathbb{P}\left(\cup_{n=N}^{\infty} A_n\right) \leqslant \sum_{n=N}^{\infty} \mathbb{P}(A_n)$ which goes to zero as $N \to \infty$, as it is the tail of a convergent series. Hence $\mathbb{P}(\limsup A_n) = 0$.

(2) $\mathbb{P}(\cup_{n=N}^{M} A_n) = 1 - \prod_{n=N}^{M} \mathbb{P}(A_n^c)$ for any $N < M$. By (1), $\mathbb{P}(A_n^c) = 1 - \mathbb{P}(A_n) \leqslant e^{-\mathbb{P}(A_n)}$. Therefore

$$\mathbb{P}(\cup_{n=N}^{M} A_n) \geqslant 1 - \prod_{n=N}^{M} e^{-\mathbb{P}(A_n)} = 1 - \exp\left\{-\sum_{n=N}^{M} \mathbb{P}(A_n)\right\}.$$

As $M \to \infty$, the left side increases to $\mathbb{P}(\cup_{n\geqslant N} A_n)$ while the right side increases to 1 (since $\sum_{n\geqslant N} \mathbb{P}(A_n) = \infty$ for any N). Therefore, $\mathbb{P}\left(\cup_{n=N}^{\infty} A_n\right) = 1$ for all N, implying that $\mathbb{P}(A_n \text{ i.o.}) = 1$. ■

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!

A useful elementary inequality: As in the proof above, we shall often encounter terms like $\prod_i (1-x_i)$ with $0 < x_i < 1$. When $x \approx 0$ is small, $1 - x \approx e^{-x}$, but when taking products of many terms, it is not clear what happens to the closeness. To carry through such operations, the following inequalities are more useful[2].

(1) $$1 - x \leqslant e^{-x} \quad \text{for all } x \in \mathbb{R}, \qquad 1 - x \geqslant e^{-x-x^2} \quad \text{for } |x| < \frac{1}{2}.$$

4. Kolmogorov's zero-one law

As in the Borel-Cantelli lemmas, many events of interest turn out to have probability 0 or 1. In any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the collection of all events having probability equal to 0 or 1 form a sigma algebra. Zero-one laws are theorems that (in special situations) identify specific sub-sigma-algebras of this sigma-algebra. Such σ-algebras (and events within them) are sometimes said to be *trivial* (w.r.t. $\mathbb{P}$). An equivalent statement is that any random variable measurable with respect to a trivial sigma algebra is an almost sure constant.

Definition 13

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 $\mathcal{T} := \bigcap_n \sigma\left(\cup_{k\geqslant n} \mathcal{F}_k\right)$. 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 $\sigma(X_n)$.

[2]For $|x| < 1$, we have the power series expansion $\log(1-x) = -x - x^2/2 - x^3/3 - \ldots$. If $|x| < \frac{1}{2}$, then $\sum_{k=3}^{\infty} |x|^3 \leqslant x^2 \sum_{k=3}^{\infty} 2^{-k} \leqslant \frac{1}{2}x^2$, hence the sum of all terms from the third one onwards is at least $-x^2/2$. This gives $\log(1-x) \geqslant -x - x^2$. The other inequality is even simpler. Consider $e^{-x} - (1-x)$ which is zero at 0 and has positive derivative for $x > 0$ and negative derivative for $x < 0$.

How to think of it? If A is in the tail of $(X_k)_{k\geqslant 1}$, 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 10: Kolmogorov's zero-one law

Let $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $\mathcal{T}_n := \sigma\left(\bigcup_{k>n} \mathcal{F}_k\right)$. 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 $\mathbb{P}(A)^2 = \mathbb{P}(A \cap A) = \mathbb{P}(A)$ which forces $\mathbb{P}(A)$ to be 0 or 1. ■

Corollary 3

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, $k \geqslant 1$ 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 $(X_k)_{k\geqslant 1}$.

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 14

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 13: 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 $n \geqslant 1$. In other words, $(\theta_k\omega)(n) = \omega(n+k)$ where $\omega = (\omega(1), \omega(2), \ldots)$.

Definition 14: 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 $k \geqslant 1$. 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 $k \geqslant 1$. 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 18

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 $|\omega_k| \leqslant 1$ for all $k \geqslant 1$. Then A is an invariant event as well as a tail event while B is an invariant event but not a tail event.

Exercise 15

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

Lemma 14: Ergodicity of i.i.d. measures

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

Proof. Let $\mu = \mathbb{P}^{\otimes\mathbb{N}}$. 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

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

$$|\mu(B)^2 - \mu(A)^2| \leqslant |\mu(B) - \mu(A)||\mu(B) + \mu(A)| \leqslant 2\varepsilon.$$

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 14: 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 $k \geqslant 1$ 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 15: Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent random variables having zero mean. Assume that $|X_k| \leqslant a_k$ 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 $\mathbb{P}\{S \geqslant tA\} \leqslant e^{-\frac{1}{2}t^2}$ 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 $\mathbb{P}\{|S| \geqslant tA\} \leqslant 2e^{-t^2/2}$.

(2) If $|X_k| \leqslant a_k$ are independent but do not necessarily have mean zero, then we can apply Hoeffding's inequality to $Y_k = X_k - \mathbb{E}[X_k]$. Since $|X_k| \leqslant a_k$, we also have $|\mathbb{E}[X_k]| \leqslant a_k$

and hence $|Y_k| \leqslant 2a_k$. This gives a conclusion that is slightly weaker but qualitatively no different: With $S = X_1 + \ldots + X_n$,

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

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

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

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

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

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

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

Consequently, we get that $\prod_{k=1}^{n}\alpha_k \leqslant e^{\theta^2 A^2/2}$. Thus, $\mathbb{P}\{S \geqslant tA\} \leqslant e^{-\theta tA + \frac{1}{2}\theta^2 A^2}$. 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.

But it sometimes falls short of what one needs. As the tail of $N(0, \sigma^2)$ behaves like $e^{-t^2/2\sigma^2}$ for the tails of S_n one might have expected $e^{-t^2/2\sigma^2}$ where $\sigma^2 = \mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$.

As $\mathrm{Var}(X_k) \leqslant a_k^2$, an upper bound of $e^{-t^2/2\sigma^2}$ for $\mathbb{P}\{S_n \geqslant t\}$ would be stronger than the Hoeffding bound. There are inequalities that address this (under more assumptions). Here are two.

Proposition 3: Bernstein's inequality

Assume that X_k are independent random variables with $|X_k| \leqslant B$ a.s. for all k and $\mathbb{E}[X_k] = 0$ and $\mathrm{Var}(X_k) = \sigma_k^2$. Let $\tau_n^2 = \sigma_1^2 + \ldots + \sigma_n^2$. Then for $t > 0$,

$$\mathbb{P}\{S_n \geqslant t\} \leqslant e^{-\frac{t^2}{2(\tau_n^2 + \frac{1}{3}Bt)}}.$$

In particular, if $0 < \varepsilon < 1$, then for $t > 0$,

$$\mathbb{P}\{S_n \geqslant t\} \leqslant \begin{cases} e^{-(1-\varepsilon)\frac{t^2}{2\tau_n^2}} & \text{if } t \leqslant \frac{\varepsilon}{B}\tau_n^2, \\ e^{-\frac{t}{4B}} & \text{if } t > \frac{\varepsilon}{B}\tau_n^2. \end{cases}$$

7. Kolmogorov's maximal inequality

Kolmogorov proved a remarkable inequality about the maximum of running sums of independent random variables. 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 $\mathbb{P}(Y_k > t) = e^{-t}$, whereas $\mathbb{P}(\max_{k\leqslant n} Y_k > t) = 1-(1-e^{-t})^n$ which is much larger (in fact converges to 1 if $n \to \infty$ with t held fixed). 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 tail of the maximum. Kolmogorov found an amazing inequality for which there seems to be no a priori reason!

Lemma 16: Kolmogorov's maximal inequality

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

Observe that the right hand side is the bound that Chebyshev's inequality gives for the probability that $|S_n| \geqslant t$. Here the same quantity is giving an upper bound for the (generally) much larger probability that one of $|S_1|, \ldots, |S_n|$ exceeds t.

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

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

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

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

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

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

In the second line we used the fact that $S_k \mathbf{1}_{\tau=k}$ depends on $X_1, \ldots, X_k$ only, while $S_n - S_k$ depends only on $X_{k+1}, \ldots, X_n$. From (4), this implies that $\mathbb{E}[S_n^2] \geqslant \mathbb{E}[S_\tau^2]$. Plug this into (3) to get $\mathbb{P}(\max_{k \leqslant n} S_k > t) \leqslant t^{-2}\mathbb{E}[S_n^2]$. ■

Remark 15

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

As observed above, the bound for $\mathbb{P}(\max_{k \leqslant n} |S_k| > t)$ given by Kolmogorov's maximal inequality is the same as the bound for $\mathbb{P}(|S_n| > t)$ given by Chebyshev's inequality. We know that the bound for $\mathbb{P}(|S_n| > t)$ can be improved (under assumptions) by applying Markov's inequality to powers or exponential function of S_n. Can we similarly improve the maximal inequality? Turns out we can, by an almost identical proof!

Claim 2: Komogorov's maximal inequality enhanced

Let X_k be independent random variables with zero mean. Assume that $\mathbb{E}[e^{\theta X_k}] < \infty$ for some $\theta > 0$. Then, for any $t > 0$,

$$\mathbb{P}\left\{\max_{0 \leqslant k \leqslant n} S_k \geqslant t\right\} \leqslant e^{-\theta t}\mathbb{E}[e^{\theta S_n}].$$

Observe that the bound is the same as the one we get for $\mathbb{P}\{S_n \geqslant t\}$ by applying Markov's inequality to $e^{\theta S_n}$.

PROOF. Let $S_n^* = \max\{S_0, \ldots, S_n\}$ (without absolute values). Fix $t > 0$ and define $\tau = \inf\{k \leqslant n : S_k \geqslant t\}$ where the infimum of the empty set is defined to be $+\infty$ (that happens precisely when $S_n^* < t$). Then,

$$\mathbb{P}\{S_n^* \geqslant t\} \leqslant e^{-\theta t}\mathbb{E}[e^{\theta S_n^*}\mathbf{1}_{S_n^* \geqslant t}] \;=\; e^{-\theta t}\sum_{k=1}^{n}\mathbb{E}[e^{\theta S_k}\mathbf{1}_{\tau=k}]$$

On the other hand

$$\begin{aligned}\mathbb{E}[e^{\theta S_n}] &\geqslant \sum_{k=1}^{n}\mathbb{E}[e^{\theta S_n}\mathbf{1}_{\tau=k}] \;=\; \sum_{k=1}^{n}\mathbb{E}[e^{\theta(S_n - S_k)}e^{\theta S_k}\mathbf{1}_{\tau=k}]\\ &= \sum_{k=1}^{n}\mathbb{E}[e^{\theta(S_n-S_k)}]\mathbb{E}[e^{\theta S_k}\mathbf{1}_{\tau=k}].\end{aligned}$$

In the last line, we used the fact that S_k and τ_k are measurable with respect to $\sigma\{X_1, \ldots, X_k\}$ while $S_n - S_k$ is measurable with respect to $\sigma\{X_{k+1}, X_{k+2}, \ldots\}$. Hence the independence and factoring of expectations.

By Jensen's inequality $\mathbb{E}[e^{\theta(S_n-S_k)}] \geqslant e^{\theta\mathbb{E}[S_n-S_k]} = 1$. Putting all this together, we have

$$\mathbb{P}\{S_n^* \geqslant t\} \leqslant e^{-\theta t}\mathbb{E}[e^{\theta S_n}].$$

This completes the proof. ■

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 $a \leqslant b$. If we could find random variables X, Y on a common probability space such that $X \leqslant Y$ a.s., and $\mathbb{E}[X] = a$ and $\mathbb{E}[Y] = b$, 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 $\mathbb{P}(A) = a$ and $\mathbb{P}(B) = b$. 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 $\mathbb{P}(A) > 0$ under some probability measure $\mathbb{P}$.

Example 19

Let $X \sim \text{Bin}(100, 3/4)$ and $Y \sim \text{Bin}(100, 1/2)$. Then it must be true that $\mathbb{P}\{X \geqslant 71\} \geqslant \mathbb{P}\{Y \geqslant 71\}$, 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 $X' = \sum_k \mathbf{1}_{U_k \leqslant 3/4}$ and $Y' = \sum_k \mathbf{1}_{U_k \leqslant 1/2}$. Then $X' \geqslant Y'$, hence the event $\{Y' \geqslant 71\}$ is a subset of $\{X' \geqslant 71\}$ showing that $\mathbb{P}\{X' \geqslant 71\} \geqslant \mathbb{P}\{Y' \geqslant 71\}$. 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 $X \geqslant Y$ a.s., then $F_\mu(t) \leqslant F_\nu(t)$ for all $t \in \mathbb{R}$. If the latter relationship holds, we say that ν is stochastically dominated by μ.

Exercise 16

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

$$F_\nu(t) = \mathbb{P}\{Y \leqslant t\} \geqslant \mathbb{P}\{X \leqslant t - r\} - \mathbb{P}\{|X - Y| > r\} \geqslant F_\mu(t - r) - r.$$

and similarly $F_\mu(t) \geqslant F_\nu(t-r) - r$. 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 $d_{TV}(\mu, \nu) = \max_{A \subseteq [n]} |\mu(A) - \nu(A)|$. 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 $\mathbb{P}\{X \neq Y\} = r$. Then $d_{TV}(\mu, \nu) \leqslant r$. Indeed, for any A, we have

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

Getting the inequality with μ and ν reversed, we see that $d_{TV}(\mu, \nu) \leqslant \mathbb{P}\{X \neq Y\}$. It is an easy fact that one can always couple random variables this way.

Exercise 17

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

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 15: 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 $T_c(\mu, \nu) := \inf\{\mathbb{E}[c(X, Y)] : X \sim \mu,\ Y \sim \nu\}$, 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 *Kantorovich metric* or *Wasserstein metric*.

CHAPTER 4

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. 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].

1. Borel-Cantelli lemmas

If X takes values in $\mathbb{R} \cup \{+\infty\}$ and $\mathbb{E}[X] < \infty$ then $X < \infty$ a.s.. That is obvious from the definition of expectation, but one may also see it as a consequence of Markov's inequality, as $\mathbb{P}\{X \geqslant t\} \leqslant t^{-1}\mathbb{E}[X] \to 0$ as $t \to \infty$. Apply this to $X = \sum_{k=1}^{\infty} \mathbf{1}_{A_k}$ which has $\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$ 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 $\mathbb{E}[X] = \sum_{k=n}^{m} \mathbb{P}(A_k)$. Also,

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

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

$$\mathbb{P}(X \geqslant 1) \geqslant \frac{\left(\sum_{k=n}^{m} \mathbb{P}(A_k)\right)^2}{\left(\sum_{k=n}^{m} \mathbb{P}(A_k)\right)^2 + \sum_{k=n}^{m} \mathbb{P}(A_k)}$$
$$= \frac{1}{1 + \left(\sum_{k=n}^{m} \mathbb{P}(A_k)\right)^{-1}}$$

[1] It is not necessary to read all the sections. On first reading one may omit the ones marked with an asterisk. Nothing is majorly wrong with them - some are incompletely written. Besides, many more applications of the basic tools are in the problem set.

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

Note that this proof used independence only to claim that $\mathbb{P}(A_k \cap A_\ell) = \mathbb{P}(A_k)\mathbb{P}(A_\ell)$. 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 $1, 2, \ldots, n$. Every night, before going to bed, you pick one of the books at random to read. The book is replaced in the shelf in the morning. How many days pass before you have picked up each of the books at least once? Let T_n denote the number of days till each book is picked at least once. We show that T_n is concentrated around $n \log n$ in a window of size n. The precise statement is in the theorem below. First let us convert the informal language to mathematics.

Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables with uniform distribution on $[n]$. Then define

$$T_n = \min\{t : \{\xi_1, \ldots, \xi_t\} = [n]\}.$$

Theorem 11: Coupon collector problem

With the above notation, for any sequence of numbers $\theta_n \to +\infty$, we have

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

Proof of Theorem 11. Fix an integer $t \geqslant 1$ and let $X_{t,k}$ be the indicator that the k^{th} book is not picked up on the first t days. Then, $\mathbb{P}(T_n > t) = \mathbb{P}(S_{t,n} \geqslant 1)$ where $S_{t,n} = X_{t,1} + \ldots + X_{t,n}$ is the number of books not yet picked in the first t days. As $\mathbb{E}[X_{t,k}] = (1 - 1/n)^t$ and $\mathbb{E}[X_{t,k}X_{t,\ell}] = (1 - 2/n)^t$ for $k \neq \ell$, we also compute that thefirst two moments of $S_{t,n}$ and use (1) to get

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

and

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

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

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

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

On the other hand, taking $t = n \log n - n\theta_n$ (where we take $\theta_n < \log n$, of course!), we now apply the second moment method. For any $n \geqslant 2$, by using (6) we get $\mathbb{E}[S_{t,n}^2] \leqslant e^{\theta_n} + e^{2\theta_n}$. The first

inequality in (5) gives $\mathbb{E}[S_{t,n}] \geqslant e^{\theta_n - \frac{\log n - \theta_n}{n}}$. Thus,

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

as $n \to \infty$. From (7) and (8), we get the sharp bounds

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

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 $k \geqslant 2$, let τ_k be the number of draws after $k-1$ distinct coupons have been seen till the next new coupon appears. Then, $T_n = \tau_1 + \ldots + \tau_n$.

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

If $\xi \sim \text{Geo}(p)$ (this means $\mathbb{P}(\xi = k) = p(1-p)^{k-1}$ for $k \geqslant 1$), then $\mathbb{E}[\xi] = \frac{1}{p}$ and $\text{Var}(\xi) = \frac{1-p}{p^2}$, by direct calculations. Therefore, remembering that $1 + \frac{1}{2} + \ldots + \frac{1}{n} = \log n + O(1)$, we get

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

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

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

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

Remark 16

One can investigate what happens when the number of days $t = N \log N + cN$ for some constant $c \in \mathbb{R}$. One can follow the first proof and show that the number of unseen books $S_{t,N}$ converges in distribution to $\text{Pois}(e^{-c})$, i.e., for each $k \geqslant 0$, as $n \to \infty$,

$$\mathbb{P}\{S_{t,n} = k\} \to e^{-e^{-c}} \times \frac{e^{-ck}}{k!}.$$

Indeed, if $X_{t,1}, \ldots, X_{t,n}$ were independent, then we would have $S_{t,n} \sim \text{Bin}(n, p_n)$ with $p_n = (1 - \frac{1}{n})^t$. When $t = n \log n + cn$, we see that $np_n \to e^{-c}$, hence the Poisson limit would follow. Although $X_{t,k}$ are not quite independent, their dependence is weak enough that the conclusion does hold. We leave it to the interested reader to do the calculations, just pointing out that it is an instance of Poisson limit law for rare events. Another point to note is that $\mathbb{P}\{S_{t,n} = 0\}$ converges to 0 as $c \to -\infty$ and to 1 as $c \to +\infty$, which is consistent with the statement of Theorem 11.

3. Branching processes

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

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

[2]For those who are not satisfied with the informal description, here is a precise definition: Let $V = \bigcup_{k=1}^{\infty} \mathbb{N}_+^k$ be the collection of all finite tuples of positive integers. For $k \geqslant 2$, 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 $v_j \leqslant \eta((v_1, \ldots, v_{j-1}))$ for $2 \leqslant j \leqslant k$. Also define $Z_{k-1}(\eta) = \#\{(v_1, \ldots, v_k) \in T\}$ for $k \geqslant 2$ 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.

It is hoped that this exorcises you of any wish for more such descriptions and convinces you of the value of the probabilists' language using random variables.

Theorem 12: The fundamental theorem on Branching processes

Let $m = \mathbb{E}[\xi]$ be the mean of the offspring distribution.

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

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

PROOF. In the proof, we compute $\mathbb{E}[Z_n]$ and $\mathrm{Var}(Z_n)$ 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 $\mathbb{E}[Z_n|Z_{n-1}] = mZ_{n-1}$ and if we assume that ξ has variance σ^2 we also get $\mathrm{Var}(Z_n|Z_{n-1}) = Z_{n-1}\sigma^2$. Therefore, $\mathbb{E}[Z_n] = \mathbb{E}[\mathbb{E}[Z_n|Z_{n-1}]] = m\mathbb{E}[Z_{n-1}]$ from which we get $\mathbb{E}[Z_n] = m^n$. Similarly, from the formula $\mathrm{Var}(Z_n) = \mathbb{E}[\mathrm{Var}(Z_n|Z_{n-1})] + \mathrm{Var}(\mathbb{E}[Z_n|Z_{n-1}])$ we can compute that

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

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

(2) If $m = \mathbb{E}[\xi] > 1$, then as before $\mathbb{E}[Z_n] = m^n$ 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

$$\mathbb{P}\{Z_n > 0\} \geqslant \frac{\mathbb{E}[Z_n]^2}{\mathrm{Var}(Z_n) + \mathbb{E}[Z_n]^2} \geqslant \frac{1}{1 + \frac{\sigma^2}{m-1}}$$

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

The assumption of finite variance of ξ can be removed as follows. Since $\mathbb{E}[\xi] = m > 1$, we can find A large so that setting $\eta = \min\{\xi, A\}$, we still have $\mathbb{E}[\eta] > 1$. 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 $\mathbb{E}[Z_n] = 1$ stays constant. Here the strengthened form of Markov's inequality (**??**) 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 $Z_n \geqslant 1$, 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 $\mathbb{E}[M_k] = 1$. 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, $\mathbb{E}[M_k] = 1$. This shows that $\mathbb{E}[Z_n \,\big|\, Z_n \geqslant 1] = n+1$ and consequently, by the strengthening of Markov's inequality given above,

$$\mathbb{P}\{Z_n \geqslant 1\} \leqslant \frac{\mathbb{E}[Z_n]}{\mathbb{E}[Z_n \,\big|\, Z_n \geqslant 1]} = \frac{1}{n+1}$$

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 $n \geqslant 1$ 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 $\mathbb{E}_n$ to denote expectation w.r.t μ_n.

Theorem 13: Hardy-Ramanujan

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

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

PROOF. (**Turan**). Fix n and for any prime p define $X_p : [n] \to \mathbb{R}$ by $X_p(k) = \mathbf{1}_{p|k}$. Then, $\nu(k) = \sum_{p \leqslant k} X_p(k)$. We define $\psi(k) := \sum_{p \leqslant \sqrt[4]{k}} X_p(k)$. Then, $\psi(k) \leqslant \nu(k) \leqslant \psi(k) + 4$ since there can be at most four primes larger than $\sqrt[4]{k}$ that divide k. From this, it is clearly enough to show (9) for ψ in place of ν (why?).

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

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

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

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

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

The second one is obvious, while the first one is not hard, (see exercise 18 below)). Thus, we get $\mathbb{E}_n[\psi] = \log\log n + O(n^{-\frac{3}{4}})$ and $\mathrm{Var}_n[\psi] = \log\log n + O(1)$. Thus, by Chebyshev's inequality,

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

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

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

Exercise 18

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

5. Connectivity of a random graph

The complete graph K_n has vertex set $[n] = \{1, 2, \ldots, n\}$ and edge set $E = \{\{i, j\} : 1 \leqslant i < j \leqslant n\}$. 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 16: Erdös-Rényi random graph

Fix $0 < p < 1$. Let $X_{i,j}$, $1 \leqslant i < j \leqslant n$, 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. What is surprising is that for large n, the change from disconnected to connected happens over a short range of p around the point $\log n/n$.

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 $n \to \infty$,

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

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,

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

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

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

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

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

which goes to 1 as $n \to \infty$ (as $p_n \to 0$ and $\frac{1}{n}e^{np_n} \to 0$). As $\mathcal{G}(n, p_n^-)$ is disconnected, when $Y \geqslant 1$, this completes the proof. ■

Of course, just using the first moment $\mathbb{E}[Y] = n(1-p)^{n-1}$ which goes to zero if $p = p_n^+$, we see by the first moment method that at p_n^+, there are no isolated vertices (with probability tending to 1). But this is not in itself of much use because absence of isolated vertices does not mean that the graph is connected. A more involved argument is needed to show that the expected number of connected components (of any size strictly smaller than n) goes to zero.

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 $|A| \leqslant n/2$. Thus, by the union bound,

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

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

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

$$\begin{aligned}\sum_{k>\varepsilon n} \binom{n}{k}(1-p)^{k(n-k)} &\leqslant n2^n e^{-\varepsilon(1-\varepsilon)n^2 p} \\ &= n2^n e^{-\varepsilon(1-\varepsilon)(1+\delta)n\log n} \text{ when } p = p_n^+.\end{aligned}$$

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

The sum over $k \leqslant \varepsilon n$ is handled by using the bounds $\binom{n}{k} \leqslant n^k$ and $1-p \leqslant e^{-p}$. We get

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

If $\varepsilon > 0$ is chosen small enough that $(1+\delta)(1-\varepsilon)-1 \geqslant \frac{1}{2}\delta$, then the above sum is bounded by a geometric series (with terms $e^{-\frac{1}{2}k\delta\log n}$) whose sum is at most

$$\frac{e^{-\frac{1}{2}\delta\log n}}{1-e^{-\frac{1}{2}\delta\log n}} = \frac{n^{-\delta/2}}{1-n^{-\delta/2}}.$$

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

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 $p \geqslant 3$ 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 $\mathbb{P}\{\xi_n = 1\}$, 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 $\alpha \leqslant 1$ and observe that $|\mathcal{S}_\alpha \cap [N]| = \xi_1 + \ldots + \xi_N$. Therefore,

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

The number of p-th powers below N grows like $N^{1/p}$. Comparing to the above, we see that $p > 3$ 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 $\alpha \geqslant \frac{2}{3}$, 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 (a, b, c) with $c \leqslant N$.

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

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

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 $\mathbb{E}[T] < \infty$ 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.

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

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 $\mathbb{E}[T_N] \sim C \sum_{c=1}^{N} \frac{1}{c} \sim C \log N$. This expectation goes to infinity and hence $\mathbb{E}[T] = \infty$. But to say that T is infinite a.s., we compute the second moment of T_N.

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

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

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

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 $\mathbb{E}[T_N] \sim C \log N$, by Chebyshev inequality we get

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

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

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

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.

Example 20

Let $X_n \sim \mathrm{Ber}(p_n)$. Then the series converges if and only if $X_n = 0$ for all but finitely many n. By the Borel-Cantelli lemmas, the event $\{X_n = 1 \text{ i.o.}\}$ has probability zero or one according as $\sum_n p_n$ converges or diverges. 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 $p_n = \mathbb{E}[X_n]$ in this example, this may give the impression that what matters is the sum of expectations. Not entirely correct. For example, let X_n be independent with $\mathbb{P}\{X_n = 1\} = \mathbb{P}\{X_n = -1\} = p_n/2$ and $\mathbb{P}\{X_n = 0\} = 1 - p_n$. 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 $\mathbb{E}[X_n] = 0$ for all n, what p_n measures is the variance. Khinchine showed that this holds in great generality.

Theorem 16: Khinchine

Let X_n be independent random variables with zero means and finite variances. Assume that $\sum_n \mathrm{Var}(X_n) < \infty$. Then $\sum X_n$ converges, a.s.

Proof. A series converges if and only if it satisfies Cauchy criterion. A sequence $(x_n)_n$ is *not* Cauchy if and only if there is some $\varepsilon > 0$ such that for any $N \geqslant 1$, there exists $k \geqslant 1$ such that $|x_{N+k} - x_N| > \varepsilon$. Translated into symbols, this means that the event E that $(S_n)_n$ is not Cauchy is given by

$$E := \bigcup_{m \geqslant 1} \bigcap_{N \geqslant 1} \bigcup_{k \geqslant 1} \left\{ |S_{N+k} - S_N| > \frac{1}{m} \right\}.$$

Thus, $\mathbb{P}(E) = 0$ if and only if $\bigcap_{N\geqslant 1}\bigcup_{k\geqslant 1}\left\{|S_{N+k} - S_N| \geqslant \frac{1}{m}\right\}$ has zero probability for each m. The intersection is smaller than each of the sets in the intersection, hence it suffices to show that

$$\begin{aligned} 0 &= \lim_{N\to\infty} \mathbb{P}\left\{\bigcup_{k\geqslant 1}\{|S_{N+k} - S_N| > \frac{1}{m}\}\right\} \\ &= \lim_{N\to\infty}\lim_{M\to\infty} \mathbb{P}\{|S_{N+k} - S_N| > \frac{1}{m} \text{ for some } 1 \leqslant k \leqslant M\}. \end{aligned}$$

Kolmogorov's maximal inequality (Lemma 16) gives

$$\mathbb{P}\{|S_{N+k} - S_N| > \frac{1}{m} \text{ for some } 1 \leqslant n \leqslant M\} \leqslant m^2 \sum_{k=N+1}^{N+M} \operatorname{Var}(X_k) \to m^2 \sum_{k=N}^{\infty} \operatorname{Var}(X_k)$$

as $M \to \infty$. The last quantity is the tail of a convergent series and hence goes to zero as $N \to \infty$. That is precisely what we wanted to show, and the proof is complete. ■

What to do if the assumptions are not exactly satisfied? First, suppose that $\sum_n \operatorname{Var}(X_n)$ is finite but $\mathbb{E}[X_n]$ may not be zero. Then, we can write $\sum X_n = \sum(X_n - \mathbb{E}[X_n]) + \sum \mathbb{E}[X_n]$. The first series on the right satisfies the assumptions of Theorem 16 and hence converges a.s. Therefore, if the deterministic series $\sum_n \mathbb{E}[X_n]$ converges, then $\sum_n X_n$ converges a.s. Observe that we are *not* asking for the absolute convergence of the series of expectations.

Next, suppose we drop the assumption that X_n has finite mean or variance. 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 $\mathbb{P}(|X_n| > A)$ is summable. Then set $Y_n = X_n \mathbf{1}_{|X_n|\leqslant A}$. 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 $\sum_n \mathbb{E}[Y_n]$ converges and $\sum_n \operatorname{Var}(Y_n) < \infty$, then it follows from the argument in the previous paragraph and Theorem 16 that $\sum Y_n$ converges a.s. Thus we have proved

Theorem 17: Kolmogorov's three series theorem - part 1

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

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

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

Kolmogorov showed the converse too! That is, if $\sum_n X_n$ converges a.s., then for *any* $A > 0$, the three series (a), (b) and (c) must converge. We skip the proof of this converse implication (but the necessity of convergence of the series (a) is a simple exercise). Although it is of great satisfaction to have found the precise conditions, the useful part is the direction that we showed,

since it allows us to show the almost sure convergence of a random series by checking convergence of three (non-random) numerical series. But we make two remarks on the necessity part.

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 $\mathbb{P}(C) = 1$).

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 $\mathbb{P}(C) = 1$ and $\mathbb{P}(ND) = 1$.

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 $\mathbb{P}(C \cap ND) = 1$, 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. $\mathrm{Exp}(1)$. 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 $\mathbb{P}(|X_n|^{\frac{1}{n}} > t) = e^{-t^n}$ for any $t > 0$. If $t = 1 + \varepsilon$ with $\varepsilon > 0$, this decays very fast and is summable. Hence, $|X_n|^{\frac{1}{n}} \leqslant 1 + \varepsilon$ a.s.. and hence $R \leqslant 1 + \varepsilon$ a.s. Take intersection over rational ε to get $R \leqslant 1$ 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 $|X_n|^{\frac{1}{n}} \geqslant 1 - \varepsilon$. Again, take intersection over rational ε to conclude that $R \geqslant 1$ 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 $\mathbb{E}[L \log_+ L] < \infty$, the generation sizes grow exponentially in the sense that

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

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 $p_k = \mathbb{P}\{L = k\}$ and $m = \sum_k k p_k$ is its mean.

Theorem 18: Growth of supercritical branching process

Assume that $p_0 = 0$ and $m > 1$ and that $\sigma^2 := \mathrm{Var}(L) < \infty$. 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 $L_{n,k}$, $n, k \geqslant 1$ 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 $\mathbb{P}\{W > 0\} > 0$.

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

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

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

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

11. Random walk on a graph

Let X_i be i.i.d. $\mathrm{Ber}_{\pm}(1/2)$ and let $S_n = X_1 + \ldots + X_n$ for $n \geqslant 1$ 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., $A = \{\omega : S_n(\omega) = 0 \text{ infinitely often}\}$. Pólya showed that

$$\mathbb{P}(A) = 1. \tag{10}$$

Observe that A is not a tail event. Indeed, suppose $X_k(\omega) = (-1)^k$ for $k \geqslant 2$. 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. Therefore Kolmogorov's zero-one law is inapplicable. Nevertheless, $\mathbb{P}(A) = 1$ as we shall show now.

Proof of the claim. Let $p_k = \mathbb{P}(S_k = 0)$. It is easy to see that $p_k = 0$ for odd k and $p_k = \binom{k}{k/2}\frac{1}{2^k}$ for even k. By Stirling's formula, one can check that $\frac{c}{\sqrt{k}} \leqslant p_{2k} \leqslant \frac{c'}{\sqrt{k}}$ for some c, c'.

Let $R_n = \sum_{k=0}^{2n} \mathbf{1}_{S_{2k}=0}$ be the number of times the random walk visits the origin in the first $2n$ steps. We see that

$$\mathbb{E}[R_n] = \sum_{k=0}^{n} p_{2k} \geqslant c \sum_{k=1}^{n} \frac{1}{\sqrt{k}} \geqslant c''\sqrt{n}.$$

On the other hand,

$$\begin{aligned}\mathbb{E}[R_n^2] = \sum_{k=0}^{n}\sum_{\ell=0}^{n} \mathbb{P}\{S_{2k} = 0, S_{2\ell} = 0\} \;&=\; \sum_{k=0}^{n} p_{2k} + 2\sum_{k=0}^{n-1}\sum_{\ell=k+1}^{n} p_{2k}p_{2(\ell-k)} \\ &= \mathbb{E}[R_n] + 2\sum_{k=0}^{n-1}\sum_{j=1}^{n-k} p_{2k}p_{2j} \;\leqslant\; \mathbb{E}[R_n] + \mathbb{E}[R_n]^2.\end{aligned}$$

Therefore, $\frac{\mathbb{E}[R_n^2]}{\mathbb{E}[R_n]^2} \to 1$ as $n \to \infty$. By the second moment method, we see that $\mathbb{P}\{R_n \geqslant r\} \to 1$ as $n \to \infty$, for any r. But $\{R_n \geqslant r\} \uparrow \{R_\infty \geqslant r\}$. Hence $\mathbb{P}\{R_\infty \geqslant r\} = 1$ for all r, which is another way of saying that the random walk returns to the origin infinitely many times, almost surely. ■

Remark 17

If you examine the proof, you see that the specifics of the random walk was not used. Although we wrote $\mathbb{E}[R_n] \geqslant c\sqrt{n}$, that was used only to say that $\mathbb{E}[R_n] \to \infty$.

If $(S_n)_{n\geqslant 0}$ is a random walk on any graph (or even a general Markov chain) started at a vertex 0, write $p_k = \mathbb{P}\{S_k\} = 0$ and $R_n = \sum_{k=0}^{n} \mathbf{1}_{S_k} = 0$. Following the same reasoning as above,

$$\mathbb{E}[R_n] = \sum_{k=0}^{n} p_k, \qquad \mathbb{E}[R_n^2] \leqslant \mathbb{E}[R_n] + \mathbb{E}[R_n]^2.$$

Thus, the second moment method shows that if $\mathbb{E}[R_n] \to \infty$, then the random walk eventually returns to the starting point, almost surely.

On the other hand, if $\mathbb{E}[R_n]$ stays bounded, then $\sum_{k=0}^{\infty} p_k < \infty$. Fund N such that $\sum_{k=N}^{\infty} p_k < 1$. This shows that $\mathbb{P}\{S_k = 0 \text{ for some } k \geqslant N\} < 1$ or equivalently, there is a positive probability for the random walk to return only finitely many times.

12. Ramsey numbers

A well-known riddle asks for a proof that among 6 people, there are three who know each other or there are three none of whom knows the other two. To generalize, let us fix $n \geqslant k \geqslant 3$. Let G be a graph with vertex set $[n]$. Let $G^\dagger$ denote the complementary graph: $\{i, j\}$ is an edge in $G^\dagger$ if and only if $\{i, j\}$ is not an edge in G. The question is: Is there necessarily a clique of size k in at least one of G and $G^\dagger$.

The smallest number n for which the answer is "Yes" for every possible graph G on $[n]$, is called the kth Ramsey number, $R(k)$. Beyond a few small values of k, the value of $R(k)$ is not known, even approximately or asymptotically for large k. Erdös used probability to get a lower bound:

$$R(k) \geqslant \frac{k}{2e} 2^{k/2}.$$

The key conclusion is that $R(k)$ grows exponentially fast in k.

Erdös' proof. Pick the graph G uniformly at random from the set of all graphs with vertex set $[n]$. This is done by sampling i.i.d. Ber(1/2) random variables $X_{i,j}$, $1 \leqslant i < j \leqslant n$ and setting the edge set of G to be the set of all $\{i, j\}$ with $X_{i,j} = 1$. The edges of $G^\dagger$ are those $\{i, j\}$ with $i < j$ for which $X_{i,j} = 0$.

Take any subset $S \subseteq [n]$ with $|S| = k$. The chance that S is a clique in G is $2^{-\binom{k}{2}}$. The chance that S is a clique in $G^\dagger$ is the same. Summing these and summing over all S, we see by the union

bound that the chance that some k-element subset of $[n]$ forms a clique in one of G or $G^\dagger$ is at most

$$2\binom{n}{k}2^{-\binom{k}{2}} \leqslant \frac{n^k}{k!\, 2^{\frac{1}{2}k(k-1)-1}}.$$

Therefore, if $n^k < k!2^{\frac{1}{2}k(k-1)-1}$, then the above probability is less than 1. Therefore, there is a positive probability that there is no k-element subset that is a clique of either G or of $G^\dagger$. Hence there must be at least one such graph G. Therefore, $R(k) \geqslant (k!2^{\frac{1}{2}k(k-1)-1})^{\frac{1}{k}} \geqslant \frac{k}{2e}2^{k/2}$ as $k! \geqslant (k/e)^k$. ■

Although the conclusion has nothing to do with probability, the probabilistic method was used. Can you do without it? All you have to do is to construct an explicit graph on c^k vertices (for some $c > 1$) such that no clique of size k exists in the graph or its complement. Apparently, no one has found such an explicit example to date! This is not an uncommon occurrence[3].

13. Percolation

Let $G = (V, E)$ be an infinite connected graph. For $0 \leqslant p \leqslant 1$ and let X_e, $e \in E$, be i.i.d. Ber(p) random variables. These random variables give rise to a random subgraph $G_p = (V, \mathcal{E})$, where $\mathcal{E} = \{e : X_e = 1\}$. Percolation is the study of connectivity properties of this random subgraph. In particular, one is interested in

(1) $\alpha(p)$, the probability that G_p has an infinite connected component.

(2) $\theta_o(p)$, the probability that the vertex o is in an infinite connected component in G_p.

The following seems intuitively obvious, but try proving it before reading the proof!

Claim 3

$\alpha(p)$ and $\theta_o(p)$ are increasing functions of p.

Indeed, if $p_1 < p_2$, we should expect more edges in G_{p_2} than in G_{p_1}, hence it seems that we must have $\alpha(p_2) \geqslant \alpha(p_1)$. But how to prove it? If there were a formula for $\alpha(p)$ or $\theta_o(p)$, we could write the formula and analytically check. But that is not the case. Fortunately, there is a beautiful probabilistic way out!

Proof. Let U_e, $e \in E$, be i.i.d. Unif$[0, 1]$ random variables. For each $p \in [0, 1]$, define the graph H_p as having vertex set V and edge set $\{e \in E : U_e \leqslant p\}$. Then H_p has the same distribution as G_p, for fixed p. Therefore

$$\mathbb{P}\{H_p \text{ has an infinite cluster}\} = \mathbb{P}\{G_p \text{ has an infinite cluster}\} = \alpha(p).$$

[3]The famous book *The probabilistic method* by Alon and Spencer has many such uses of probability.

But H_p are all constructed on the same probability space ("coupled") in such a way that the set of edges of H_{p_1} is a subset of the set of edges of H_{p_2}, if $p_1 < p_2$.

Therefore, if the event "H_{p_1} has an infinite cluster" occurs, so does the event "H_{p_2} has an infinite cluster". Hence $\alpha(p_1) \leqslant \alpha(p_2)$. Similarly, if the event "o is in an infinite cluster of H_{p_1}" occurs, so does the event "o is in an infinite cluster of H_{p_2}". Hence $\theta_0(p_1) \leqslant \theta_o(p_2)$. ■

There is another surprise now.

Claim 4

$\alpha(p)$ is 0 or 1, for any p.

After reading the proof below, think why it does not apply to $\theta_o(p)$.

Proof. Arrange the edges of G in a sequence $e_1, e_2, \ldots$. Then $X_{p_1}, X_{p_2}, \ldots$ are independent random variables, hence any tail event has probability 0 or 1, by Kolmogorov's law. But the event that G_p has an infinite cluster is a tail event, since changing the status of finitely many edges cannot create or destroy and infinite component. Therefore, $\alpha(p) \in \{0, 1\}$. ■

When one combines the two claims, it follows that there must be some $p_c \in [0, 1]$ (that may depend on the graph G) such that $\alpha(p) = 0$ if $p \in [0, p_c)$ and $\alpha(p) = 1$ for $p \in (p_c, 1]$. In many graphs (including $\mathbb{Z}^d$ for any $d \geqslant 2$), one can show[4] that p_c is strictly between 0 and 1. This is very interesting and raised many questions, including what the value of p_c is for specific graphs and what is the value of $\alpha(p_c)$, etc.

But the most interesting take-away for now is that something discontinuous has popped up from a model where no discontinuity was thrown into the definition. There are phenomena in physics called *phase transitions* that are points of discontinuity of some quantity. For example, when ice changes to water or water changes to steam, there is a drastic and sudden change in the intermolecular distances. How can mathematics lead to discontinuous phenomena, unless it is already built into the model definition? Percolation probability $\alpha(p)$ shows us that it is indeed possible! The study of phase transitions is a very active area of research in probability today.

14. Cycles in a random permutation

This section can be said to be an application of coupling, but in a somewhat different sense than we did before. The point we wish to convey is that building a random object using appropriate independent random variables illuminates the random object and makes many computations easier. The random object we choose to illustrate this is a *random permutation*.

[4] While not too difficult, this is a digression that we don't take now. What it entails is showing that $\alpha(p) = 0$ for sufficiently small $p > 0$ and that $\alpha(p) > 0$ for sufficiently large $p < 1$.

Let $\mathcal{S}_n$ denote the set of permutations of $[n]$. Let $\Pi \sim \text{Unif}(\mathcal{S}_n)$ denote a uniformly sampled random permutation. One can ask many questions about Π, we stick to the following one: How many cycles does Π typically have?

It is possible to approach this question in many ways, using recursions, generating functions, etc. You are encouraged to try your hand at it. The method below is exquisitely beautiful and depends on building Π using independent random variables in an ingenious way.

The Chinese restaurant process of Dubins and Pitman: Imagine a restaurant with infinitely many circular tables numbered $1, 2, 3 \ldots$. Initially, all tables are empty. People $P_1, P_2, \ldots$ enter the restaurant one after another and sit as follows.

- P_1 sits at table $V_1 = 1$.
- For $k \geqslant 2$, when P_k arrives, she picks a number $V_k \sim \text{Unif}\{1, 2, \ldots, k\}$ (independently of $V_1, \ldots, V_{k-1}$). If $V_k = j < k$, then P_k sits to the immediate left of the person P_j (if P_i is already sitting to the left of P_j, then P_k sits between P_i and P_j). But if $V_k = k$, then P_k sits in the first vacant table available.
- Interpret each table as a cycle going clockwise. For example, if a table has P_4, P_7, P_{11}, P_{12} seated clocwise in that order, we interpret it as the cycle $(4, 7, 11, 12)$. Taking the different tables as a product of disjoint cycles, we get a permutation Π.

For example, if $(V_1, \ldots, V_9) = (1, 1, 2, 1, 5, 4, 7, 7, 6)$, then $\Pi = (1, 9, 6, 4, 2, 3)(5)(7, 8)$.

Claim 5

After n people arrive, the permutation Π_n built from the Chinese restaurant process has uniform distribution on $\mathcal{S}_n$.

PROOF. $V = (V_1, \ldots, V_n)$ has uniform distribution on the set $\hat{\mathcal{S}}_n := [1] \times [2] \times \ldots \times [n]$. Further, the CRP sets up a bijection $V \leftrightarrow \Pi$ between $\hat{\mathcal{S}}_n$ and $\mathcal{S}_n$. Hence $\Pi_n \sim \text{Unif}(\mathcal{S}_n)$. ■

How is this useful? Let us address the above question on the number of cycles $\mathcal{C}_n$ of Π_n. Observe that

$$\mathcal{C}_n = \sum_{k=1}^{n} \mathbf{1}_{V_k = k}$$

since a new table is started by P_k if and only if $V_k = k$ (and $\mathcal{C}_n$ is just the number of tables occupied after n people have arrived). From this we immediately arrive at

$$\mathbb{E}[\mathcal{C}_n] = \sum_{k=1}^{n} \mathbb{P}\{V_k = 1\} = \sum_{k=1}^{n} \frac{1}{k} = \log n + O(1).$$

Thus, a random permutation has about $\log n$ number of cycles in expectation. Observe that the independence of V_ks was not used, just the linearity of expectation. Using independence, one can show that typically (not just on average) the number of cycles is close to $\log n$.

Exercise 19

Calculate $\mathrm{Var}(\mathcal{C}_n)$ and show that $\mathbb{P}\{(1-\delta)\log n \leqslant \mathcal{C}_n \leqslant (1+\delta)\log n\} \to 1$ as $n \to \infty$.

Remark 18

The Chinese restaurant way is particularly suited to study the cycle structure of a random permutation. For example one can use it to study the distribution of the cycle sizes. It is not well-suited if one is interested in some other feature such as the number of descents or the length of the longest increasing subsequence (both have been studied by combinatorists and probabilists). One can try to find other ways of constructing or representing Π to study such features, but there is no guarantee that so illuminating a representation exists!

CHAPTER 5

Laws of large numbers

The time it takes to drive from Bangalore to Dharwad is fairly stable, although the details of traffic is different every day. How many passengers will board the metro in a day is also fairly stable, which allows the planning of frequency of trains. It is possible to predict with great certainty how long 2 grams of a radio active material is reduced to 1 gram of it, although it is impossible to predict when a particular atom will decay and the best model is that it is a random variable with an exponential distribution.

In the broadest sense, a law of large numbers is the phenomenon of deterministic behaviour emerging from the combination of many random ingredients. In this chapter we shall see a few theorems that try to capture this in simple, yet important, situations.

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 19: Kolmogorov's weak law of large numbers

Let $X_1, X_2 \ldots$ be i.i.d random variables. If $\mathbb{E}[|X_1|] < \infty$, then for any $\delta > 0$,

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

Let us introduce some terminology. If Y_n, Y are random variables on a probability space and $\mathbb{P}\{|Y_n - Y| \geqslant \delta\} \to 0$ as $n \to \infty$ for every $\delta > 0$, then we say that Y_n converges to Y *in probability* and write $Y_n \xrightarrow{P} Y$. In this language, the conclusion of the weak law of large numbers is that $\frac{1}{n} S_n \xrightarrow{P} \mathbb{E}[X_1]$ (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 $\mathbb{E}[X_1] = 0$ (or else replace X_i by $X_i - \mathbb{E}[X_1]$). By Chebyshev's inequality, $\mathbb{P}(|n^{-1} S_n| > \delta) \leqslant$

$n^{-2}\delta^{-2}\mathrm{Var}(S_n)$. By the independence of X_is, we see that $\mathrm{Var}(S_n) = n\sigma^2$. Thus, $\mathbb{P}(|\frac{S_n}{n}| > \delta) \leqslant \frac{\sigma^2}{n\delta^2}$ 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 $Y_k := X_k \mathbf{1}_{|X_k| \leqslant A_n}$ 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 $\mu_n := \mathbb{E}[Y_1] = -\mathbb{E}[Z_1]$ 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 $n \geqslant n_0$.

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

$$\mathbb{P}\left\{ \Big| \frac{S_n^Y}{n} - \mu_n \Big| > \delta \right\} \leqslant \frac{\mathrm{Var}(Y_1)}{n\delta^2} \leqslant \frac{A_n \mathbb{E}[|X_1|]}{n\delta^2}. \tag{11}$$

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

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

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

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

Now, we take $A_n = \alpha n$ with $\alpha := \delta^3 \mathbb{E}[|X_1|]^{-1}$. The first term clearly becomes less than δ. The second term is bounded by $\alpha^{-1}\mathbb{E}[|X_1| \, \mathbf{1}_{|X_1| > \alpha n}]$, which goes to zero as $n \to \infty$ (for any fixed choise of $\alpha > 0$). Thus, we see that

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

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 $\mathrm{Var}(S_n^Y) = n\mathrm{Var}(Y_1)$ 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 $\mathbb{E}[|X|^p] < \infty$ 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 $\mathbb{E}[|X_1|] = \infty$), 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 $a_n = \mathbb{E}[X]$.

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

Note that the Cauchy distribution violates this condition.

Exercise 20

Find a distribution which satisfies the condition $t\mathbb{P}\{|X| \geqslant t\} \to 0$ 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 proof of Weierstrass approximation theorem. Recall the Weierstrass' approximation theorem.

Theorem 20: 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 $n \geqslant 1$, 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 $Q_{f,n}(p) = \mathbb{E}[f(n^{-1}S_n)]$, where S_n has $\text{Bin}(n,p)$ distribution. Law of large numbers enters, because Binomial may be thought of as a sum of i.i.d Bernoullis.

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

$$\begin{aligned}
\left| \mathbb{E}_p \left[f\left(\frac{S_n}{n}\right)\right] - f(p) \right| &\leqslant \mathbb{E}_p \left[\left| f\left(\frac{S_n}{n}\right) - f(p) \right| \right] \\
&= \mathbb{E}_p \left[\left| f\left(\frac{S_n}{n}\right) - f(p) \right| \mathbf{1}_{|\frac{S_n}{n}-p| \leqslant \delta} \right] + \mathbb{E}_p \left[\left| f\left(\frac{S_n}{n}\right) - f(p) \right| \mathbf{1}_{|\frac{S_n}{n}-p| > \delta} \right] \\
&\leqslant \omega_f(\delta) + 2\|f\| \mathbb{P}_p \left\{ \left| \frac{S_n}{n} - p \right| > \delta \right\} \qquad (12)
\end{aligned}$$

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

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

Plugging this into (12) and recalling that $Q_{f,n}(p) = \mathbb{E}_p \left[f\left(\frac{S_n}{n}\right)\right]$, we get

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

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 19

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 $\mathrm{Bin}(n,p)$ 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 $\mathbb{E}[Y_1] = \int_a^b f(x)dx$. Therefore, by WLLN,

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

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

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

Exercise 21

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 $\sqrt{\mathrm{Var}(\bar{X}_n)} = \sqrt{p(1-p)}n^{-\frac{1}{2}}$. 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 $n^{-1}S_n \xrightarrow{P} \mathbb{E}[X_1]$. The strong law strengthens it to almost sure convergence.

Theorem 21: Kolmogorov's strong law of large numbers

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

The proof of this theorem is somewhat complicated. First of all, we should ask if WLLN implies SLLN? From Lemma 8 we see that this can be done if $\mathbb{P}\left(|n^{-1}S_n - \mathbb{E}[X_1]| > \delta\right)$ is summable, for every $\delta > 0$. Even assuming finite variance $\mathrm{Var}(X_1) = \sigma^2$, Chebyshev's inequality only gives a

bound of $\sigma^2\delta^{-2}n^{-1}$ for this probability and this is not summable. Since this is at the borderline of summability, if we assume that pth moment exists for some $p > 2$, we may expect to carry out this proof. Suppose we assume that $\alpha_4 := \mathbb{E}[X_1^4] < \infty$ (of course 4 is not the smallest number bigger than 2, but how do we compute $\mathbb{E}[|S_n|^p]$ in terms of moments of X_1 unless p is an even integer?). Then, we may compute that (assume $\mathbb{E}[X_1] = 0$ without loss of generality)

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

Thus $\mathbb{P}\left(|n^{-1}S_n| > \delta\right) \leqslant n^{-4}\delta^{-4}\mathbb{E}[S_n^4] = O(n^{-2})$ which is summable, and by Lemma 8 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].

Theorem 22: SLLN under second moment assumption

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

Proof. Assume $\mathbb{E}[X_1] = 0$ without loss of generality and let $\sigma^2 = \mathrm{Var}(X_1)$. By Chebyshev's inequality, $\mathbb{P}\{|\frac{1}{n}S_n| \geqslant t\} \leqslant \frac{\sigma^2}{nt^2}$ since $\mathrm{Var}(S_n) = n\sigma^2$. 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 $|\frac{1}{n_k}S_{n_k}| \leqslant \delta$ for all but finitely many k, almost surely. If this even be denoted E_δ, then $\mathbb{P}(E_\delta) = 1$, 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

$$\frac{1}{n_k}U_{n_k} \to \mathbb{E}[X_1^+] \quad \text{and} \quad \frac{1}{n_k}V_{n_k} \to \mathbb{E}[X_1^-], \ \text{a.s.} \tag{13}$$

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

Now for any n, let k be such that $n_k \leqslant n < n_{k+1}$. Clearly $U_{n_k} \leqslant U_n < U_{n_{k+1}}$ and $V_{n_k} \leqslant V_n < V_{n_{k+1}}$, 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,

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

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

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

[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.

we see that on the event in (13), we also have $\frac{1}{n}U_n \to \mathbb{E}[X_1^+]$ and $\frac{1}{n}V_n \to \mathbb{E}[X_1^-]$. Putting these together with the almost sure assertion of (13), and recalling that $S_n = U_n - V_n$, we conclude that $\frac{1}{n}S_n \overset{a.s.}{\to} \mathbb{E}[X_1^+] - \mathbb{E}[X_1^-] = \mathbb{E}[X_1]$. ■

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, 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 21. **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 $X_n \geqslant 0$ and $\mu = \mathbb{E}[X_1] < \infty$ (Caution: Don't also assume zero mean in addition to non-negativity!). One consequence of non-negativity is that

$$\frac{S_{N_1}}{N_2} \leqslant \frac{S_n}{n} \leqslant \frac{S_{N_2}}{N_1} \text{ if } N_1 \leqslant n \leqslant N_2. \tag{14}$$

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 6

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

Proof of the claim Fix j and for $1 \leqslant k \leqslant n_j$ write $X_k = Y_k + Z_k$ where $Y_k = X_k \mathbf{1}_{X_k \leqslant n_j}$ 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 $j \geqslant J_\delta$, we have $\mathbb{E}[Z_1] \leqslant \delta$. 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

[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.

$S_{n_j} = S^Y_{n_j} + S^Z_{n_j}$ and $\mathbb{E}[X_1] = \mathbb{E}[Y_1] + \mathbb{E}[Z_1]$, we get

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{S_{n_j}}{n_j} - \mathbb{E}[X_1]\right| > 2\delta\right\} &\leqslant \mathbb{P}\left\{\left|\frac{S^Y_{n_j}}{n_j} - \mathbb{E}[Y_1]\right| + \left|\frac{S^Z_{n_j}}{n_j} - \mathbb{E}[Z_1]\right| > 2\delta\right\} \\ &\leqslant \mathbb{P}\left\{\left|\frac{S^Y_{n_j}}{n_j} - \mathbb{E}[Y_1]\right| > \delta\right\} + \mathbb{P}\left\{\left|\frac{S^Z_{n_j}}{n_j} - \mathbb{E}[Z_1]\right| > \delta\right\} \\ &\leqslant \mathbb{P}\left\{\left|\frac{S^Y_{n_j}}{n_j} - \mathbb{E}[Y_1]\right| > \delta\right\} + \mathbb{P}\left\{\frac{S^Z_{n_j}}{n_j} \neq 0\right\}. \qquad (15)\end{aligned}$$

We shall show that both terms in (15) are summable over j. The first term can be bounded by Chebyshev's inequality

$$\mathbb{P}\left\{\left|\frac{S^Y_{n_j}}{n_j} - \mathbb{E}[Y_1]\right| > \delta\right\} \leqslant \frac{1}{\delta^2 n_j}\mathbb{E}[Y_1^2] = \frac{1}{\delta^2 n_j}\mathbb{E}[X_1^2 \mathbf{1}_{X_1 \leqslant n_j}]. \qquad (16)$$

while the second term is bounded by the union bound

$$\mathbb{P}\left\{\frac{S^Z_{n_j}}{n_j} \neq 0\right\} \leqslant n_j \mathbb{P}(X_1 > n_j). \qquad (17)$$

The right hand sides of (16) and (17) are both summable. To see this, observe that for any positive x, there is a unique k such that $n_k < x \leqslant n_{k+1}$, and then

$$\sum_{j=1}^{\infty} \frac{1}{n_j} x^2 \mathbf{1}_{x \leqslant n_j} \leqslant x^2 \sum_{j=k+1}^{\infty} \frac{1}{\lambda^j} = x^2 \frac{C_\lambda}{\lambda^{k+1}} \leqslant C_\lambda x,$$

$$\sum_{j=1}^{\infty} n_j \mathbf{1}_{x > n_j} \leqslant \sum_{j=1}^{k} \lambda^j \leqslant \lambda^k \sum_{j \geqslant 0} \frac{1}{\lambda^j} \leqslant C_\lambda x.$$

Here, we may take $C_\lambda = \sum_{j \geqslant 0} \lambda^{-j} = \frac{\lambda}{\lambda - 1}$, 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

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

As $\mathbb{E}[X_1] < \infty$, the probabilities on the left hand side of (16) and (17) are summable in j, and hence it also follows that $\mathbb{P}\left\{\left|\frac{S_{n_j}}{n_j} - \mathbb{E}[X_1]\right| > 2\delta\right\}$ is summable. This happens for every $\delta > 0$ and hence Lemma 8 implies that $\frac{S_{n_j}}{n_j} \overset{a.s.}{\to} \mathbb{E}[X_1]$ a.s. This proves the claim.

Step 3: Fix $\lambda > 1$. Then, for any n, find k such that $\lambda^k < n \leqslant \lambda^{k+1}$, and then, from (14) we get

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

Take intersection of the above event over all $\lambda = 1 + \frac{1}{m}$, $m \geqslant 1$ to get $\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}[X_1]$ 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 17: A maximal inequality

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

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

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.

Proof of SLLN assuming Lemma 17. Fix $A > 0$ and define $Y_n = X_n \mathbf{1}_{|X_n| \leqslant A}$ 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) $\frac{1}{n}S_n^Y \overset{a.s.}{\to} \mathbb{E}[X_1 \mathbf{1}_{|X_1| \leqslant A}]$ by the SLLN for bounded random variables

(2) For any $\varepsilon > 0$, by Lemma 17,

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

Putting these together, we have

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

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

It remains to prove the maximal inequality.

Proof of Lemma 17. Define

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

[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.

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

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

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

If X_k are i.i.d. with finite mean, we have $M_n \stackrel{d}{=} M_n'$ and hence have the same expectation (why does $\mathbb{E}[M_n]$ exist?). Hence $\mathbb{E}[X_1 \mathbf{1}_{M_n > 0}] \geqslant 0$.

Fix $t > 0$ and apply this to the variables $X_i - t$. The corresponding quantities $M_{n,t}, M_{n,t}'$ satisfy $M_{n,t} \leqslant (M_n - t)_+$ and $M_{n,t}' \leqslant (M_n' - t)_+$. Therefore,

$$\mathbb{E}[(X_1 - t)\mathbf{1}_{M_n > t}] \geqslant \mathbb{E}[(X_1 - t)\mathbf{1}_{M_{n,t} > 0}] \geqslant 0.$$

Therefore, $\mathbb{E}[X_1 \mathbf{1}_{M_n > t}] \geqslant t\mathbb{P}\{M_n > t\}$, and the left side is clearly bounded by $\mathbb{E}[|X_1|]$. This gives the inequality

$$\mathbb{P}\{M_n > t\} \leqslant \frac{1}{t}\mathbb{E}[|X_1|].$$

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. In particular, there are two directions, both of which could be made precise in different ways. Overall, the interest is in getting stronger conclusions by making stronger assumptions as necessary.

Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero means.

(1) For what α does $\frac{S_n}{n^\alpha} \stackrel{a.s.}{\to} 0$ or $\frac{S_n}{n^\alpha} \stackrel{P}{\to} 0$? Clearly if $\alpha \geqslant 1$, these are true, but the interest is in $0 < \alpha < 1$. This leads to what is known as the *law of iterated logarithm*.

(2) We know that $\mathbb{P}\{|\frac{S_n}{n}| \geqslant t\} \to 0$ for any $t > 0$? Can we say more? One may ask for upper bounds valid for all n and t or one may ask for the rate at which the probability goes to zero as $n \to \infty$. The first (exact bounds) kind are called *concentration inequalities* and the second kind (asymptotic rate) are called *large deviations*.

5.1. The law of iterated logarithm. For simplicity, assume that X_i are bounded random variables with mean zero. Say $|X_k| \leqslant B$ a.s. Then Hoeffding's inequality says that

$$\mathbb{P}\{|S_n| \geqslant u\} \leqslant 2e^{-\frac{u^2}{2B^2 n}}$$

for any $n \geqslant 1$ and any $u > 0$. Clearly this goes to zero if $u = u_n$ and $\frac{u_n}{\sqrt{n}} \to \infty$. Thus, for any such u_n, we get $\frac{S_n}{u_n} \xrightarrow{P} 0$. In particular, this holds for $u_n = n^\alpha$ with $\alpha > \frac{1}{2}$, but one can also take $u_n = \sqrt{n}h(n)$, where $h(n) \to \infty$ arbitrarily slowly. When we prove CLT, it will be clear that the probability does not go to zero if $h(n)$ stays bounded. So we have a complete answer for the convergence in probability question.

The story is a little more interesting when it comes to almost sure convergence. Now we should ask for summability of the deviation probabilities. If $u = u_n$ where $u_n = B\sqrt{2(1+\varepsilon)n\log n}$, then

$$e^{-\frac{u^2}{2B^2 n}} = e^{-(1+\varepsilon)\log n} = \frac{1}{n^{1+\varepsilon}}$$

which is summable. Therefore, $\limsup \frac{S_n}{\sqrt{n\log n}} \leqslant B\sqrt{2}$ a.s. In particular, if we take $u_n = h(n)\sqrt{n\log n}$ where $h(n) \to \infty$ as $n \to \infty$, then $\frac{S_n}{\sqrt{n\log n}h(n)} \xrightarrow{a.s.} 0$. However, this is not the optimal answer. The precise answer is given by the following theorem, first proved by Khinchine for Bernoulli distribution, and extended by Komogorov and then Hartman and Wintner to more general distributions.

Theorem 23: Law of iterated logarithm

Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero means and unit variances. Then

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

This is the sharp answer, as dividing by anything growing faster than $\sqrt{n\log\log n}$ will obviously give zero in the limit.

This cannot be proved by a naive application of Borel-Cantelli lemmas. We know that if A_n are independent events, then $\mathbb{P}\{A_n \text{ i.o.}\}$ is 0 or 1 according as $\sum_n \mathbb{P}(A_n)$ is finite or infinite. However, for non-independent events, only one side of the implication is correct. Consider the following example.

Example 21: Borel-Cantelli after blocking

Let A_n be independent events in a probability space and let $B_1 = A_1$, $B_2 = B_3 = A_2$, $B_4 = B_5 = B_6 = A_3$ and so on (n many B_is are equal to A_n). To show that only finitely many B_ns occur a.s., if we apply Borel-Cantelli lemma naively, we get the sufficient condition $\sum n\mathbb{P}(A_n) < \infty$. This is clearly foolish, as the event $\{B_n \text{ i.o.}\}$ is the same as $\{A_n \text{ i.o.}\}$, and the latter has zero probability whenever $\sum \mathbb{P}(A_n) < \infty$, a much weaker condition!

Although the situation in the example may look artificial, it is the general nature of things. Often we have a sequence of events $B_1, B_2, \ldots$ where B_n and B_{n+1} are very nearly the same event, but B_n and B_m are nearly independent if $|n-m|$ is large. This is clearly so for $B_n = \{S_n \geqslant g(n)\}$ for some smooth, polynomially growing function. Khinchine's idea is that there is some way to

make them into blocks $C_k = \cup_{n_k \leqslant n < n_{k+1}} B_n$, so that C_k are nearly independent, C_k is almost the same as B_{n_k}. This way, the event $\{B_n \text{ i.o}\}$ is nearly the same as $\{C_k \text{ i.o}\}$ and that has (nearly) zero or one probability according as $\sum_k \mathbb{P}(B_{n_k})$ converges or diverges. There are many details glossed over here, but the key point is that of applying Borel-Cantelli lemma after appropriate blocking. We give proof of the upper bound in LIL in Section 2.

5.2. Large deviations and concentration inequalities. Now we come to the second question of getting bounds for deviation probabilities $\mathbb{P}\{|\frac{S_n}{n}| \geqslant t\}$. We already know that if the X_k are bounded by B and have zero means, then Hoeffding's inequality gives

$$\mathbb{P}\{|S_n| \geqslant tn\} \leqslant 2e^{-\frac{t^2}{2B^2}}.$$

Such inequalities are called concentration inequalities. Recall that the starting point of the proof of Hoeffding's inequality was the application of Markov's inequality to $e^{\theta S_n}$:

$$\begin{aligned}\mathbb{P}\{S_n \geqslant tn\} &\leqslant e^{-\theta nt}\mathbb{E}[e^{\theta S_n}] = e^{-\theta nt}\mathbb{E}[e^{\theta X_1}]^n \\ &= e^{-cn}\end{aligned}$$

where $c = \theta t + \log \mathbb{E}[e^{\theta X_1}]$. Similar, but weaker polynomially decaying bounds can be derived assuming only a few moments for the X_ks. All these are concentration inequalities.

If we are interested in the asymptotics of the deviation probabilities as $n \to \infty$, more precise things can be said.

Assume that $X_1, X_2, \ldots$ are i.i.d. random variables such that $\psi(\theta) = \mathbb{E}[e^{\theta X_1}] < \infty$ for all $\theta \in \mathbb{R}$ (satisfied by bounded random variables, for example). Then we take the bound obtained above

$$\mathbb{P}\{S_n \geqslant tn\} \leqslant e^{-n[\theta t - \log \psi(\theta)]}.$$

For fixed $t > 0$, the best bound is got by optimizing over θ. Let $I(t) = \sup_\theta(\theta t - \log \psi(\theta))$. The supremum can be shown to be finite by some convexity observations, but just assume it for now. Then we get

$$\mathbb{P}\{S_n \geqslant tn\} \leqslant e^{-nI(t)}.$$

This is still valid for all n and $t > 0$. What is remarkable is that the bound becomes tight as $n \to \infty$, at least on the logarithmic scale.

Theorem 24: Cramer's theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variable with $\psi(\theta) < \infty$ for all $\theta \in \mathbb{R}$. Then

$$\frac{1}{n}\log \mathbb{P}\{S_n/n \geqslant t\} = -I(t) \quad \text{for any } t > 0.$$

Because we take logarithms and divide by n, this statement is *not* saying that we can reverse the upper bound and write $\mathbb{P}\{S_n \geqslant tn\} \leqslant ce^{-nI(t)}$ for some constant. In fact, it could be well be

that $\mathbb{P}\{S_n \geqslant tn\} \leqslant \frac{1}{n^{10}} e^{-nI(t)}$. But on the log-scale, asymptotically, we get a sharp estimate for the probability of deviation.

We have already proved the upper bound. We shall not prove the lower bound here. We work it out for the special case of Bernoulli random variables, where precise computations are possible.

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

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

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 $n \geqslant N$ and $K \leqslant k \leqslant N - K$, we have

$$\frac{(1-\varepsilon)}{2\sqrt{n}} e^{-nI(k/n)} \leqslant p_n(k) \leqslant (1+\varepsilon) e^{-nI(k/n)}. \tag{18}$$

where we used the fact that $k(n-k)$ is largest when $k = n/2$ and smallest when $k = 1$ (we anyway have $k \geqslant K$) 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 $2(x-\frac{1}{2})^2 \leqslant I(x) \leqslant 3(x-\frac{1}{2})^2$ for $0 \leqslant x \leqslant 1$.

Large deviations: If $x > \frac{1}{2}$, then take $\varepsilon = 1/2$ (or any fixed number in $(0,1)$) and use (18) to get

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

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

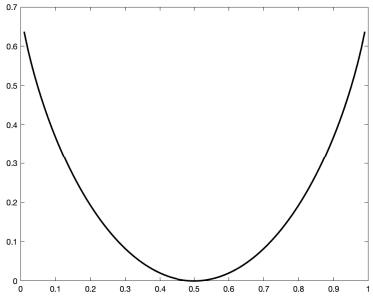

FIGURE 1. Graph of the function $x \mapsto I(x)$

compared to the exponential term and we can simply write $\mathbb{P}\{S_n > nx\} \approx e^{-nI(x)}$ in the sense that

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

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

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

$$\mathbb{P}\{S_n > nx\} \leqslant ne^{-nI(x)} \;\leqslant\; ne^{-2n(x-\frac{1}{2})^2}$$

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

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

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*

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

6. Empirical distribution converges to true distribution*

Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables with distribution μ. The *empirical distribution* based on the first n samples is defined as the random probability measure

$$L_n = \frac{1}{n}(\delta_{X_1} + \ldots + \delta_{X_n}).$$

This is the probability distribution whose CDF F_{L_n} has jumps of size $\frac{1}{n}$ at each of the sample points X_k, $k \leqslant n$, counted with multiplicity (meaning that if a particular value occurs p times, then the jump at that location if $\frac{p}{n}$). Thus for a fixed $x \in \mathbb{R}$, we see that

$$F_{L_n}(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X_k \leqslant x} \overset{a.s.}{\to} F_\mu(x)$$

by SLLN, since $\mathbf{1}_{X_k \leqslant x}$ are i.i.d. with $\mathrm{Ber}(F_\mu(x))$ distribution. If D is any countable dense set such as rationals, we can take the intersection of the almost sure event above and say that

$$F_{L_n}(x) \to F_\mu(x) \text{ for all } x \in D, \qquad \text{a.s.} \tag{19}$$

Pay attention to the placement of the "a.s."; for example, it is not evident that we can write the above statement with "for all $x \in \mathbb{R}$" as it involves taking uncountable intersection. Nevertheless, the above statement is sufficient to say that $L_n \overset{d}{\to} \mu$. This is because (it was an exercise in the problem set), convergence of the CDF at a countable dense set of points implies convergence in distribution. Applying this to each ω in the good set in (19), we get that[4] $L_n \overset{d}{\to} \mu$ a.s.

But in fact, the convergence holds in the stronger Kolmogorov-Smirnov metric! In particular, that implies that in (19)one can write "for all $x \in \mathbb{R}$".

Theorem 25: Glivenko-Cantelli

Let $X_1, X_2, \ldots$ be i.i.d. random variables with distribution μ. Then

$$\|F_{L_n} - F_\mu\|_{\sup} \overset{a.s.}{\to} 0.$$

Proof. First let us do it assuming that $\mu = \lambda$ is the uniform distribution on $[0,1]$. Fix integer $M \geqslant 1$ and let $N(\omega)$ be such that $|F_{L_n(\omega)}(k/M) - (k/M)| \leqslant \varepsilon$ for all $k \in \{0, 1, \ldots, M\}$ for all $n \geqslant N(\omega)$. SLLN shows that $N(\omega) < \infty$ a.s., for any $M < \infty$. But then, for any $x \in [0,1]$, we can find k such that $\frac{k}{M} \leqslant x \leqslant \frac{k+1}{M}$ and then

$$\begin{aligned} F_{L_n(\omega)}(x) - F_\lambda(x) &\leqslant F_{L_n(\omega)}(\frac{k+1}{M}) - F_\lambda(\frac{k}{M}) \\ &= F_{L_n(\omega)}(\frac{k+1}{M}) - \frac{k+1}{M} + \frac{1}{M} \leqslant \varepsilon + \frac{1}{M}. \end{aligned}$$

Similarly

$$\begin{aligned} F_\lambda(x) - F_{L_n(\omega)}(x) &\leqslant \frac{k+1}{M} - F_{L_n(\omega)}(\frac{k}{M}) \\ &= \frac{k}{M} - F_{L_n(\omega)}(\frac{k}{M}) + \frac{1}{M} \leqslant \varepsilon + \frac{1}{M}. \end{aligned}$$

[4] Be careful in reading this statement. What it means is that for almost every ω, the sequence of measures $L_n(\omega)$ converge to μ in the Lévy metric.

Together, this shows that $\|F_{L_n(\omega)} - F_\lambda\|_{\sup} \leqslant \varepsilon + \frac{1}{M}$ for all $n \geqslant N(\omega)$. As $\varepsilon > 0$ and $M < \infty$ are arbitrary, we see that $\|F_{L_n} - F_\lambda\|_{\sup} \to 0$ a.s.

Now if μ is a general distribution, without loss of generality, we assume that $X_k = G_\mu(U_k)$, where U_k are i.i.d. Unif$[0,1]$ and G_μ is the generalized inverse of F_μ that satisfies

$$G_\mu(u) \leqslant x \qquad \text{if and only if} \qquad u \leqslant F_\mu(x)$$

for $u \in (0,1)$ and $x \in \mathbb{R}$. Therefore $\mathbf{1}_{X_k \leqslant x} = \mathbf{1}_{U_k \leqslant F_\mu(x)}$, and hence $F_{L_n}(x) = F_{L_n'}(F_\mu(x))$, where L_n' is the empirical distribution of $U_1, \ldots, U_n$. Therefore (as $F_\lambda(F_\mu(x)) = F_\mu(x)$ for all x),

$$\|F_{L_n} - F_\mu\|_{\sup} = \|F_{L_n'} - F_\lambda\|_{\sup}$$

and we have shown that the latter goes to 0 almost surely, as $n \to \infty$. ■

Observe that the key element in the proof was applying SLLN to Bernoulli random variables. We have already seen in that case how the closeness of the sample mean to expectation can be strengthened using Bernstein's/Hoeffding's inequality. Following it up, one can strengthen the Glivenko-Cantelli theorem to show that

$$n^p \|F_{L_n} - F_\mu\|_{\sup} \xrightarrow{a.s.} 0$$

provided $p < \frac{1}{2}$. We leave this as exercise.

7. Using characteristic functions to prove laws of large numbers

Let $X_1, X_2, \ldots$ be i.i.d. with finite expectation μ. Let $\psi(t) = \mathbb{E}[e^{itX_1}]$ be the characteristic function of X_ks. Then,

$$\mathbb{E}[e^{it\frac{S_n}{n}}] = \prod_{k=1}^{n} \mathbb{E}[e^{itX_k/n}] = \varphi(t/n)^n.$$

As $\mathbb{E}[X_1]$ exists, by Theorem 39 in the appendix, it follows that φ is C^1-smooth and $\varphi'(t) = \mathbb{E}[iX_1 e^{itX_1}]$. Therefore, using the Taylor expansion of φ near 0, we get $\varphi(u) = 1 + i\mu u + o(u)$ as $u \to 0$. Therefore, for fixed t,

$$\mathbb{E}[e^{it\frac{S_n}{n}}] = \left(1 + i\mu\frac{t}{n} + o(1/n)\right)^n \to e^{i\mu t}$$

as $n \to \infty$. But $t \mapsto e^{i\mu t}$ is the characteristic function of δ_μ. Lévy's continuity theorem implies that $\frac{S_n}{n} \xrightarrow{d} \delta_\mu$. As the limiting distribution is degenerate, it follows that $\frac{S_n}{n} \xrightarrow{P} \mu$. Thus we have proved WLLN under first moment assumption.

Assume that $\mu = 0$ and that $\mathbb{E}[X_1^2] < \infty$, so that $\varphi(u) = 1 + O(u^2)$ as $u \to 0$. Let $h_n \to \infty$ as $n \to \infty$ and consider S_n/h_n. Its characteristic function is

$$\begin{aligned}\mathbb{E}[e^{itS_n/h_n}] &= \prod_{k=1}^{n} \mathbb{E}[e^{itX_k/h_n}] = \varphi(t/h_n)^n \\ &= \Big(1 + O(1/h_n^2)\Big)^n \to 1\end{aligned}$$

if $nh_n^{-2} \to 0$. As the constant function 1 is the characteristic function of the zero random variable, by Lévy's continuity theorem we get $\frac{S_n}{h_n} \xrightarrow{d} 0$. As the limit is constant, $\frac{S_n}{h_n} \xrightarrow{P} 0$ for any h_n such that $h_n^2/n \to \infty$.

This finishes alternate proofs of several WLLN type results we had seen earlier. I am not aware of any approach to the strong laws using characteristic functions.

8. Strong law for certain non-independent random variables

The techniques that we used allow us to prove strong law for certain sequences of random variables even if there is no independence. Such questions may arise in contexts that have no explicit mention of probability. For example, in analysis, the sequence $(\sin(n\theta))_n$ looks like a sequence of random numbers (for a.e. θ, it fails spectacularly if you take $\theta = \pi/2$ for example). In number theory, many arithmetical sequences like the Möbius function may satisfy laws of large numbers (on average equal number of zeros and ones occur in the sequence).

For example, let $X_n = \alpha^n$ for $n \geqslant 0$, where α is picked uniformly at random from the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ (this just means that $\alpha = e^{2\pi i V}$ where $V \sim \text{Unif}[0,1]$). Clearly, X_n are not independent. But we can still see that

$$\frac{1}{N}\sum_{k=1}^{N} X_k = \begin{cases} \frac{1-\alpha^{N+1}}{N(1-\alpha)} & \text{if } \alpha \neq 1, \\ 1 & \text{if } \alpha = 1. \end{cases}$$

As $|1 - \alpha^{N+1}| \leqslant 2$, it follows that $\frac{1}{N}\sum_{k=1}^{N} X_k \to 0$ for all $\alpha \neq 1$. Extending this, we can see that if α is not a root of unity, then $\frac{1}{N}\sum_{k=1}^{N} X_k^p \to 0$ for all $p \geqslant 1$. In particular

$$\frac{1}{N}\sum_{k=1}^{N} X_k^p \xrightarrow{a.s.} 0 \text{ for all } p \geqslant 1.$$

What is $n_1 < n_2 < n_3 < \ldots$ is a fixed sequence. Is it true that $\frac{1}{N}\sum_{k=1}^{N} X_{n_k}^p \xrightarrow{a.s.} 0$? The geometric series formula does not apply, and the answer is not clear. It suffices to take $p = 1$ (just replace α by α^p which is also uniform on S^1). We show that it is true[5].

[5] Abhishek Khetan raised this question to me, and while we were both certain it must be somewhere in the literature, we could not locate a proof. We worked out for ourselves the proof given here.

Theorem 26

Let $n_1 < n_2 < n_3 < \ldots$. Then $\frac{1}{N}\sum_{k=1}^{N} X_{n_k} \overset{a.s.}{\to} 0$.

Now we do not have independence. But the random variables (complex-valued, but that is no big deal[6]) are bounded. The more basic techniques based on Chebyshev inequality are more amenable, as they only require pairwise correlations, and it is easy to see that

$$\mathbb{E}[X_n \bar{X}_m] = \int_0^{2\pi} e^{2\pi i(n-m)\nu} \frac{d\nu}{2\pi} = \begin{cases} 0 & \text{if } m \neq n, \\ 1 & \text{if } m = n. \end{cases}$$

Proof. Let $Y_N = \frac{1}{N}\sum_{k=1}^{N} X_k$. Then $\mathbb{E}[Y_N] = 0$ and $\mathbb{E}[|Y_N|^2] = \frac{1}{N}$. Therefore,

$$\mathbb{P}\{|Y_N| \geqslant \delta\} \leqslant \frac{\mathbb{E}[|Y_N|^2]}{\delta^2} \leqslant \frac{1}{N\delta^2}.$$

This is summable, for any $\delta > 0$, over the subsequence $N_k = k^2$. Therefore, by Borel-Cantelli we see that $Y_{k^2} \overset{a.s.}{\to} 0$.

For general n, find k such that $k^2 \leqslant n < (k+1)^2$ and observe that

$$\begin{aligned} |Y_n| &\leqslant \Big|\frac{1}{n}\sum_{j=1}^{k^2} X_j\Big| + \frac{1}{n}\sum_{j=k^2+1}^{n} |X_j| \\ &\leqslant |Y_{k^2}| + \frac{2k+1}{n}. \end{aligned}$$

This shows that $Y_n \overset{a.s.}{\to} 0$ as $n \to \infty$. ■

Notice the similarity to the proof of SLLN under second moment assumption that we gave earlier. The difference is that there we used the positivity of X_js to sandwich S_n between S_{k^2} and $S_{(k+1)^2}$ but here we used the boundedness of X_js to get that control.

[6] If $X = X_1 + iX_2$ is a complex valued random variable, then $\mathbb{E}[X]$ just means $\mathbb{E}[X_1] + i\mathbb{E}[X_2]$ etc. No new definitions are needed and we can write everything in terms of real-valued random variables. It is just that the complex notation could be more convenient.

CHAPTER 6

Central limit theorems

Laws of large numbers apply when there is deterministic behaviour arising out of randomness. As we saw, this happens when the system size goes to infinity. But at any finite size of the system, there are fluctuations from the deterministic behaviour, and they may be important.

For example, traveling to the airport may take 60 minutes on average, but one must make allowance for random happenstances that increase the time above average. In planning frequency of buses or trains, average numbers of passengers is a key input, but the system for allow for more or less people showing up at a particular time and day. A dangerous chemical may be packed in a tight container, but some of the molecules are sure to leak out by chance - how many and is it safe?

Central limit theorems (or more general convergence in distribution statements) describe such fluctuations. A one line summary would be that in laws of large numbers, all probabilities of interest are close to 0 or to 1, whereas in central limit behaviour, events of all probabilities between 0 and 1 feature. Nevertheless, there is a remarkable regularity or universality in that while there are a great many different ways to change the "microscopic details" (the random variables that make up the system), but only a few distinct behaviours for the fluctuations. We shall see a few theorems about sums of random variables, in most of which the fluctuations turn out to be Gaussian or Poisson, even though the model description does not have anything to do with these distributions!

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 $\mathbb{E}[S_n^2] = n\sigma^2$, 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 27: Central limit theorem for i.i.d. variables

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

Informally, letting Z denote a standard Normal variable, we may write $S_n \approx n\mu + \sigma\sqrt{n}Z$. More precisely, $\mathbb{P}\{S_n \leqslant n\mu + \sigma\sqrt{n}t\} \to \mathbb{P}\{Z \leqslant t\}$ 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 $(2n)^{-\frac{1}{2}}S_{2n} \xrightarrow{d} Y$ and further,

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

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

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

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

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

Thus we must have $Y_1 + Y_2 \stackrel{d}{=} \sqrt{2}Y$. If $Y_1 \sim N(0, \sigma^2)$, then certainly it is true that $Y_1 + Y_2 \stackrel{d}{=} \sqrt{2}Y$. We claim that $N(0, \sigma^2)$ are the only distributions that have this property. If so, then it gives a strong heuristic that the central limit theorem is true. 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 $\mathbb{E}[X_1] = 1$ and $\mathrm{Var}(X_1) = 1$. Then $S_n \sim \mathrm{Gamma}(n, 1)$ and hence $\frac{S_n - n}{\sqrt{n}}$ has density

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

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

Exercise 22

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

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

Fix $k = (k_1, \dots, k_{2p})$ and consider the corresponding summand. The expectation factors as a product of $\mathbb{E}[X^{\ell_i}]$, $1 \leqslant i \leqslant n$, 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 $n(n-1)\dots(n-p+1)$, and the number of ways that these indices may be chosen is $(2p-1)(2p-3)\dots(3)(1)$.

Next those terms in which at least one ℓ_i is equal to 4. 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 $\int f(x/\sqrt{t}) p_t(x)dx = \mathbb{E}[f(Z)]$ 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*

$$\mathbb{E}[Zf'(Z)] = \mathbb{E}[f''(Z)]. \tag{20}$$

We leave it as an exercise to justify the differentiation under integral and the integration by parts. If we set $h = f'$, then (20) transforms to

$$\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)], \tag{21}$$

which is often called *Stein's identity*[1]. With a bit more care, one can prove that (21) holds for any $h : \mathbb{R} \to \mathbb{R}$ that is absolutely continuous with $h' \in L^1(\gamma)$ (this means that $h(x) - h(0) = \int_0^x g(t)dt$ 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 (20) with $f(x) = x^{2p}$ we get $\mathbb{E}[Z^{2p}] = (2p-1)\mathbb{E}[Z^{2p-2}]$, from which it follows that

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

2.4. Characteristic function. Formally one can see that $\mathbb{E}[e^{itZ}] = e^{-\frac{1}{2}t^2}$ by substituting it in the moment generating function. That can be made into an honest proof by first arguing that $w \mapsto \mathbb{E}[e^{wZ}]$ is an entire function (which is equal to the characteristic function on the imaginary axis and equal to the moment generating function on the real axis). Two entire functions that agree on the real line must agree everywhere, hence the claim follows.

Another way (avoiding complex analysis) is to apply the integration by parts formula to $f(x) = e^{itx}$ to get $\mathbb{E}[itZe^{itZ}] = -t^2\mathbb{E}[e^{itZ}]$. Setting $\varphi(t) = \mathbb{E}[e^{itZ}]$ we see (again, differentiating under the expectation) that $\varphi'(t) = -t\varphi(t)$, for which the unique solution satisfying $\varphi(0) = 1$ is

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

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 (20) and (21) are also characterizations of the standard Gaussian distribution. This means that if $\mathbb{E}[h'(W)] = \mathbb{E}[Wh(W)]$ 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 $W + W' \stackrel{d}{=} \sqrt{2}W$, then $W \sim N(0, \sigma^2)$ for some $\sigma^2 \geqslant 0$. To see this, suppose $\psi(\cdot)$ denotes the characteristic function of W,

[1]As Arka Das pointed out in class, (21) can be got directly by writing $\mathbb{E}[f'(Z)] = \int f'(x)\varphi(x)dx$ 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 (20) is related to the Ornstein-Uhlenbeck process, a Markov process with stationary distribution $N(0,1)$.

then

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

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 28

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

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

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

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

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 $W_n \xrightarrow{d} W$, so we prove the relevant one-way implications in the first and third statements.

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

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

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

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

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

(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 $\mathbb{E}[e_t(S_n/\sqrt{n})] \to e^{-t^2/2}$ 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 $\mathbb{E}[e_t(X_1/\sqrt{n})] \approx 1 - \frac{t^2}{2n}$ if $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$. Hence $\mathbb{E}[e_t(S_n/\sqrt{n})] \approx (1 - \frac{t^2}{2n})^n \approx e^{-t^2/2}$. 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 $\mathbb{E}[f(S_n^X/\sqrt{n})] \approx \mathbb{E}[f(S_n^Z/\sqrt{n})]$. As the latter is the same as $\mathbb{E}[f(Z)]$, 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 29

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 27.

4.1. Proof via characteristic functions. We shall need the following facts.

Exercise 23

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

PROOF OF THEOREM 29. 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 $\psi_n(t) := \mathbb{E}\left[e^{itS_n/\sqrt{n}}\right]$. Writing $S_n = X_1 + \ldots + X_n$ and using independence,

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

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 $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = 1$ to get

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

Clearly, $|R_n(t)| \leqslant C_t n^{-3/2}$ for a constant C_t (that depends on t but not n). Hence $nR_n(t) \to 0$ and by Exercise 23 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 $0 \leqslant k \leqslant n$ 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,

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

By Taylor expansion, we see that

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

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

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

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

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

Add up over k from 1 to n to get

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

which goes to zero as $n \to \infty$. Thus, $\mathbb{E}[f(S_n/\sqrt{n})] \to \mathbb{E}[f(Y_1)]$ 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 30: Lindeberg-Feller CLT

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

(3) For any $\delta > 0$, we have $\sum_{k=1}^{n} \mathbb{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] \to 0$ 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 27 from Theorem 30. Let $X_{n,k} = n^{-\frac{1}{2}} X_k$ for $k = 1, 2, \ldots, n$. Then, $\mathbb{E}[X_{n,k}] = 0$ while $\sum_{k=1}^{n} \mathbb{E}[X_{n,k}^2] = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[X_1^2] = \sigma^2$, for each n. Further, $\sum_{k=1}^{n} \mathbb{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] = \mathbb{E}[X_1^2 \mathbf{1}_{|X_1|>\delta\sqrt{n}}]$ which goes to zero as $n \to \infty$ by DCT, since $\mathbb{E}[X_1^2] < \infty$. 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 22

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 $(S_n - \mathbb{E}[S_n])/\sqrt{\text{Var}(S_n)}$ converge in distribution to $N(0,1)$? Obviously the standard CLT does not apply.

To fit it in the framework of Theorem 30, 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 30 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 $1 \leqslant k \leqslant n$). 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 $|X_k - p_k| \leqslant 1$). 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 $(S_n - \mathbb{E}[S_n])/\tau_n$. Changing X_1 from 0 to 1 or vice versa will induce a big change in the value of $(S_n - \mathbb{E}[S_n])/\tau_n$ 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 24: Lyapunov's central limit theorem

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

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

Example 23

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 $\mathbb{E}[T_n] = 0$ and $\mathbb{E}[T_n^2] = \sum_{k=1}^n k^2 \sim \frac{n^3}{3}$. 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 30.

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

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

Exercise 25

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 26

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

Proof of Theorem 30. The characteristic function of $S_n = X_{n,1} + \ldots + X_{n,n}$ is given by $\psi_n(t) = \prod_{k=1}^n \mathbb{E}\left[e^{itX_{n,k}}\right]$. 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 $|x| \leqslant \delta$ and the second one for $|x| > \delta$ to write

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

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

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

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

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

We want to apply Exercise 26 to $z_k = \mathbb{E}\left[e^{itX_{n,k}}\right]$ and $w_k = 1 - \frac{1}{2}\sigma_{n,k}^2t^2$. Clearly $|z_k| \leqslant 1$ by properties of c.f. If we prove that $\max_{k\leqslant n}\sigma_{n,k}^2 \to 0$, then it will follow that $|w_k| \leqslant 1$ and hence with $\theta = 1$ in Exercise 26, we get

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

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

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

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

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

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

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

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

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

From 23 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 $\mathbb{E}[Y_{n,k}] = \mathbb{E}[X_{n,k}]$ and $\mathbb{E}[Y_{n,k}^2] = \mathbb{E}[X_{n,k}^2]$.

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 $0 \leqslant k \leqslant n$. 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,

$$\mathbb{E}[f(V_n)] - \mathbb{E}[f(V_0)] = \sum_{k=1}^{n} \mathbb{E}[f(V_k) - f(V_{k-1})] \tag{24}$$
$$= \sum_{k=1}^{n} \mathbb{E}[f(V_k) - f(U_k)] - \sum_{k=1}^{n} \mathbb{E}[f(V_{k-1}) - f(U_k)].$$

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

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

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

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

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

$$|\mathbb{E}[f(V_k)] - \mathbb{E}[f(V_{k-1})]| \leqslant \frac{1}{2}\mathbb{E}[|f''(U_k)|] \left|\mathbb{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|\leqslant\delta}] - \mathbb{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|\leqslant\delta}]\right| \tag{25}$$
$$+ \frac{1}{2}\left|\mathbb{E}[|f''(U_k^\#)|X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}]\right| + \frac{1}{2}\left|\mathbb{E}[|f''(U_k^{\#\#})|Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\right| \tag{26}$$
$$+ \frac{1}{6}\left|\mathbb{E}[|f'''(U_k^*)||X_{n,k}|^3 \mathbf{1}_{|X_{n,k}|\leqslant\delta}]\right| + \frac{1}{6}\left|\mathbb{E}[|f'''(U_k^{**})||Y_{n,k}|^3 \mathbf{1}_{|Y_{n,k}|\leqslant\delta}]\right| \tag{27}$$

Since $\mathbb{E}[X_{n,k}^2] = \mathbb{E}[Y_{n,k}^2]$, the term in the first line (25) is the same as $\frac{1}{2}\mathbb{E}[|f''(U_k)|] \left|\mathbb{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] - \mathbb{E}[Y_{n,k}^2 \mathbf{1}_{|Y_{n,k}|>\delta}]\right|$ which in turn is bounded by

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

The terms in (26) are also bounded by

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

To bound the two terms in (27), we show how to deal with the first.

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

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

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

Add up over k and use (24) to get

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

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

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

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

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 $n\mathbb{E}[X_1]$) 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 24

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 $n\mathbb{E}[X_1]$, scaling is by n instead of $\sqrt{n}$ and the limit is Cauchy instead of Normal.

This raises the following questions.

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

(2) What are the possible limiting distributions?

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

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

Symmetric α-stable distributions Fix $\alpha > 0$. Do there exist i.i.d. random variables X, Y such that $X + Y \stackrel{d}{=} 2^{\frac{1}{\alpha}}X$? 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 31: Symmetric stable distributions

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

The proof that $e^{-|t|^\alpha}$ is a valid characteristic function for $0 < \alpha \leqslant 2$ is explained in Example **??**. That it fails to be a characteristic function for $\alpha > 2$ is explained in Example **??**. Let us give a second proof of the latter fact.

Proof of non-existence for $\alpha > 2$. If $\alpha \geqslant 2$, then $t \mapsto e^{-|t|^\alpha}$ is a C^2 function, with a maximum at 0. If a probability measure 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 $X + Y \stackrel{d}{=} 2^{1/\alpha}X$ where X, Y are i.i.d. μ_α, we see that $2\mathrm{Var}(X) = 2^{2/\alpha}\mathrm{Var}(X)$. Either $\mathrm{Var}(X) = 0$, in which case $X = 0$ a.s., or $\alpha = 2$, in which case $X \sim N(0, \sigma^2)$ for some $\sigma \geqslant 0$. ■

Henceforth, we shall write μ_α for the distribution with characteristic function $e^{-|t|^\alpha}$, for $0 < \alpha < 2$ (our convention is to keep the $\alpha = 2$ case of Gaussian outside the class of stable distributions). These distributions are heavy tailed. The proof above in fact shows that none of them can have finite variance.

Theorem 32: 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:

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

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

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

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 $n^{-\frac{1}{\alpha}} S_n \xrightarrow{d} \mu_\alpha$. 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 33: 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 $\mathrm{Bin}(n, \lambda/n) \xrightarrow{d} \mathrm{Pois}(\lambda)$ 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 34: Poisson convergence of sums of independent Bernoullis

Let $\xi_{n,j} \sim \mathrm{Ber}(p_{n,j})$, $1 \leqslant j \leqslant n$, 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) $p_n^* := \max_{j \leqslant n} p_{n,j} \to 0$ as $n \to \infty$. Then $S_n := \xi_{n,1} + \ldots + \xi_{n,n}$ converges in distribution to $\mathrm{Pois}(\lambda)$.

Proof. By a direct calculation,

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

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

$$\begin{aligned}e^{-\sum_{j=1}^{n}(p_{n,j}+p_{n,j}^2)} \leqslant \prod_{i=1}^{n}(1 - p_{n,j}) \leqslant e^{-\sum_{j=1}^{n} p_{n,j}}, \\ e^{p_n^*} \leqslant \frac{1}{1 - p_{n,j_r}} \leqslant e^{p_n^*(1+p_n^*)}.\end{aligned}$$

Thus,

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

Now, $\sum_{j=1}^{n} p_{n,j} \to \lambda$ and $\sum_{j=1}^{n} p_{n,j}^2 \leqslant p_n^* \sum_{j=1}^{n} p_{n,j} \to 0$. 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

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

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

Alternate proof. For $t \in \mathbb{R}$,

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

By Exercise 26,

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

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 $\text{Pois}(\lambda)$, we see that $S_n \xrightarrow{d} \text{Pois}(\lambda)$. ■

7.2. Proof of Theorem 33. 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 $\mathbb{P}\{X_1 \geqslant x\} \sim Cx^{-\alpha}$, 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 33 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 33, but technically much simpler.

Theorem 35: 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 $b(u) = \inf\{t : \mu([-t,t]^c) \leqslant u\}$. Then

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

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

$$\mathbb{E}\left[e^{itS_{K_n}/a_n}\right] = \exp\left\{\int_0^\infty (\cos(tu)-1)\,d\mu_n(u)\right\}. \tag{28}$$

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 $\varphi_\delta = \sum_{j\geqslant 1} j\delta(\mathbf{1}_{I_{j,\delta}} - \mathbf{1}_{-I_{j,\delta}})$. Then $\varphi_\delta(t) \to t$ as $\delta \downarrow 0$, and $|\varphi_\delta(t)| \leqslant t$. Hence, by DCT,

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

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 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 (28).

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

$$2\int_0^\infty (\cos t - 1)\,d\mu_n(t) \to \int_0^\infty (\cos t - 1)\frac{\alpha}{t^{\alpha+1}}dt. \tag{29}$$

Once we justify (29), 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 (29), 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)

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

[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.

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

More about sums of independent random variables

Sums of independent random variables are very important, and have played a central role in the development of the concepts of probability theory. People have delved far more deeply into this topic than anyone today wants to know[1]! In this chapter we give a few isolated snippets. Some of these are usually taught in probability courses, some not so much.

(1) Law of iterated logarithm.

(2) Cramer's theorem of large deviations.

(3) Anti-concentration inequalities.

(4) Error estimates in the central limit theorem.

1. The law of iterated logarithm

If $a_n \uparrow \infty$ is a deterministic sequence, then Kolmogorov's zero-one law implies that $\limsup \frac{S_n}{a_n}$ is constant a.s. What is this constant?

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

$$\mathbb{P}\{S_n \geqslant t_n\} \leqslant \mathbb{E}[S_n^{2p}] t_n^{-2p} \quad \leqslant \quad C_p n^p t_n^{-2p}$$

where we used the fact that $\mathbb{E}[S_n^{2p}] \leqslant C_p n^p$ 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 $n^{-\alpha} S_n \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$.

To see that $\mathbb{E}[S_n^{2p}] \leqslant C_p n^p$, 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 $\mathbb{E}[X_1^{k_1}] \dots \mathbb{E}[X_n^{k_n}]$. 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

[1] A great deal of it was developed in the Soviet union. One particular reference is Petrov's *Sums of independent random variables*

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 $\binom{n}{p} \leqslant n^p$ 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 $\mathbb{E}[S_n^{2p}] \leqslant C_p n^p$, 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 lim sup is a constant a.s., we can scale by that and reformulate the question as follows.

Question: Let X_i be i.i.d $\mathrm{Ber}_{\pm}(1/2)$ 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 36: Khinchine's law of iterated logarithm

Let X_i be i.i.d. $\mathrm{Ber}_{\pm}(1/2)$ 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 1: Hartman-Wintner law of iterated logarithm

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

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

2. Proof of LIL for Bernoulli random variables

Let $X_1, X_2, \ldots$ be i.i.d. $\mathrm{Ber}_{\pm}(1/2)$ random variables. Theorem 36 follows from the following two statements. For any $\delta > 0$, we have

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \leqslant 1+\delta \quad \text{a.s.} \tag{30}$$

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \geqslant 1-\delta \quad \text{a.s.} \tag{31}$$

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

Example 25: 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 $\sum n\mathbb{P}(B_n) < \infty$. 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 $\sum \mathbb{P}(B_n) < \infty$, a much weaker condition!

What this suggests is that when we have a sequence of A_ns and want to show that $\mathbb{P}\{A_n \text{ i.o.}\} = 0$, 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 $n_k \leqslant n < n_{k+1}$. If only finitely many C_ks occur, the only finitely many A_ns occur, and thus it suffices to show that $\sum_k \mathbb{P}(C_k) < \infty$. The naive union bound $\mathbb{P}(C_k) \leqslant \sum_{n=n_k}^{n_{k+1}} \mathbb{P}(A_n)$ takes us back to the condition $\sum_n \mathbb{P}(A_n) < \infty$, but the point is that there may be better bounds for $\mathbb{P}(C_n)$ than the union bound.

Proof of the upper bound (30). 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

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

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

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

$$\mathbb{P}(D_k) \leqslant C_\lambda k^{-(1+\delta)^2/\lambda} \quad \text{where } C_\lambda < \infty \text{ for any } \lambda > 1. \tag{32}$$

Granting this, it is clear that choosing $1 < \lambda < (1+\delta)^2$ ensures summability of $\mathbb{P}(D_k)$. We give two proofs of the inequality (32) below, which completes the proof. ■

Proof of (32) via the reflection principle: The following lemma is of interest in itself and useful.

Lemma 18: Reflection principle/Ballot problem

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

$$2\mathbb{P}\{S_n > a\} \leqslant \mathbb{P}\{\max\{S_0, \ldots, S_n\} \geqslant a\} \leqslant 2\mathbb{P}\{S_n \geqslant a\}.$$

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 $\max\{S_0, \ldots, S_n\} \geqslant a$ as a union of pairwise disjoint events

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

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

$$\begin{aligned} \mathbb{P}(\{S_n \geqslant a\} \cap A_k) &= \mathbb{P}(\{S_n - S_k \geqslant 0\} \cap A_k) \\ &= \mathbb{P}\{S_n - S_k \geqslant 0\}\mathbb{P}\{A_k\} \ \geqslant \ \frac{1}{2}\mathbb{P}(A_k). \end{aligned} \tag{33}$$

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

$$\begin{aligned} \mathbb{P}(\{S_n > a\} \cap A_k) &= \mathbb{P}(\{S_n - S_k > 0\} \cap A_k) \\ &= \mathbb{P}\{S_n - S_k > 0\}\mathbb{P}\{A_k\} \ \leqslant \ \frac{1}{2}\mathbb{P}(A_k). \end{aligned} \tag{34}$$

Add up over k to get $2\mathbb{P}\{S_n > a\} \leqslant \mathbb{P}\{\max\{S_0, \ldots, S_n\} \geqslant a\}$.

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

Returning to the proof of (32), if D_k occurs, then there is some $n \leqslant n_{k+1}$ (in fact some $n \geqslant n_k$) such that $S_n \geqslant a_{n_k}(1+\delta)$. The reflection principle in Lemma 18 applies to give the bound

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

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

$$\frac{(1+\delta)^2 2\lambda^k \log\log\lambda^k}{2\lambda^{k+1}} = \frac{(1+\delta)^2}{\lambda}\log(k\log\lambda) \tag{35}$$

from which (32) immediately follows. ■

Proof of (32) via the modified Markov inequality (??): 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 $X_k \geqslant 1$. Apply the strengthened form of Markov's inequality (**??**) to write

$$\mathbb{P}(D_k) = \mathbb{P}\{X_k \geqslant 1\} \leqslant \frac{\mathbb{E}[X_k]}{\mathbb{E}[X_k \mid X_k \geqslant 1]}.$$

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

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

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

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

Next we claim that $c(n_{k+1}-n_k)$ (for some $c>0$) is a lower bound for $\mathbb{E}[X_k \mid X_k \geqslant 1]$. The heuristic idea is that if $X_k \geqslant 1$, there is some (random) $N \in [n_k, n_{k+1})$ for which $S_N \geqslant a_{n_k}(1+\delta)$. 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 $\mathbb{P}\{S_n - S_N \geqslant 0\} \geqslant \frac{1}{2}$, which would imply that $\mathbb{E}[X_k \mid X_k \geqslant 1] \geqslant \frac{1}{2}(n_{k+1}-n_k)$. 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 $X_k \mathbf{1}_{X_k \geqslant 1} \geqslant \frac{1}{2}(Y_k \mathbf{1}_{Z_k \geqslant 1} + Z_k \mathbf{1}_{Y_k \geqslant 1})$ and $\{X_k \geqslant 1\} \subseteq \{Y_k \geqslant 1\} \cup \{Z_k \geqslant 1\}$. Consequently,

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

In the second line we used the elementary inequality $\frac{a+b}{c+d} \geqslant \min\{\frac{a}{c}, \frac{b}{d}\}$ valid for any non-negative numbers a, b, c, d. Now consider the second term inside the minimum. Since $Y_k \geqslant 1$, 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 $S_n - S_N$, $n \geqslant N$, 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, $\mathbb{E}[Z_k \mid Y_k \geqslant 1] \geqslant \frac{1}{4}(n_{k+1} - n_k)$. Similarly (considering the random walk in backwards direction starting from n_{k+1}), reason that $\mathbb{E}[Y_k \mid Z_k \geqslant 1] \geqslant \frac{1}{4}(n_{k+1} - n_k)$. Putting all this together, $\mathbb{E}[X_k \mid X_k \geqslant 1] \geqslant \frac{1}{8}(n_{k+1} - n_k)$.

Thus,

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

By the computation shown in (35), this is of the form given in (32). ■

2.1. Proof of the lower bound (31). 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

$$\mathbb{P}\{S_{n_k} \geqslant (1-2\delta)a_{n_k} \text{ i.o.}\} = 1 \tag{36}$$

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, $S_{n_k} \geqslant -2a_{n_k}$ for all but finitely many k. Also, $a_{n_k} \leqslant a_{n_{k+1}}/\sqrt{\lambda}$, hence

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

for all but finitely many k, a.s. Therefore, (36) follows if we choose $\lambda > 4/\delta^2$ and show that

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

These events are independent across k, and hence a good lower bound on the individual probabilities is sufficient. The one given below in Claim 7 gives

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

Claim 7: An estimate for binomial coefficients

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

$$\mathbb{P}\{S_n \geqslant k\} \geqslant e^{-\frac{1}{2}\frac{k^2}{2n}}$$

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. ■

3. Law of iterated logarithm for general i.i.d. random variables

Hartman and Wintner showed in that if X_k are i.i.d. with zero mean and unit variance, then

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

This extends Khinchine's LIL for Bernoulli random variables. It also immediately implies that the lim inf of the same quantities is -1 and that the set of limit points of the sequence $\{S_n/\sqrt{2n\log\log n}\}$ is equal to $[-1, 1]$ a.s. The easiest way to prove this is using Brownian motion. For now, we present an earlier law of iterated logarithm due to Kolmogorov, which is restrictive in asking for the X_k to be bounded, but relaxes the requirement of identical distribution.

Theorem 37: Kolmogorov (1929)

Let X_k be independent random variables with $\mathbb{E}[X_k] = 0$, $\mathrm{Var}(X_k) = \sigma_k^2$ and $|X_k| \leqslant B_k$ a.s. Let $\tau_n^2 = \sigma_1^2 + \ldots + \sigma_n^2$ and assume that (1) $\tau_n^2 \to \infty$ as $n \to \infty$ and (2) $B_n = o(\tau_n^2/\sqrt{\log\log\tau_n})$. Then,

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

The excuse for discussing this theorem is to show the techniques. In particular, pay attention to how Kolmogorov finds substitutes for the estimates that were easily obtained for Bernoulli random variables, some of which look delicate and not easy to generalize. These are (1) Bernstein like estimate for the probability that S_n is large. (2) Corresponding lower bound of Gaussian type. (3) Reflection principle that allowed to control the maximum of S_k, $k \leqslant n$, by S_n.

The key probability estimates are in the following lemma.

Lemma 19: Gaussian tail bounds analogue

Let X_k be as in the statement of Theorem 37. Fix $n \geqslant 1$. Let $B_n^* = \max_{k\leqslant n} B_k$.

(1) Upper bound:

$$\mathbb{P}\{S_n \geqslant t\} \leqslant \begin{cases} e^{-(1-\varepsilon)\frac{t^2}{2\tau_n^2}} & \text{for } t \leqslant \varepsilon\frac{\tau_n^2}{B_n^*}, \text{ where } \varepsilon \leqslant 1. \\ e^{-\frac{t}{4B_n^*}} & \text{for } t \geqslant \frac{\tau_n^2}{B_n^*}. \end{cases}$$

(2) $\mathbb{P}\{S_n \geqslant t\} \geqslant e^{-(1+\varepsilon)\frac{t^2}{2\tau_n^2}}$ for $a_\varepsilon\tau_n \leqslant t \leqslant b_\varepsilon\frac{\tau_n^2}{B_n}$ where $a_\varepsilon \to \infty$ and $b_\varepsilon \to 0$ as $\varepsilon \downarrow 0$. In fact, we may take $b_\varepsilon = \Theta(\varepsilon^2)$ and $a_\varepsilon = \Theta(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon})$.

Let $S_n^* = \max\{S_0, S_1, \ldots, S_n\}$.

Lemma 20: Reflection principle analogue

$\mathbb{P}\{S_n^* \geqslant t\} \leqslant 2\mathbb{P}\{S_n \geqslant t - \sqrt{2}\tau_n\}$.

Proof. Let $J = \min\{k \leqslant n : S_k \geqslant t\}$, which is well-defined on the event $\{S_n^* \geqslant t\}$. Conditional on $J = j$, we know that $S_j \leqslant t + B_j$. By Chebyshev's inequality, $S_n - S_j \geqslant -\sqrt{2}(\tau_n - \tau_j)$ with probability at least $\frac{1}{2}$. Thus $\mathbb{P}\{S_n \geqslant t + B_j - \sqrt{2}(\tau_n - \tau_j) \,\big|\, J = j\} \geqslant \frac{1}{2}$. Dropping the $B_j + \sqrt{2}\tau_j$ term which is positive, $\mathbb{P}\{S_n \geqslant t - \sqrt{2}\tau_n \,\big|\, J = j\} \geqslant \frac{1}{2}$. Multiply by $\mathbb{P}\{J = j\}$ and sum over $1 \leqslant j \leqslant n$ to get

$$\mathbb{P}\{S_n \geqslant t - \sqrt{2}\tau_n\} \geqslant \frac{1}{2}\sum_{j=1}^{n} \mathbb{P}\{J = j\} = \frac{1}{2}\mathbb{P}\{S_n^* \geqslant t\}.$$

This proves the lemma. ■

Now we come to the proof of the LIL. Let $\varphi(n) = \sqrt{2\tau_n \log\log \tau_n}$. The second assumption can be written as $\frac{B_n \varphi(n)}{\tau_n^2} \to 0$.

Proof of the upper bound in the LIL. Fix $0 < \delta < 1$ and choose $n_0 > \frac{1}{\delta}$ such that $\sqrt{\log\log \tau_{n_0}} > \frac{8}{\delta}$ and $2B_n\varphi(n) \leqslant \delta\tau_n^2$ for $n \geqslant n_0$. Then define $n_0 < n_1 < n_2 < \ldots$ successively by choosing $n_{k+1} = \lfloor (1+\delta)n_k \rfloor$. The condition $\delta n_0 > 1$ ensures that $n_{k+1} > n_k$.

Let $E_n = \{S_n \geqslant (1+\delta)^2\varphi(n)\}$ and let $E_k^* = \cup_{n=n_k+1}^{n_{k+1}} E_n$. The goal is to show that $\mathbb{P}(E_k^*)$ is summable, which implies that only finitely many E_k^*s occur a.s., hence only finitely many E_ns occur a.s.

Observe that if E_k^* occurs then $S_{n_{k+1}}^* \geqslant (1+\delta)^2\varphi(n_k)$. By Lemma 20,

$$\begin{aligned}
\mathbb{P}\{E_k^*\} &\leqslant \mathbb{P}\{S_{n_{k+1}} \geqslant (1+\delta)^2\varphi(n_k) - \sqrt{2}\tau_{n_{k+1}}\} \\
&\leqslant \mathbb{P}\{S_{n_{k+1}} \geqslant (1+\delta)\varphi(n_{k+1}) - \frac{\varphi(n_{k+1})}{\sqrt{\log\log \tau_{n_{k+1}}}}\} \\
&\leqslant \mathbb{P}\{S_{n_{k+1}} \geqslant (1+\frac{7\delta}{8})\varphi(n_{k+1})\}
\end{aligned}$$

as $\sqrt{\log\log \tau_{n_0}} > \frac{8}{\delta}$. By the first part of Lemma 19, with $\varepsilon = \delta$ (satisfied by the requirement $2B_n\varphi(n) \leqslant \delta\tau_n^2$ for $n \geqslant n_0$), we see that this probability is bounded by

$$\begin{aligned}
\exp\left\{-(1-\delta)(1+\frac{7}{8}\delta)^2\frac{\varphi(n_{k+1})^2}{2\tau_{n_{k+1}}^2}\right\} &\leqslant \exp\{-(1+\frac{\delta}{2})\log\log \tau_{n_{k+1}}\} \quad \text{(if } \delta \text{ if small enough)} \\
&= \frac{1}{(\log \tau_{n_{k+1}})^{1+\frac{\delta}{2}}}.
\end{aligned}$$

As $\log \tau_{n_k} \asymp k$, this is summable. Hence E_k^* occurs only finitely many times, a.s. ■

PROOF OF THE LOWER BOUND IN THE LIL. Fix $\theta > 1$ and $\delta > 0$. Inductively choose n_{k+1} to be the smallest $n > n_k$ so that $\theta\tau_{n_k} \leqslant \tau_n$. Then $\log \tau_{n_k} \sim k \log \theta$. By the limsup upper bound and negating the X_is, we see that $S_n \geqslant -(1+\delta)\varphi(n)$ for all large n.

Now fix some k and suppose that $S_{n_k} \geqslant -(1+\delta)\varphi(n_k)$. We apply the second part of Lemma 19 with $n = n_{k+1} - n_k$ and $t = (1-\delta)\varphi(n_{k+1})$. Then

(1) $\frac{t}{\tau_n} \geqslant c\sqrt{\log\log \tau_n} \to \infty$ and

(2) $\frac{t}{\tau_n^2/B_n} \leqslant \frac{B_n\sqrt{2\log\log \tau_n}}{\tau_n} \to 0$ by assumption.

Therefore, Lemma 19 applies for large enough k, and

$$\mathbb{P}\{S_{n_{k+1}} - S_{n_k} \geqslant (1+\delta)\varphi(n_{k+1})\} \geqslant e^{-(1+\delta)\frac{\varphi(n_{k+1})^2}{2\tau_{n_{k+1}} - \tau_{n_k}}}.$$

The exponent is

$$\frac{\tau_{n_{k+1}} \log\log \tau_{n_{k+1}}}{\tau_{n_{k+1}} - \tau_{n_k}} \geqslant \log k$$

■

4. Anti-concentration

CHAPTER 8

Appendix: Characteristic functions and their properties

Definition 17

Let $\mu \in \mathcal{P}(\mathbb{R})$. The function $\hat{\mu} : \mathbb{R}^d \to \mathbb{C}$ defined by $\hat{\mu}(t) := \int_{\mathbb{R}} e^{itx} d\mu(x)$ is called the *characteristic function* or the *Fourier transform* of μ. If X is a random variable whose distribution is μ, we also refer to $\hat{\mu}$ as the "characteristic function of X" and denote it ψ_X.

There are various other "integral transforms" of a measure that are closely related to the c.f. For example, $t \mapsto \int e^{tx} d\mu(x)$ (if it exists) is called the moment generating function of μ . The probability generating function of a probability measure μ supported on $\mathbb{N}$ is defined by $t \mapsto \int t^x d\mu(x) = \sum_{k \geqslant 0} \mu\{k\} t^k$ (which exists for $|t| < 1$), and so on. 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 $\mu_n \xrightarrow{d} \mu$. The forward implication is the key property that is used in proving central limit theorems.

(D) There exist necessary and sufficient conditions for a complex valued function on the real line 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. We state some basic properties of characteristic functions.

[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 38

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

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

Lemma 21

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

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

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

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 22: Parseval's identity

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

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

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 39: 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 $h^{-1}|e^{i(t+h)x} - e^{itx}| \leqslant |x|$ 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 40: 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 $\|f - g\|_{L^1(\mathbb{R})} \leqslant \varepsilon$. Then $\|\hat{f} - \hat{g}\|_{\sup} \leqslant \varepsilon$ (we use $\hat{f}(t)$ for $\int f(x)e^{itx} dx$). Therefore,

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

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 41: Smoothness of the characteristic function to the decay of the measure

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

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

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

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

This is the claim. ■

Theorem 42: 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 $\hat{\mu}(t) = p_1 e^{ita_1} + \ldots + p_k e^{ita_k}$.

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

(3) Let $X, X' \sim \text{unif}[-1, 1]$ 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 $|x| \leqslant 2$. 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 $\text{Pois}(\lambda)$ distribution is

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

(5) If $X \sim \text{Exp}(\lambda)$, then $\psi_X(t) = \int_0^\infty \lambda e^{-\lambda x} e^{itx} dx = \frac{\lambda}{\lambda - it}$. If $Y \sim \text{Gamma}(\nu, \lambda)$, then if ν is an integer, then $Y \stackrel{d}{=} X_1 + \ldots + X_\nu$ where X_k are i.i.d $\text{Exp}(\lambda)$. 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. The idea is that if $\psi(t) = \mathbb{E}[e^{itZ}]$ where $Z \sim N(0,1)$, then differentiating under the integral,

$$\psi'(t) = \mathbb{E}[iZe^{itZ}] = -t^2\mathbb{E}[e^{itZ}] = -t^2\psi(t)$$

where the second equality uses Stein's identity ($\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)]$ for all reasonable h). The only solution to this differential equation satisfying $\psi(0) = 0$ is $\psi(t) = e^{-t^2/2}$.

(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 $-R \leqslant u \leqslant R$ and $\eta(u) = Re^{is}$ for $0 \leqslant s \leqslant \pi$. 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 $\frac{e^{-t\,\text{Im}\,z}}{|1+z^2|} \leqslant \frac{1}{R^2-1}$, 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 $\psi(t) = e^{-|t|}$.

[2]Here is the argument: Fix $R > 0$ and let $\gamma(u) = u$ and $\eta(t) = u + it$ for $-R \leqslant u \leqslant R$ and let $\eta'_x(s) = x + is$ for $0 \leqslant s \leqslant t$. 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 $e^{-\frac{1}{2}(R^2+s^2)} \leqslant e^{-R^2/2}$ for any $0 \leqslant s \leqslant t$. Thus the two integrals are bounded in absolute value by $e^{-R^2/2}|t|$ which goes to 0 as $R \to \infty$.

0.4. Inversion formulas. We now come to one of the most important reasons why characteristic function is a useful tool. Characteristic function determines the measure and we can write formulas for recovering a measure from the characteristic function[3].

Theorem 43

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

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

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

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

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

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

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

[3]The idea behind these arguments may not be clear unless one starts with a simpler situation. Consider a function $f \in L^1(\mathbb{T})$, where $\mathbb{T} = [-\pi, \pi]$ with the measure $\frac{d\theta}{2\pi}$. Then $e_n(\theta) = e^{in\theta}$, $n \in \mathbb{Z}$, form an orthonormal basis for $L^2(\mathbb{T})$, and hence we have the L^2-expansion $f = \sum_{n\in\mathbb{Z}} \hat{f}(n)e_n$, where $\hat{f}(n) = \langle f, e_n\rangle = \int_{\mathbb{T}} f(\theta)e^{-in\theta}\frac{d\theta}{2\pi}$. If we change the interval to $[-\pi L, \pi L]$ with uniform probability measure, then the orthonormal basis is $\{e_{n/L} : n \in \mathbb{Z}\}$. When $L \to \infty$, we may expect to get all $\{e_t : t \in \mathbb{R}\}$, and try to expand $f \in L^2(\mathbb{R})$ as a superposition of these complex exponentials. However, $e_t \notin L^2(\mathbb{R})$, and as there are uncountably many, they cannot possibly form an orthonormal basis either. A related point is that there is no uniform probability distribution on $\mathbb{R}$. However, the fact that $\frac{1}{2\pi(2L)}\lim_{L\to\infty}\int_{-L}^{L} e_t(x)\overline{e_s(x)}dx = \delta_{t-s}$ can be thought of as a form of orthonormality. And then we should expect that if $\hat{f}(t) = \int_{\mathbb{R}} f(x)\overline{e_t}(x)dx$, then $f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(t)e_t(x)dt$. This is indeed the Fourier inversion formula, but a proper proof will uncover the precise conditions on f and $\hat{f}$ that are needed. It is also not easy to justify the interchange the $L \to \infty$ limit with the inversion formula for finite L. The proofs here in some sense do that, by first multiplying f or $\hat{f}$ by $\mathbf{1}_{[-L,L]}$, or even better, by multiplying or convolving with Gaussian.

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 4: 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 $\lim_{\sigma\to\infty}\int_a^b f_\sigma$ 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

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

In the second line, Fubini's theorem was applicable as $(t,\alpha)\mapsto|\hat{\mu}(t)|e^{-\frac{t^2}{2\sigma^2}}$ is integrable over $\mathbb{R}\times[a,b]$, for $\sigma>0$. 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 $|f_\sigma(\alpha)|\leqslant C$ 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 43 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 44: Lévy's continuity theorem

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

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

(2) If $\hat{\mu}_n(t)\to\psi(t)$ pointwise for all t and ψ is continuous at 0, then $\psi=\hat{\mu}$ for some $\mu\in\mathcal{P}(\mathbb{R})$ and $\mu_n\xrightarrow{d}\mu$.

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 $\mu_n\xrightarrow{d}\mu$.

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

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

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

It remains to show tightness[4]. From Lemma 23 below, as $n \to \infty$,

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

where the last implication follows by DCT (since $1-\hat{\mu}_n(t) \to 1-\psi(t)$ for each t and also $|1-\hat{\mu}_n(t)| \leqslant 2$ 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 $n \geqslant N$. Now, find $A > 2/\delta$ such that for any $n \leqslant N$, 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. ■

Lemma 23

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

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

Proof. We write

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

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

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

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

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.

Example 26: Symmetric stable distributions

As $X \sim \text{Pois}(\lambda)$ has characteristic function $\exp\{\lambda(e^{it} - 1)\}$, it follows that uX has characteristic function $\exp\{\lambda(e^{iut} - 1)\}$. Adding independent copies of such variables, we see that $\exp\{\sum_{j=1}^{N} \lambda_j(e^{iu_j t} - 1)\}$ 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 $\exp\{\sum_{j=1}^{N} \lambda_j(2\cos(u_j t) - 2)\}$. 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 $I(bt) = b^{\alpha}I(t)$ for any $b > 0$. Therefore, $I(t) = C|t|^{\alpha}$ for $C = I(1)$. We have proved that $\exp\{-|t|^{\alpha}\}$ is a characteristic function for $0 < \alpha < 2$.

For $0 < \alpha < 2$, the distribution μ_α with characteristic function $\hat{\mu}_\alpha(t) = e^{-|t|^{\alpha}}$ is called the *symmetric α-stable distribution*. If we set $\alpha = 2$, we get the Gaussian distribution. But $e^{-|t|^{\alpha}}$ is not a characteristic function for $\alpha > 2$, as we shall see later and in the problem sets.

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 $m \geqslant 1$ and any complex numbers $c_1, \ldots, c_m$ and any real numbers $t_1, \ldots t_m$, we must have

$$0 \leqslant \int \Big|\sum_{k=1}^{m} c_k e^{it_k x}\Big|^2 d\mu(x) = \sum_{k,\ell=1}^{n} c_k \bar{c}_\ell \int e^{i(t_k - t_\ell)x} d\mu(x)$$

$$= \sum_{k,\ell=1}^{n} c_k \bar{c}_\ell \varphi(t_k - t_\ell).$$

This motivates the following definition.

Definition 18: Positive definite functions

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

Before going further, let us see why symmetric α-stable distributions do not exist for $\alpha > 2$.

Example 27

Suppose X is a random variable with characteristic function $e^{-|t|^\alpha}$. Fix $t > 0$ and let $Y = 2 - e^{-itX} - e^{-itX}$. Then

$$\begin{aligned}
\mathbb{E}[|Y|^2] &= 6 - 8e^{-t^\alpha} + 2e^{-(2t)^\alpha} \\
&= 6 - 8(1 - t^\alpha + O(t^{2\alpha})) + 2(1 - 2^p t^p + O(t^{2\alpha})) \qquad (\text{as } t \to 0) \\
&= 2(4 - 2^\alpha)t^p + O(t^{2\alpha})
\end{aligned}$$

Thus if $\alpha > 2$, then for small enough t we get $\mathbb{E}[|Y|^2] < 0$, which is impossible! Hence $e^{-|t|^\alpha}$ is not a characteristic function for $\alpha > 2$.

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

Theorem 45: Bochner's theorem

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

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

- If φ is positive definite, then $|\varphi| \leqslant 1$. Indeed, for any t, the positive semi-definiteness of $M_\varphi[0, t]$ shows that $1 - |\varphi(t)|^2 \geqslant 0$ (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 $a_{i,j} = \mathbb{E}[X_i X_j]$. Similarly $B = \mathbb{E}[Y_i Y_j]$ 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 $Z_i = X_i Y_i$, $1 \leqslant i \leqslant n$, is precisely C. Hence C is positive semi-definite.

- For any nice function $c : \mathbb{R} \mapsto \mathbb{C}$, we have

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

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

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

The first integral is positive by (2) (take $c(t) = e^{ixt}\mathbf{1}_{|t|\leqslant M}$). As for the second integral, we claim that it goes to zero as $M \to \infty$. Indeed, fix $\delta > 0$ and observe that for $|s| \leqslant (1-\delta)M$, the inner integral is less than $c_M := \int_{I_{\delta M^c}} |\varphi(u)|du$ (as $|t-s| \geqslant \delta M$ 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

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

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 $f(x) \geqslant 0$ for all x. We now claim that $\int f(x)dx = 1$. To start with, since $|f| \leqslant \|\varphi\|_1$, for

[5] 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

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

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

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 $\mathbb{E}[\varphi(\sigma Z)]$ 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 $f \geqslant 0$) to $f(x)$. 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 fis 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 $\varphi_\sigma(t) = \varphi(t)e^{-\sigma^2t^2/2}$ (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 $\hat{\mu}(t) = \int e^{i\langle t,x\rangle}d\mu(x) = \mathbb{E}[e^{i\langle t,X\rangle}]$. 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.

Here is an interesting fact that is totally non-trivial if we do not use characteristic functions (I don't know any such proof).

Proposition 4: Cramer-Wold device

Suppose $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have equal 1-dimensional marginals in all directions. Then $\mu = \nu$.

To be clear, what the equality of marginals means that if $X \sim \mu$ and $Y \sim \nu$ and $\langle X, v\rangle \stackrel{d}{=} \langle Y, v\rangle$ for each $v \in \mathbb{R}^d$. The conclusion is that $X \stackrel{d}{=} Y$. As we know very well, equality of 1-dimensional marginals in the co-ordinate (or any finite set of) directions is not enough to claim equality of joint distributions.

Proof. Since $\langle X, v\rangle \stackrel{d}{=} \langle Y, v\rangle$ for each v, we see that $\mathbb{E}[e^{i\langle X,v\rangle}] = \mathbb{E}[e^{i\langle Y,v\rangle}]$, hence the characteristic functions of X and Y coincide. Therefore, they have the same distribution on $\mathbb{R}^d$. ■

Remark 20

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.