Probability theory

Manjunath Krishnapur

DEPARTMENT OF MATHEMATICS, INDIAN INSTITUTE OF SCIENCE

2000 *Mathematics Subject Classification.* Primary

ABSTRACT. These are lecture notes from the spring 2010 Probability theory class at IISc. There are so many books on this topic that it is pointless to add any more, so these are not really a substitute for a good (or even bad) book, but a record of the lectures for quick reference. I have freely borrowed a lot of material from various sources, like Durrett, Rogers and Williams, Kallenberg, etc. Thanks to all students who pointed out many mistakes in the notes/lectures.

Contents

Chapter 1. Measure theory 1

1.1. Probability space 1

1.2. The ‘standard trick of measure theory’! 4

1.3. Lebesgue measure 6

1.4. Non-measurable sets 8

1.5. Random variables 9

1.6. Borel Probability measures on Euclidean spaces 10

1.7. Examples of probability measures on the line 11

1.8. A metric on the space of probability measures on $\mathbb{R}^d$ 12

1.9. Compact subsets of $\mathscr{P}(\mathbb{R}^d)$ 13

1.10. Absolute continuity and singularity 14

1.11. Expectation 16

1.12. Limit theorems for Expectation 17

1.13. Lebesgue integral versus Riemann integral 17

1.14. Lebesgue spaces: 18

1.15. Some inequalities for expectations 18

1.16. Change of variables 19

1.17. Distribution of the sum, product etc. 21

1.18. Mean, variance, moments 22

Chapter 2. Independent random variables 25

2.1. Product measures 25

2.2. Independence 26

2.3. Independent sequences of random variables 27

2.4. Some probability estimates 30

2.5. Applications of first and second moment methods 30

2.6. Weak law of large numbers 34

2.7. Applications of weak law of large numbers 35

2.8. Modes of convergence 37

2.9. Uniform integrability 38

2.10. Strong law of large numbers 40

2.11. Kolmogorov’s zero-one law 41

2.12. The law of iterated logarithm 42

2.13. Hoeffding’s inequality 43

2.14. Random series with independent terms 44

2.15. Kolmogorov’s maximal inequality 46

2.16. Central limit theorem - statement, heuristics and discussion 46

2.17. Central limit theorem - Proof using characteristic functions 48

2.18. CLT for triangular arrays 48

2.19. Limits of sums of random variables 50
2.20. Poisson convergence for rare events 51

Chapter 3. Brownian motion 53
3.1. Brownian motion and Winer measure 53
3.2. Some continuity properties of Brownian paths - Negative results 54
3.3. Some continuity properties of Brownian paths - Positive results 55
3.4. Lévy's construction of Brownian motion 56

Appendix A. Characteristic functions as tool for studying weak convergence 59
Defintions and basic properties 59
(A) Transformation rules 59
(B) Inversion formulas 60
(C) Continuity theorem 62

CHAPTER 1

Measure theory

1.1. Probability space

“Random experiment” was a non-mathematical term used to describe physical situations with more than one possible outcome, for instance, “toss a fair coin and observe the outcome”. In probability, although we sometimes use the same language, it is only as a quick substitute for a mathematically meaningful and precise phrasing. Consider the following examples.

(1) *“Draw a random integer from 1 to 100. What is the probability that you get a prime number?”* Mathematically, we just mean the following. Let $\Omega = \{1,2\ldots,100\}$, and for each $\omega \in \Omega$, we set $p_\omega = \frac{1}{100}$. Subsets $A \subset \Omega$ are called 'events' and for each subset we define $\mathbf{P}(A) = \sum_{\omega \in A} p_\omega$. In particular, for $A = \{2,3,5,\ldots,97\}$, we get $\mathbf{P}(A) = \frac{1}{4}$.

This is the setting for all of discrete probability. We have a finite or countable set Ω called sample space, and for each $\omega \in \Omega$ a number $p_\omega \geq 0$ is specified, so that $\sum_\omega p_\omega = 1$. For any $A \subset \Omega$, one defines its probability to be $\mathbf{P}(A) := \sum_{\omega \in A} p_\omega$. The whole game is to calculate probabilities of interesting events! The difficulty is of course that the set Ω and probabilities p_ω may be defined by a property which makes it hard to calculate probabilities.

Example 1.1. Fix $n \geq 1$ and let Ω be the set of all self-avoiding paths on length n in $\mathbb{Z}^2$ starting from $(0,0)$. That is,

$$\Omega = \{\omega = (\omega_0, \ldots, \omega_n) : \omega_i \in \mathbb{Z}^2, \omega_0 = (0,0), \omega_i - \omega_{i-1} \in \{\pm \mathbf{e}_1, \pm \mathbf{e}_2\}\}.$$

Then let $p_\omega = \frac{1}{\#\Omega}$. One interesting event is $A = \{\omega : \|\omega_n\| < n^{0.6}\}$. Far from finding $\mathbf{P}(A)$, it has not been proved to this day whether for large n, the value $P(A)$ is close to zero or one!

(2) *“Draw a number at random from the interval* $[0,1]$. *What is the probability that it is less than* $\frac{1}{2}$*? That it is rational? That its decimal expansion contains no 7?”* For the first question it seems that the answer must be $\frac{1}{2}$, but the next two questions motivate us to think more deeply about the meaning of such an assertion.

Like before we may set $\Omega = [0,1]$. What is p_ω? 'Intutively' it seems that the probability of the number falling in an interval $[a,b] \subset [0,1]$ should be $b-a$ and that forces us to set $p_\omega = 0$ for every ω. But then we cannot possibly get $\mathbf{P}([a,b])$ as $\sum_{\omega \in [a,b]} p_\omega$, even if such an uncountable sum had any meaning! So what is the basis for asserting that $\mathbf{P}[a,b] = b-a$?! Understanding this will be the first task.

A first attempt: Let us *define* the probability of any set $A \subset [0,1]$ to be the length of that set. We understand the length of an interval, but what is the length of the set of rational numbers? irrational numbers? A seemingly reasonable idea is to set

$$\mathbf{P}_*(A) = \inf\left\{\sum_k |I_k| \ : \ \text{each } I_k \text{ is an interval and } \{I_k\} \text{ a countable cover for } A\right\}.$$

Then perhaps, $\mathbf{P}_*(A)$ should be the probability of A for every subset $A \subset [0,1]$. This is at least reasonable in that $\mathbf{P}_*[a,b] = b - a$ for any $[a,b] \subset [0,1]$. But we face an unexpected problem. One can find[1] A such that $\mathbf{P}_*(A) = 1$ and $\mathbf{P}_*(A^c) = 1$ and that violates one of the basic requirements of probability, that $\mathbf{P}_*(A \cup A^c)$ be equal to $\mathbf{P}_*(A) + \mathbf{P}_*(A^c)$! You may object that our definition of $\mathbf{P}_*$ was arbitrary, may be another definition works? Before tackling that question, we should make precise what all we properties we require probabilities to satisfy. This we do next, but let us record here that there will be two sharp differences from discrete probability.

(a) One cannot hope to define $\mathbf{P}(A)$ for all $A \subset [0,1]$, but only for a rich enough class of subsets! These will be called events.

(b) One does not start with elementary probabilities p_ω and then compute $\mathbf{P}(A)$, but probabilities of all events are part of the specification of the probability space! (If all probabilities are specified at the outset, what does a probabilist do for a living? Hold that thought till the next lecture!).

Now we define the setting of probability in abstract and then return to the second situation above.

Definition 1.2. A probability space is a triple $(\Omega, \mathscr{F}, \mathbf{P})$ where

(1) The *sample space* Ω is an arbitrary set.

(2) The *σ-field* or *σ-algebra* $\mathscr{F}$ is a set of subsets of Ω such that (i) $\phi, \Omega \in \mathscr{F}$, (ii) if $A \in \mathscr{F}$, then $A^c \in \mathscr{F}$, (iii) if $A_n \in \mathscr{F}$ for $n = 1,2\ldots$, then $\cup A_n \in \mathscr{F}$. In words, $\mathscr{F}$ is closed under complementation and under countable unions, and contains the empty set. Elements of $\mathscr{F}$ are called *measurable sets*.

(3) The *probability measure* is any function $\mathbf{P} : \mathscr{F} \to [0,1]$ is such that if $A_n \in \mathscr{F}$ and are pairwise disjoint, then $\mathbf{P}(\cup A_n) = \sum \mathbf{P}(A_n)$ (countable additivity) and such that $\mathbf{P}(\Omega) = 1$. $\mathbf{P}(A)$ is called *the probability of A*.

Measurable sets are what we call *events* in probability theory. It is meaningless to ask for the probability of a subset of Ω that is not measurable. The σ-field is closed under many set operations and the usual rules of probability also hold. If one allows $\mathbf{P}$ to take values in $[0,\infty]$ and drops the condition $\mathbf{P}(\Omega) = 1$, then it is just called a *measure*. Measures have the same basic properties as probability measures, but probabilistically crucial concepts of *independence* and *conditional probabilities* (to come later) don't carry over to general measures and that is mainly what makes probability theory much richer than general measure theory.

Exercise 1.3. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space.

(1) $\mathscr{F}$ is closed under finite and countable unions, intersections, differences, symmetric differences. Also $\Omega \in \mathscr{F}$.

(2) If $A_n \in \mathscr{F}$, then $\limsup A_n := \{\omega \ : \ \omega \text{ belongs to infinitely many } A_n\}$ and $\liminf A_n := \{\omega \ : \ \omega \text{ belongs to all but finitely many } A_n\}$ are also in $\mathscr{F}$. In particular, if A_n increases or decreases to A, then $A \in \mathscr{F}$.

[1] Not obvious!

(3) $\mathbf{P}(\phi)=0$, $\mathbf{P}(\Omega)=1$. For any $A,B\in\mathscr{F}$ we have $\mathbf{P}(A\cup B)=\mathbf{P}(A)+\mathbf{P}(B)-\mathbf{P}(A\cap B)$. If $A_n\in\mathscr{F}$, then $\mathbf{P}(\cup A_n)\le\sum\mathbf{P}(A_n)$.

(4) If $A_n\in\mathscr{F}$ and A_n increases (decreases) to A, the $\mathbf{P}(A_n)$ increases (decreases) to $\mathbf{P}(A)$.

Some examples of probability spaces.

Example 1.4. Let Ω be a finite or countable set. Let $\mathscr{F}$ be the collection of all subsets of Ω. Then $\mathscr{F}$ is a σ-field. Given any numbers p_ω, $\omega\in\Omega$ that add to 1, we set $\mathbf{P}(A)=\sum_{\omega\in A}p_\omega$. Then $\mathbf{P}$ is a probability measure. More generally, let Ω be any set and let $R\subset\Omega$ be a countable set. Let $\mathscr{F}$ be the powerset of Ω. Fix nonnegative numbers p_x, $x\in R$ that add to 1. Then define $\mathbf{P}(A)=\sum_{x\in R\cap A}p_x$. This is a probability measure on $\mathscr{F}$.

This means that a discrete measure, say Binomial distribution, can be considered as a p.m. on $\{1,2,\dots,n\}$ or on $\mathbb{R}$. The problem of not being able to define probability for all subsets does not arise when the p.m. is so simple.

Example 1.5. Let Ω be an arbitrary set. Let $\mathscr{F}=\{A\subset\Omega:\ \text{either } A \text{ or } A^c \text{ is countable}\}$ (where 'countable' includes finite and empty sets). Define $\mathbf{P}(A)=0$ if A is countable and $\mathbf{P}(A)=1$ if A^c is countable. This is just a frivolous example of no particular importance.

Exercise 1.6. Check that $\mathscr{F}$ is a σ-field and that $\mathbf{P}$ is a probability measure on $\mathscr{F}$.

In the most interesting cases, one cannot explicitly say what the elements of $\mathscr{F}$ are, but only require that it is rich enough. Here is an exercise to introduce the important idea of a σ-field generated by a collection of sets.

Exercise 1.7. Let Ω be a set and let S be a collection of subsets of Ω. Show that there is a smallest sigma filed $\mathscr{F}$ containing all elements of S. That is, if $\mathscr{G}$ is any σ-field of subsets of Ω and $\mathscr{G}\supset S$, then $\mathscr{G}\supset\mathscr{F}$. $\mathscr{F}$ is called *the σ-field generated by S* and often denoted $\sigma(S)$.

Now we come to the most interesting probability spaces for Probability theory.

Example 1.8. Let $\Omega=[0,1]$. Let S be the collection of all intervals, to be precise let us take all right-closed, left-open intervals $(a,b]$, with $0\le a<b\le 1$ as well as intervals $[0,b]$, $b\le 1$. If we are trying to make precise the notion of '*drawing a number at random from* $[0,1]$', then we would want $\mathbf{P}(a,b]=b-a$ and $\mathbf{P}[0,b]=b$. The precise mathematical questions can now be formulated as follows. (i) Let $\mathscr{G}$ be the σ-field of all subsets of $[0,1]$. Is there a p.m. $\mathbf{P}$ on $\mathscr{G}$ such that $\mathbf{P}(a,b]=b-a$ and $\mathbf{P}[0,b]=b$ for all $0\le a<b\le 1$? If the answer is 'No', we ask for the less ambitious (ii) Is there a smaller σ-field large enough to contain all interval $(a,b]$, say $\mathscr{F}=\sigma(S)$ such that $\mathbf{P}(a,b]=b-a$?

The answer to the first question is 'No', which is why we need the notion of σ-fields, and the answer to the second question is 'Yes', which is why probabilists still have their jobs. Neither answer is obvious, but we shall answer them in coming lectures.

Example 1.9. Let $\Omega=\{0,1\}^{\mathbb{N}}=\{\omega=(\omega_1,\omega_2,\dots):\omega_i\in\{0,1\}\}$. Let S be the collection of all subsets of Ω that depend on only finitely many co-ordinates (such sets are called cylinders). More precisely, a cylinder set is of the form $A=\{\omega:\omega_{k_1}=\epsilon_1,\dots\omega_{k_n}=\epsilon_n\}$ for some given $n\ge 1$, $k_1<k_2<\ldots<k_n$ and $\epsilon_i\in\{0,1\}$ for $i\le n$.

What are we talking about? If we want to make precise the notion of '*toss a coin infinitely many times*', then clearly Ω is the sample space to look at. It is also desirable that elements of S be in the σ-field as we should be able to ask questions such as 'what is the chance that the fifth, seventh and thirtieth tosses are head, tail and head respectively' which is precisely asking for the probability of a cylinder set.

If we are '*tossing a coin with probability p of turning up Head*', then for a cylinder set $A = \{\omega : \omega_{k_1} = \epsilon_1, \ldots \omega_{k_n} = \epsilon_n\}$, it is clear that we would like to assign $\mathbf{P}(A) = \prod_{i=1}^{n} p_i^{\epsilon_i} q^{1-\epsilon_i}$ where $q = 1-p$. So the mathematical questions are: (i) If we take $\mathscr{F}$ to be the σ-field of all subsets of Ω, does there exist a p.m. $\mathbf{P}$ on $\mathscr{F}$ such that for cylinder sets $\mathbf{P}(A)$ is as previously specified. (ii) If the answer to (i) is 'No', is there a smaller σ-field, say the one generated by all cylinder sets and a p.m. $\mathbf{P}$ on it with probabilities as previously specified for cylinders?

Again, the answers are 'No' and 'Yes', respectively.

The σ-fields in these two examples can be captured under a common definition.

Definition 1.10. Let (X,d) be a metric space. The σ-field $\mathscr{B}$ generated by all open balls in X is called the Borel sigma-field of X.

First consider $[0,1]$ or $\mathbb{R}$. Let $S = \{(a,b]\} \cup \{[0,b]\}$ and let $T = \{(a,b)\} \cup \{[0,b)\} \cup \{(a,1]\}$. We could also simply write $S = \{(a,b]\cap[0,1] : a < b \in \mathbb{R}\}$ and $T = \{(a,b)\cap[0,1] : a < b \in \mathbb{R}\}$. Let the sigma-fields generated by S and T be denoted $\mathscr{F}$ (see example above) and $\mathscr{B}$ (Borel σ-field), respectively. Since

$$(a,b) = \cup_n (a, b-1/n], \qquad (a,b] = \cap_n (a, b+1/n], \qquad [0,b] = \cap_n$$

it is clear that $S \subset \mathscr{B}$ and $T \subset \mathscr{F}$. Hence $\mathscr{F} = \mathscr{B}$.

In the countable product space $\Omega = \{0,1\}^{\mathbb{N}}$ or more generally $\Omega = X^{\mathbb{N}}$, the topology is the one generated by all sets of the form $U_1 \times \ldots \times U_n \times X \times X \times \ldots$ where U_i are open sets in X. Clearly each of these sets is a cylinder set. Conversely, each cylinder set is an open set. Hence $\mathscr{G} = \mathscr{B}$. More generally, if $\Omega = X^{\mathbb{N}}$, then cylinders are sets of the form $A = \{\omega \in \Omega : \omega_{k_i} \in B_i, i \leq n\}$ for some $n \geq 1$ and $k_i \in \mathbb{N}$ and some Borel subsets B_i of X. It is easy to see that the σ-field generated by cylinder sets is exactly the Borel σ-field.

1.2. The 'standard trick of measure theory'!

While we care about sigma fields only, there are smaller sub-classes that are useful in elucidating the proofs. Here we define some of these.

Definition 1.11. Let S be a collection of subsets of Ω. We say that S is a

(1) π**-system** if $A,B \in S \implies A \cap B \in S$.
(2) λ**-system** if (i) $\Omega \in S$. (ii) $A,B \in S$ and $A \subseteq B \implies B \backslash A \in S$. (iii) $A_n \uparrow A$ and $A_n \in S \implies A \in S$.
(3) **Algebra** if (i) $\phi, \Omega \in S$. (ii) $A \in S \implies A^c \in S$. (iii) $A,B \in S \implies A \cup B \in S$.
(4) σ**-algebra** if (i) $\phi, \Omega \in S$. (ii) $A \in S \implies A^c \in S$. (iii) $A_n \in S \implies \cup A_n \in S$.

We have included the last one again for comparision. Note that the difference between algebras and σ-algebras is just that the latter is closed under countable unions while the former is closed only under finite unions. As with σ-algebras, arbitrary intersections of algebras/λ-systems/π-systems are again algebras/λ-systems/π-systems and hence one can talk of the algebra generated by a collection of subsets etc.

Example 1.12. The table below exhibits some examples.

Ω	S ($\pi-$system)	$\mathscr{A}(S)$ (algebra generated by S)	$\sigma(S)$
$(0,1]$	$\{(a,b] : 0<a\le b\le 1\}$	$\{\cup_{k=1}^{N}(a_k,b_k] : 0<a_1\le b_1\le a_2\le b_2\ldots\le b_N\le 1\}$	$\mathscr{B}(0,1]$
$[0,1]$	$\{(a,b]\cap[0,1] : a\le b\}$	$\{\cup_{k=1}^{N}R_k : R_k\in S$ are pairwise disjoint$\}$	$\mathscr{B}[0,1]$
$\mathbb{R}^d$	$\{\prod_{i=1}^{d}(a_i,b_i] : a_i\le b_i\}$	$\{\cup_{k=1}^{N}R_k : R_k\in S$ are pairwise disjoint$\}$	$\mathscr{B}_{\mathbb{R}^d}$
$\{0,1\}^{\mathbb{N}}$	collection of all cylinder sets	finite disjoint unions of cylinders	$\mathscr{B}(\{0,1\}^{\mathbb{N}})$

Often, as in these examples, sets in a π-system and in the algebra generated by the π-system can be described explicitly, but not so the sets in the generated σ-algebra.

Clearly, a σ-algebra is an algebra is a π-systemas well as a λ-system. The following converse will be useful. Plus, the proof exhibits a basic trick of measure theory!

Lemma 1.13 (Sierpinski-Dynkin $\pi-\lambda$ theorem)**.** *Let Ω be a set and let $\mathscr{F}$ be a set of subsets of Ω.*

(1) *$\mathscr{F}$ is a σ-algebra if and only if it is a π-system as well as a λ-system.*
(2) *If S is a π-system, then $\lambda(S)=\sigma(S)$.*

PROOF. (1) One way is clear. For the other way, suppose $\mathscr{F}$ is a π-system as well as a λ-system. Then, $\phi,\Omega\in\mathscr{F}$. If $A\in\mathscr{F}$, then $A^c=\Omega\backslash A\in\mathscr{F}$. If $A_n\in\mathscr{F}$, then the finite unions $B_n:=\cup_{k=1}^{n}A_k=\left(\cap_{k=1}^{n}A_k^c\right)^c$ belong to $\mathscr{F}$ as $\mathscr{F}$ is a π-system. The countable union $\cup A_n$ is the increasing limit of B_n and hence belongs to $\mathscr{F}$ by the λ-property.

(2) By part (i), it suffices to show that $\mathscr{F}:=\lambda(S)$ is a π-system, that is, we only need show that if $A,B\in\mathscr{F}$, then $A\cap B\in\mathscr{F}$. This is the tricky part of the proof!

Fix $A\in S$ and let $\mathscr{F}_A:=\{B\in\mathscr{F} : B\cap A\in\mathscr{F}\}$. S is a π-system, hence $\mathscr{F}_A\supset S$. We claim that $\mathscr{F}_A$ is a λ-system. Clearly, $\Omega\in\mathscr{F}_A$. If $B,C\in\mathscr{F}_A$ and $B\subset C$, then $(C\backslash B)\cap A=(C\cap A)\backslash(B\cap A)\in\mathscr{F}$ because $\mathscr{F}$ is a λ-system containing $C\cap A$ and $B\cap A$. Thus $(C\backslash B)\in\mathscr{F}_A$. Lastly, if $B_n\in\mathscr{F}_A$ and $B_n\uparrow B$, then $B_n\cap A\in\mathscr{F}_A$ and $B_n\cap A\uparrow B\cap A$. Thus $B\in\mathscr{F}_A$. This means that $\mathscr{F}_A$ is a λ-system containing S and hence $\mathscr{F}_A\supset\mathscr{F}$. In other words, $A\cap B\in\mathscr{F}$ for all $A\in S$ and all $B\in\mathscr{F}$.

Now fix any $A\in\mathscr{F}$. And again define $\mathscr{F}_A:=\{B\in\mathscr{F} : B\cap A\in\mathscr{F}\}$. *Because of what we have already shown*, $\mathscr{F}_A\supset S$. Show by the same arguments that $\mathscr{F}_A$ is a λ-system and conclude that $\mathscr{F}_A=\mathscr{F}$ for all $A\in\mathscr{F}$. This is another way of saying that $\mathscr{F}$ is a π-system. ■

As an application, we prove a certain uniqueness of extension of measures.

Lemma 1.14. *Let S be a π-system of subsets of Ω and let $\mathscr{F}=\sigma(S)$. If $\mathbf{P}$ and $\mathbf{Q}$ are two probability measures on $\mathscr{F}$ such that $\mathbf{P}(A)=\mathbf{Q}(A)$ for all $A\in S$, then $\mathbf{P}(A)=\mathbf{Q}(A)$ for all $A\in\mathscr{F}$.*

PROOF. Let $T=\{A\in\mathscr{F} : \mathbf{P}(A)=\mathbf{Q}(A)\}$. By the hypothesis $T\supset S$. We claim that T is a λ-system. Clearly, $\Omega\in T$. If $A,B\in T$ and $A\supset B$, then $\mathbf{P}(A\backslash B)=\mathbf{P}(A)-\mathbf{P}(B)=\mathbf{Q}(A)-\mathbf{Q}(B)=\mathbf{Q}(A\backslash B)$ implying that $A\backslash B\in T$. Lastly, if $A_n\in T$ and $A_n\uparrow A$, then $\mathbf{P}(A)=\lim_{n\to\infty}\mathbf{P}(A_n)=\lim_{n\to\infty}\mathbf{Q}(A_n)=\mathbf{Q}(A)$. Thus $T\supset\lambda(S)$ which is equal to $\sigma(S)$ by Dynkin's $\pi-\lambda$ theorem. Thus $\mathbf{P}=\mathbf{Q}$ on $\mathscr{F}$. ■

1.3. Lebesgue measure

Theorem 1.15. *There exists a unique Borel probability measure* $\mathbf{m}$ *on* $[0,1]$ *such that* $\mathbf{m}(I) = |I|$ *for any interval* I.

[**Sketch of the proof**] Note that $S = \{(a,b] \cap [0,1]\}$ is a π-system that generate $\mathscr{B}$. Therefore by Lemma 1.14, uniqueness follows. Existence is all we need to show. There are two steps.

Step 1 - Construction of the outer measure $\mathbf{m}_*$ Recall that we define $\mathbf{m}_*(A)$ for any subset by

$$\mathbf{m}_*(A) = \inf\left\{\sum_k |I_k| \,:\, \text{ each } I_k \text{ is an open interval and } \{I_k\} \text{ a countable cover for } A\right\}.$$

$\mathbf{m}_*$ has the following properties. (i) $\mathbf{m}_*$ is a $[0,1]$-valued function defined on all subsets $A \subset \Omega$. (ii) $\mathbf{m}_*(A \cup B) \le \mathbf{m}_*(A) + \mathbf{m}_*(B)$ for any $A, B \subset \Omega$. (iii) $\mathbf{m}_*(\Omega) = 1$.

These properties constitute the definition of an **outer measure**. In the case at hand, the last property follows from the following exercise.

Exercise 1.16. Show that $\mathbf{m}_*(a,b] = b - a$ if $0 < a \le b \le 1$.

Clearly, we also get countable subadditivity $\mathbf{m}_*(\cup A_n) \le \sum \mathbf{m}_*(A_n)$. The difference from a measure is that equality might not hold, even if the sets are pairwise disjoint.

Step-2 - The σ-field on which $\mathbf{m}_*$ is a measure

Let $\mathbf{m}_*$ be an outer measure on a set Ω. Then by restricting $\mathbf{m}_*$ to an appropriate σ-fields one gets a measure. We would also like this σ-field to be large (not the sigma algebra $\{\emptyset, \Omega\}$ please!).

Cartheodary's brilliant definition is to set

$$\mathscr{F} := \left\{A \subset \Omega \,:\, \mathbf{m}_*(E) = \mathbf{m}_*(A \cap E) + \mathbf{m}_*(A^c \cap E) \text{ for any } E\right\}.$$

Note that subadditivity implies $\mathbf{m}_*(E) \le \mathbf{m}_*(A \cap E) + \mathbf{m}_*(A^c \cap E)$ for any E for any A, E. The non-trivial inequality is the other way.

Theorem 1.17. *Then,* $\mathscr{F}$ *is a sigma algebra and* μ_* *restricted to* $\mathscr{F}$ *is a p.m.*

Proof. It is clear that $\emptyset, \Omega \in \mathscr{F}$ and $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$. Next, suppose $A, B \in \mathscr{F}$. Then for any E,

$$\mathbf{m}_*(E) = \mathbf{m}_*(E \cap A) + \mathbf{m}_*(E \cap A^c) = \mathbf{m}_*(E \cap A \cap B) + \{\mathbf{m}_*(E \cap A \cap B^c) + \mathbf{m}_*(E \cap A^c)\} \ge \mathbf{m}_*(E \cap A \cap B) + \mathbf{m}_*(E \cap (A \cap B)^c))$$

where the last inequality holds by subadditivity of $\mathbf{m}_*$ and $(E \cap A \cap B^c) \cup (E \cap A^c) = E \cap (A \cap B)^c$. Hence $\mathscr{F}$ is a π-system.

As $A \cup B = (A^c \cap B^c)^c$, it also follows that $\mathscr{F}$ is an algebra. For future use, note that $\mathbf{m}_*(A \cup B) = \mathbf{m}_*(A) + \mathbf{m}_*(B)$ if A, B are disjoint sets in $\mathscr{F}$. To see this apply the definition of $A \in \mathscr{F}$ with $E = A \cup B$.

It suffices to show that $\mathscr{F}$ is a λ-system. Suppose $A, B \in \mathscr{F}$ and $A \supset B$. Then

$$\mathbf{m}_*(E) = \mathbf{m}_*(E \cap B^c) + \mathbf{m}_*(E \cap B) = \mathbf{m}_*(E \cap B^c \cap A) + \mathbf{m}_*(E \cap B^c \cap A^c) + \mathbf{m}_*(E \cap B) \ge \mathbf{m}_*(E \cap (A \setminus B)) + \mathbf{m}_*(E \cap (A \setminus B)^c).$$

Before showing closure under increasing limits, Next suppose $A_n \in \mathscr{F}$ and $A_n \uparrow A$. Then $\mathbf{m}_*(A) \ge \mathbf{m}_*(A_n) = \sum_{k=1}^n \mathbf{m}_*(A_k \setminus A_{k-1})$ by finite additivity of $\mathbf{m}_*$. Hence

$\mathbf{m}_*(A) \geq \sum \mathbf{m}_*(A_k \backslash A_{k-1})$. The other way inequality follows by subadditivity of $\mathbf{m}_*$ and we get $\mathbf{m}_*(A) = \sum \mathbf{m}_*(A_k \backslash A_{k-1})$. Then for any E we get

$$\mathbf{m}_*(E) = \mathbf{m}_*(E \cap A_n) + \mathbf{m}_*(E \cap A_n^c) \geq \mathbf{m}_*(E \cap A_n) + \mathbf{m}_*(E \cap A^c) = \sum_{k=1}^{n} \mathbf{m}_*(E \cap (A_k \backslash A_{k-1})) + \mathbf{m}_*(E \cap A^c).$$

The last equality follows by finite additivity of $\mathbf{m}_*$ on $\mathscr{F}$. Let $n \to \infty$ and use subadditivity to see that

$$\mathbf{m}_*(E) \geq \sum_{k=1}^{\infty} \mathbf{m}_*(E \cap (A_k \backslash A_{k-1})) + \mathbf{m}_*(E \cap A^c) \geq \mathbf{m}_*(E \cap A) + \mathbf{m}_*(E \cap A^c).$$

Thus, $A \in \mathscr{F}$ and it follows that $\mathscr{F}$ is a λ-system too and hence a σ-algebra.

Lastly, if $A_n \in \mathscr{F}$ are pairwise disjoint with union A, then $\mathbf{m}_*(A) \geq \mathbf{m}_*(A_n) = \sum_{k=1}^{n} \mathbf{m}_*(A_k) \to \sum_k \mathbf{m}_*(A_k)$ while the other way inequality follows by subadditivity of $\mathbf{m}_*$ and we see that $\mathbf{m}_*|_{\mathscr{F}}$ is a measure.

Step-3 - $\mathscr{F}$ is large enough!

Let $A = (a,b]$. For any $E \subset [0,1]$, let $\{I_n\}$ be an open cover such that $\mathbf{m}_*(E) \geq \sum |I_n|$. Then, note that $\{I_n \cap (a,b)\}$ and $\{I_n \cap [a,b]^c\}$ are open covers for $A \cap E$ and $A^c \cap E$, respectively ($I_n \cap [a,b]^c$ may be a union of two intervals, but that does not change anything essential). It is also clear that $|I_n| = |I_n \cap (a,b)| + |I_n \cap (a,b)^c|$. Hence we get

$$\mathbf{m}_*(E) \geq \sum |I_n \cap (a,b)| + \sum |I_n \cap (a,b)^c| \geq \mathbf{m}_*(A \cap E) + \mathbf{m}_*(A^c \cap E).$$

The other inequality follows by subadditivity and we see that $A \in \mathscr{F}$. Since the intervals $(a,b]$ generate $\mathscr{B}$, and $\mathscr{F}$ is a sigma algebra, we get $\mathscr{F} \supset \mathscr{B}$. Thus, restricted to $\mathscr{B}$ also, $\mathbf{m}_*$ gives a p.m. ■

Remark 1.18. (1) We got a σ-algebra $\mathscr{F}$ that is larger than $\mathscr{B}$. Two natural questions. Does $\mathscr{F}$ or $\mathscr{B}$ contain all subsets of $[0,1]$? Is $\mathscr{F}$ strictly larger than $\mathscr{B}$? We show that $\mathscr{F}$ does not contain all subsets. One of the homework problems deals with the relationship between $\mathscr{B}$ and $\mathscr{F}$.

(2) $\mathbf{m}$, called the Lebesgue measure on $[0,1]$, is the only probability space one ever needs. In fact, all probabilities ever calculated can be seen, in principle, as calculating the Lebsgue measure of some Borel subset of $[0,1]$!

Generalities The construction of Lebesgue measure can be made into a general procedure for constructing interesting measures, starting from measures of some rich enough class of sets. The steps are as follows.

(1) Given an algebra $\mathscr{A}$ (in this case finite unions of $(a,b]$), and a *countably additive p.m* $\mathbf{P}$ on $\mathscr{A}$, define an outer measure $\mathbf{P}_*$ on all subsets by taking infimum over countable covers by sets in $\mathscr{A}$.
(2) Then define $\mathscr{F}$ exactly as above, and prove that $\mathscr{F} \supset \mathscr{A}$ is a σ-algebra and $\mathbf{P}_*$ is a p.m. on $\mathscr{A}$.
(3) Show that $\mathbf{P}_* = \mathbf{P}$ on $\mathscr{A}$.

Proofs are quite the same. Except, in $[0,1]$ we started with $\mathbf{m}$ defined on a π-system S rather than an algebra. But in this case the generated algebra consists precisely of disjoint unions of sets in S, and hence we knew how to define $\mathbf{m}$ on $\mathscr{A}(S)$. When can we start with $\mathbf{P}$ defined ona π-system? The crucial point in $[0,1]$ was that for any $A \in S$, one can write A^c as a finite union of sets in S. In such cases (which includes examples from the previous lecture) the generated algebra is precisely the set of disjoint finite unions of sets in S and we define $\mathbf{P}$ on $\mathscr{A}(S)$ and then proceed to step one above.

Exercise 1.19. Use the general procedure as described here, to construct the following measures.

(a) A p.m. on $([0,1]^d, \mathscr{B})$ such that $\mathbf{P}([a_1,b_1] \times \ldots \times [a_d,b_d]) = \prod_{k=1}^{d}(b_k - a_k)$ for all cubes contained in $[0,1]^d$. This is the d-dimensional Lebesgue measure.

(b) A p.m. on $\{0,1\}^{\mathbb{N}}$ such that for any cylinder set $A = \{\omega : \omega_{k_j} = \epsilon_j,\ j = 1,\ldots,n\}$ (any $n \geq 1$ and $k_j \in \mathbb{N}$ and $\epsilon_j \in \{0,1\}$) we have (for a fixed $p \in [0,1]$ and $q = 1-p$)

$$\mathbf{P}(A) = \prod_{j=1}^{n} p^{\epsilon_j} q^{1-\epsilon_j}.$$

[**Hint:** Start with the algebra generated by cylinder sets].

1.4. Non-measurable sets

We have not yet shown the necessity for σ-fields. Restrict attention to $([0,1], \mathscr{F}, \mathbf{m})$ where $\mathscr{F}$ is either (i) $\mathscr{B}$, the Borel σ-algebra or (ii) $\overline{\mathscr{B}}$ the possibly larger σ-algebra of Lebesgue measurable sets (as defined by Caratheodary). This consists of two distinct issues.

1. Showing that $\overline{\mathscr{B}}$ (hence $\mathscr{B}$) does not contain all subsets of $[0,1]$.
2. Showing that it is not possible at all to define a p.m. $\mathbf{P}$ on the σ-field of all subsets so that $\mathbf{P}[a,b] = b-a$ for all $0 \leq a \leq b \leq 1$. In other words, one cannot consistently extend $\mathbf{m}$ from $\overline{\mathscr{B}}$ (on which it is uniquely determined by the condition $\mathbf{m}[a,b] = b-a$) to a p.m. $\mathbf{P}$ on the σ-algebra of all subsets.

(1) $\overline{\mathscr{B}}$ **does not contain all subsets of** $[0,1]$**:** We shall need the following 'translation invariance property' of $\mathbf{m}$ on $\overline{\mathscr{B}}$.

Exercise 1.20. For any $A \subset [0,1]$ and any $x \in [0,1]$, $\mathbf{m}(A+x) = \mathbf{m}(A)$, where $A + x := \{y + x (\text{mod } 1) : y \in A\}$ (eg: $[0.4, 0.9] + 0.2 = [0, 0.1] \cup [0.6, 1]$). Show that for any $A \in \overline{\mathscr{B}}$ and $x \in [0,1]$ that $A + x \in \overline{\mathscr{B}}$ and that $\mathbf{m}(A+x) = \mathbf{m}(A)$.

Now we construct a subset $A \subset [0,1]$ and countably (infinitely) many $x_k \in [0,1]$ such that the sets $A + x_k$ are pairwise disjoint and $\cup_k (A + x_k)$ is the whole of $[0,1]$. Then, if A were in $\overline{\mathscr{B}}$, by the exercise $A + x_k$ would have the same probability as A. But $\sum \mathbf{m}(A + x_k)$ must be equal to $\mathbf{m}[0,1] = 1$, which is impossible! Hence $A \notin \overline{\mathscr{B}}$.

How to construct such a set A and $\{x_k\}$? Define an equivalence relation on $[0,1]$ by $x \sim y$ if $x - y \in \mathbb{Q}$ (check that this is indeed an equivalence relation). Then, $[0,1]$ splits into pairwise disjoint equivalence classes whose union is the whole of $[0,1]$.

Invoke *axiom of choice* to get a set A that has exactly one point from each equivalence class. Consider $A + r$, $r \in \mathbb{Q} \cap [0,1)$. If $A + r$ and $A + s$ intersect then we get an $x \in [0,1]$ such that $x = y + r = z + s$ (mod 1) for some $y, z \in A$. This implies that $y - z = r - s$ (mod 1) and hence that $y \sim z$. So we must have $y = z$ (as A has only one element from each equivalence class) and that forces $r = s$ (why?). Thus $A + r$, $r \in \mathbb{Q} \cap [0,1)$ are pairwise disjoint. Further given $x \in [0,1]$, there is a $y \in A$ belonging to the $[[x]]$. Therefore $x \in A + r$ where $r = y - x$ or $y - x + 1$. Thus we have constructed the set A whose countably many translates $A + r$, $r \in \mathbb{Q} \cap [0,1)$ are pairwise disjoint and exhaustive! This answers question (1).

Remark 1.21. There is a theorem to the effect that the axiom of choice is necessary to show the existence of a non-measurable set (as an aside, we should perhaps not have used the word 'construct' given that we invoke the axiom of choice).

(2) **m does not extend to all subsets:** The proof above shows in fact that **m** cannot be extended to a *translation invariant* p.m. on all subsets. If we do not require translation invariance for the extended measure, the question becomes more difficult.

Note that there do exist probability measures on the σ-algebra of all subsets of $[0,1]$, so one cannot say that there are no measures on all subsets. For example, define $\mathbf{Q}(A) = 1$ if $0.4 \in A$ and $\mathbf{Q}(A) = 0$ otherwise. Then $\mathbf{Q}$ is a p.m. on the space of all subsets of $[0,1]$. $\mathbf{Q}$ is a discrete p.m. in hiding! If we exclude such measures, then it is true that some subsets have to be omitted to define a p.m. You may find the proof for the following general theorem in Billingsley, p. 46 (uses *axiom of choice* and *continuum hypothesis*).

Fact 1.22. There is no p.m. on the σ-algebra of all subsets of $[0,1]$ that gives zero probability to singletons.

Say that x is an atom of $\mathbf{P}$ if $\mathbf{P}(\{x\}) > 0$ and that $\mathbf{P}$ is purely atomic if $\sum_{\text{atoms}} \mathbf{P}(\{x\}) = 1$. The above fact says that if $\mathbf{P}$ is defined on the σ-algebra of all subsets of $[0,1]$, then $\mathbf{P}$ must be have atoms. It is not hard to see that in fact $\mathbf{P}$ must be purely atomic. To see this let $\mathbf{Q}(A) = \mathbf{P}(A) - \sum_{x \in A} \mathbf{P}(\{x\})$. Then $\mathbf{Q}$ is a non-negative measure without atoms. If $\mathbf{Q}$ is not identically zero, then with $c = \mathbf{Q}([0,1])^{-1}$, we see that $c\mathbf{Q}$ is a p.m. without atoms, and defined on all subsets of $[0,1]$, contradicting the stated fact.

Remark 1.23. This last manipulation is often useful and shows that we can write any probability measure as a convex combination of a purely atomic p.m. and a completely nonatomic p.m.

(3) **Finitely additive measures** If we relax countable additivity, strange things happen. For example, there does exist a *translation invariant* ($\mu(A+x) = \mu(A)$ for all $A \subset [0,1]$, $x \in [0,1]$, in particular, $\mu(I) = |I|$) *finitely additive* ($\mu(A \cup B) = \mu(A) + \mu(B)$ for all A, B disjoint) p.m. defined on all subsets of $[0,1]$! In higher dimensions, even this fails, as shown by the mind-boggling

Banach-Tarski "paradox": The unit ball in $\mathbb{R}^3$ can be divided into finitely many (five, in fact) disjoint pieces and rearranged (only translating and rotating each piece) into a ball of twice the original radius!!

1.5. Random variables

Definition 1.24. Let $(\Omega_i, \mathscr{F}_i, \mathbf{P}_i)$, $i = 1,2$, be two probability spaces. A function $T : \Omega_1 \to \Omega_2$ is called an Ω_2-valued random variable if $T^{-1}A \in \mathscr{F}_1$ for any $A \in \mathscr{F}_2$. Here $T^{-1}(A) := \{\omega \in \Omega_1 : T(\omega) \in A\}$ for any $A \subset \Omega_2$.

Important cases are when $\Omega_2 = \mathbb{R}$ and $\mathscr{F}_2 = \mathscr{B}(\mathbb{R})$ (we just say "random variable") or $\Omega_2 = \mathbb{R}^d$ and $\mathscr{F}_2 = \mathscr{B}(\mathbb{R}^d)$ ("random vector"). When $\Omega_2 = C[0,1]$ with $\mathscr{F}_2$ its Borel sigma algebra (under the sup-norm metric), T is called a "stochastic process". When Ω_2 is itself the space of all locally finite countable subsets of $\mathbb{R}^d$ (with Borel sigma algebra in an appropriate metric) , we call T a "point process". In genetics or population biology one looks at genealogies, and then we have tree-valued random variables etc. etc.

Remark 1.25. Some remarks.

(1) If $T : \Omega_1 \to \Omega_2$ is any function, then given a σ-algebra $\mathscr{G}$ on Ω_2, the "pull-back" $\{T^{-1}A \ : \ A \in \mathscr{G}\}$ is the smallest σ-algebra on Ω_1 w.r.t. which T is measurable (if we fix $\mathscr{G}$ on Ω_2) . Conversely, given a σ-algebra $\mathscr{F}$ on Ω_1, the "push-forward" $\{A \subset \Omega_2 \ : \ T^{-1}A \in \mathscr{F}\}$ is the largest σ-algebra on Ω_2 w.r.t. which T is measurable (if we fix $\mathscr{F}$ on Ω_1). These properties are simple consequences of the fact that $T^{-1}(A)^c = T^{-1}(A^c)$ and $T^{-1}(\cup A_n) = \cup_n T^{-1}(A_n)$.

(2) If S generates $\mathscr{F}_2$, i.e., $\sigma(S) = \mathscr{F}_2$, then it suffices to check that $T^{-1}A \in \mathscr{F}_1$ for any $A \in S$.

Example 1.26. Consider $([0,1],\mathscr{B})$. Any continuous function $T : [0,1] \to \mathbb{R}$ is a random variable. This is because $T^{-1}(\text{open}) = \text{open}$ and open sets generate $\mathscr{B}(\mathbb{R})$. **Exercise:** Show that T is measurable if it is any of the following. (a) Lower semicontinuous, (b) Right continuous, (c) Non-decreasing, (d) Linear combination of measurable functions, (e) $\limsup$ of a countable sequence of measurable functions. (a) supremum of a countable family of measurable functions.

Push forward of a measure: If $T : \Omega_1 \to \Omega_2$ is a random variable, and $\mathbf{P}$ is a p.m. on $(\Omega_1,\mathscr{F}_1)$, then defining $\mathbf{Q}(A) = \mathbf{P}(T^{-1}A)$, we get a p.m $\mathbf{Q}$, on $(\Omega_2,\mathscr{F}_2)$. $\mathbf{Q}$, often denoted $\mathbf{P}T^{-1}$ is called the push-forward of $\mathbf{P}$ under T.

The reason why $\mathbf{Q}$ is a measure is that if A_n are pairwise disjoint, then $T^{-1}A_n$ are pairwise disjoint. However, note that if B_n are pairwise disjoint in Ω_1, then $T(B_n)$ are in general not disjoint. This is why there is no "pull-back measure" in general (unless T is one-one, in which case the pull-back is just the push-forward under T^{-1}!)

When $(\Omega_2,\mathscr{F}_2) = (\mathbb{R},\mathscr{B})$, the push forward (a Borel p.m on $\mathbb{R}$) is called the *distribution* of the r.v. T. If $T = (T_1,\dots,T_d)$ is a random vector, then the pushforward, a Borel p.m. on $\mathbb{R}^d$ is called the distribution of T or as the *joint distribution* of $T_1,\dots,T_d$.

1.6. Borel Probability measures on Euclidean spaces

Given a Borel p.m. μ on $\mathbb{R}^d$, we define its cumulative distribution functions (CDF) to be $F_\mu(x_1,\dots,x_d) = \mu((-\infty,x_1] \times \dots \times (-\infty,x_d])$. Then, by basic properties of probability measures, $F_\mu : \mathbb{R}^d \to [0,1]$ (i) is non-decreasing in each co-ordinate, (ii) $F_\mu(x) \to 0$ if $\max_i x_i \to -\infty$, $F_\mu(x) \to 1$ if $\min_i x_i \to +\infty$, (iii) F_μ is right continuous in each co-ordinate.

Two natural questions. Given an $F : \mathbb{R}^d \to [0,1]$ satisfying (i)-(iii), is there necessarily a Borel p.m. with F as its CDF? If yes, is it unique?

If μ and ν both have CDF F, then for any rectangle $R = (a_1,b_1] \times \dots \times (a_d,b_d]$, $\mu(R) = \nu(R)$ because they are both determined by F. Since these rectangles form a π-system that generate the Borel σ-algebra, $\mu = \nu$ on $\mathscr{B}$.

What about existence of a p.m. with CDF equal to F? For simplicity take $d = 1$. One boring way is to define $\mu(a,b] = F(b) - F(a)$ and then go through Caratheodary construction. But all the hard work has been done in construction of Lebesgue measure, so no need to repeat it!

Consider the probability space $((0,1),\mathscr{B},m)$ and define the function $T : (0,1) \to \mathbb{R}$ by $T(u) := \inf\{x \ : \ F(x) \geq u\}$. When F is strictly increasing and continuous, T is just the inverse of F. In general, T is non-decreasing, left continuous. Most importantly,

$T(u) \le x$ if and only if $F(x) \ge u$. Let $\mu := m\ T^{-1}$ be the push-forward of the Lebesgue measure under T. Then,

$$\mu(-\infty,x] = m\{u\ :\ T(u) \le x\} = m\{u\ :\ F(x) \ge u\} = m(0,F(x)] = F(x).$$

Thus, we have produced a p.m. μ with CDF equal to F. Thus p.m.s on the line are in bijective correspondence with functions satisfying (i)-(iii). Distribution functions (CDFs) are a useful but dispensable tool to study measures on the line, because we have better intuition in working with functions than with measures.

Exercise 1.27. Do the same for Borel probability measures on $\mathbb{R}^d$.

1.7. Examples of probability measures on the line

There are many important probability measures that occur frequently in probability and in the real world. We give some examples below and expect you to familiarize yourself with each of them.

Example 1.28. The examples below have CDFs of the form $F(x) = \int_{-\infty}^{x} f(t)dt$ where f is a non-negative integrable function with $\int f = 1$. In such cases f is called the *density* or pdf (probability density function). Clearly F is continuous and non-decreasing and tends to 0 and 1 at ∞ and $-\infty$ respectively. Hence, there do exist probability measures on $\mathbb{R}$ with the corresponding density.

(1) *Normal distribution.* For fixed $a \in \mathbb{R}$ and $\sigma^2 > 0$, $N(a,\sigma^2)$ is the p.m. on $\mathbb{R}$ with density $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/2\sigma^2} du$. F is clearly increasing and continuous and $F(-\infty) = 0$. That $F(+\infty) = 1$ is not so obvious but true!

(2) *Gamma distribution* with shape parameter $\alpha > -1$ and scale parameter $\lambda > 0$ is the p.m. with density $f(x) = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$.

(3) *Exponential distribution.* Exponential(λ) is the p.m. with density $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$ and $f(x) = 0$ if $x < 0$. This is a special case of Gamma distribution, but important enough to have its own name.

(4) *Beta distribution.* For parameters $a > -1$, $b > -1$, the Beta(a,b) distribution is the p.m. with density $B(a,b)^{-1} x^{a-1}(1-x)^{b-1}$ for $x \in [0,1]$. Here $B(a,b)$ is the beta function, equal to $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$. (Why does it integrate to 1?).

(5) *Uniform distribution* on $[a,b]$ is the p.m. with density $f(x) = \frac{1}{b-a}$ for $x \in [a,b]$. For example, with $a = 0, b = 1$, this is a special case of the Beta distribution.

(6) *Cauchy distribution.* This is the p.m. with density $\frac{1}{\pi(1+x^2)}$ on the whole line. Unlike all the previous examples, this distribution has "heavy tails"

You may have seen the following discrete probability measures. They are very important too and will recur often.

Example 1.29. The examples below have CDFs of the form $F(x) = \sum_{u_i \le x} p(x_i)dt$, where $\{x_i\}$ is a fixed countable set, and $p(x_i)$ are non-negative numbers that add to one. In such cases p is called the pmf (probability density function). and from what we have shown, there do exist probability measures on $\mathbb{R}$ with the corresponding density or CDF.

(1) *Binomial distribution.* Binomial(n,p), with $n \in \mathbb{N}$ and $p \in [0,1]$, has the pmf $p(k) = \binom{n}{k} p^k q^{n-k}$ for $k = 0,1,\dots,n$.

(2) *Bernoulli distribution.* $p(1) = p$ and $p(0) = 1-p$ for some $p \in [0,1]$. Same as Binomail($1,p$).

(3) *Poisson(λ)* distribution with parameter $\lambda \geq 0$ has p.m.f $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$.

(4) *Geometric(p)* distribution with parameter $p \in [0,1]$ has p.m.f $p(k) = q^k p$ for $k = 0, 1, 2, \ldots$.

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

What kind of space is $\mathscr{P}(\mathbb{R}^d)$ (the space of p.m.s on $\mathbb{R}^d$)? It is clearly a convex set (this is true for p.m.s on any sample space and σ-algebra).

We saw that for every Borel p.m. on $\mathbb{R}^d$ there is associated a unique CDF. This suggests a way of defining a distance function on $\mathscr{P}(\mathbb{R}^d)$ using their CDFs. Let $D(\mu, \nu) = \sup_{x \in \mathbb{R}^d} |F_\mu(x) - F_\nu(x)|$. Since CDFs are bounded between 0 and 1, this is well-defined and one can easily check that it gives a metric on $\mathscr{P}(\mathbb{R}^d)$.

Is this the metric we want to live with? For $a \in \mathbb{R}^d$, we denote by δ_a the p.m. for which $\delta_a(A) = 1$ if $A \ni a$ and 0 otherwise (although this p.m. can be defined on all subsets, we just look at it as a Borel measure). If $a \neq b$, it is easy to see that $D(\delta_a, \delta_b) = 1$. Thus, even when $a_n \to a$ in $\mathbb{R}^d$, we do not get convergence of δ_{a_n} to δ_a. This is an undesirable feature and hence we would like a weaker metric.

Definition 1.30. For $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, define the Lévy distance between them as (here $\mathbf{1} = (1, 1, \ldots, 1)$)

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

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

First of all, $d(\mu, \nu) \leq 1$. That d is indeed a metric is an easy exercise. If $a_n \to a$ in $\mathbb{R}^d$, does δ_{a_n} converge to δ_a? Indeed $d(\delta_a, \delta_b) = (\max_i |b_i - a_i|) \wedge 1$ and hence $d(\delta_{a_n}, a) \to 0$.

Exercise 1.31. Let $\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{k/n}$. Show directly by definition that $d(\mu_n, m) \to 0$. What about $D(\mu_n, \mu)$?

How does convergence in the metric d show in terms of CDFs?

Proposition 1.32. *$\mu_n \overset{d}{\to} \mu$ if and only if $F_{\mu_n}(x) \to F_\mu(x)$ for all continuity points x of F_μ.*

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

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

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

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

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

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

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

1.9. Compact subsets of $\mathscr{P}(\mathbb{R}^d)$

Often we face problems like the following. A functional $L : \mathscr{P}(\mathbb{R}^d) \to \mathbb{R}$ is given, and we would like to find the p.m. μ that minimizes $L(\mu)$. By definition, we can find nearly optimal p.m.s μ_n satisfying $L(\mu_n) - \frac{1}{n} \le \inf_\nu L(\nu)$. Then we might expect that if some subsequence μ_{n_k} converged to a p.m. μ, then that μ might be the optimal solution we are searching for. Thus we are faced with the question of characterizing compact subsets of $\mathscr{P}(\mathbb{R}^d)$, so that existence of convergent subsequences can be asserted.

Looking for a convergent subsequence: Let μ_n be a sequence in $\mathscr{P}(\mathbb{R}^d)$. We would like to see if a convergent subsequence can be extracted. Write F_n for F_{μ_n}. For any fixed $x \in \mathbb{R}^d$, $F_n(x)$ is a bounded sequence of reals and hence we can find a subsequence $\{n_k\}$ such that $F_{n_k}(x)$ converges.

Fix a dense subset $S = \{x_1, x_2, \dots\}$ of $\mathbb{R}^d$. Then, by the observation above, we can find a subsequence $\{n_{1,k}\}_k$ such that $F_{n_{1,k}}(x_1)$ converges to some number in $[0,1]$ that we shall denote $G(x_1)$. Then extract a further subsequence $\{n_{2,k}\}_k \subset \{n_{1,k}\}_k$ such that $F_{n_{2,k}}(x_2) \to G(x_2)$, another number in $[0,1]$. Of course, we also have $F_{n_{2,k}}(x_1) \to G(x_1)$. Continuing this way, we get subsequences $\{n_{1,k}\} \supset \{n_{2,k}\} \supset \dots \{n_{\ell,k}\} \dots$ such that for each ℓ, as $k \to \infty$, we have $F_{n_{\ell,j}}(x_j) \to G(x_j)$ for each $j \le \ell$.

The *diagonal sbsequence* $\{n_{\ell,\ell}\}$ is ultimately the subsequence of each of the above obtained subsequences and therefore, $F_{n_{\ell,\ell}}(x_j) \to G(x_j)$ for all j.

To define the limiting function on the whole line, set $F(x) := \inf\{G(x_j) : j \text{ for which } x_j > x\}$. F is well defined, takes values in $[0,1]$ and is non-decreasing. It is also right-continuous, because if $y_n \downarrow y$, then for any j for which $x_j > y$, it is also true that $x_j > y_n$ for sufficiently large n. Thus $\liminf_{n\to\infty} G(y_n) \le \inf\limits_{x_j > y} G(x_j) = F(y)$. Lastly, if y is any continuity point of F, then for any $\delta > 0$, we can find i, j such that $y - \delta < x_i < y < x_j < y + \delta$. Therefore

$$F(y-\delta) \le G(x_i) = \lim F_{n_{\ell,\ell}}(x_i) \le \liminf F_{n_{\ell,\ell}}(y) \le \limsup F_{n_{\ell,\ell}}(y) \le \lim F_{n_{\ell,\ell}}(x_j) = G(x_j) \le F(y+\delta).$$

The equalities are by prperty of the subsequence $\{n_{\ell,\ell}\}$, the inner two inequalities are obvious, and the outer two inequalities follow from the definition of F in terms of G (and the fact that G is nondecreasing). Since F is continuous at y, we get $\lim F_{n_{\ell,\ell}}(y) = F(y)$.

If only we could show that $F(+\infty) = 1$ and $F(-\infty) = 0$, then F would be the CDF of some p.m. μ and we would immediately get $\mu_n \xrightarrow{d} \mu$. But this is false in general!

Example 1.33. 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 p.m. 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 1.34. A family $\{\mu_\alpha\}_{\alpha \in I} \subset \mathscr{P}(\mathbb{R}^d)$ is said to be *tight* if for any $\epsilon > 0$, there is a compact set $K_\epsilon \subset \mathbb{R}^d$ such that $\mu_\alpha(K_\epsilon) \ge 1 - \epsilon$ for all $\alpha \in I$.

Example 1.35. Suppose the family has only one p.m. μ. Since $[-n,n]^d$ increase to $\mathbb{R}^d$, given $\epsilon > 0$, for a large enough n, we have $\mu([-n,n]^d) \geq 1-\epsilon$. Hence $\{\mu\}$ is tight. If the family is finite, tightness is again clear.

Take $d = 1$ and let μ_n be p.m.s 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.

Theorem 1.36 (Helly's selection principle). *(a) A sequence of probability measures on $\mathbb{R}^d$ is tight if and only if every subsequence has a further subsequence that converges weakly. (b) Equivalently a subset of $\mathscr{P}(\mathbb{R}^d)$ is precompact if and only if it is tight.*

PROOF. (a) If μ_n is a tight sequence in $\mathscr{P}(\mathbb{R}^d)$, then any subsequence is also tight. By the earlier discussion, given any subsequence $\{n_k\}$, we may extract a further subsequence $n_{\ell,k}$ and find a non-decreasing right continuous function F (taking values in $[0,1]$) such that $F_{n_{\ell,k}}(x) \to F(x)$ for all continuity points x of F. Fix $A > 0$ such that $\mu_n[-A,A] \geq 1-\epsilon$ and such that A is a continuity point of F. Then $F(A) = \lim_{k\to\infty} F_{n_{\ell,k}}(A) \geq 1-\epsilon$. Thus $F(+\infty) = 1$. Similarly one can show that $F(-\infty) = 0$. This shows that $F = F_\mu$ for some $\mu \in \mathscr{P}(\mathbb{R}^d)$ and thus $\mu_{n_{\ell,k}} \overset{d}{\to} \mu$ as $k \to \infty$.

Conversely, if the sequence $\{\mu_n\}$ is not tight, then for any $A > 0$, we can find an infinite sequence n_k such that $\mu_{n_k}(-A,A) < 1-\epsilon$ (why?). Then, either $F_{n_k}(A) < 1-\frac{\epsilon}{2}$ for infinitely many k or $F_{n_k}(-A) < \frac{\epsilon}{2}$. Thus, for any $A > 0$, we have $F(A) < 1-\frac{\epsilon}{2}$ or $F(-A) < \frac{\epsilon}{2}$. Thus F is not a CDF of a p.m., and we see that the subsequence $\{n_k\}$ has no further subsequence than can converge to a probability measure.

(b) Standard facts about convergence in metric spaces and part (a).

■

1.10. Absolute continuity and singularity

How wild can the jumps of a CDF be? If μ is a p.m. on $\mathbb{R}$ with CD F that has a jump at x, that means $\mu x = F(x) - F(x-) > 0$. Since the total probability is one, there can be atmost n jumps of size $\frac{1}{n}$ or more. Putting them together, there can be atmost countably many jumps. In particular F is continuous on a dense set. Let J be the set of all jumps of F. Then, $F = F_{\text{atom}} + F_{\text{cts}}$ where $F_{\text{atom}}(x) := \sum_{x\in J}(F(x) - F(x-))$ and $F_{\text{cts}} = F - F_{\text{atom}}$. Clearly, F_{cts} is a continuous non-decreasing function, while F_{atom} is a non-decreasing continuous function that increases only in jumps (if $J \cap [a,b] = \emptyset$, then $F_{\text{atom}}(a) = F_{\text{atom}}(b)$).

If F_{atom} is not identically zero, then we can scale it up by $c = (F_{\text{atom}}(+\infty) - F_{\text{atom}}(-\infty))^{-1}$ to make it a CDF of a p.m. on $\mathbb{R}$. Similarly for F_{cts}. This means, we can write μ as $c\mu_{\text{atom}} + (1-c)\mu_{\text{cts}}$ where $c \in [0,1]$ and μ_{atom} is a purely atomic measure (its CDF increases only in jumps) and μ_{cts} has a continuous CDF.

Definition 1.37. Two measures μ and ν on the same $(\Omega, \mathscr{F})$ are said to be *mutually singular* and write $\mu \perp \nu$ if there is a set $A \in \mathscr{F}$ such that $\mu(A) = 0$ and $\nu(A^c) = 0$. We say that μ is *absolutely continuous to* ν and write $\mu \ll \mu$ if $\mu(A) = 0$ whenever $\nu(A) = 0$.

Remark 1.38. (i) Singularity is reflexive, absolute continuity is not. If $\mu \ll \nu$ and $\nu \ll \mu$, then we say that μ and ν are *mutually absolutely continuous*. (ii) If $\mu \perp \nu$, then we cannot also have $\mu \ll \nu$ (unless $\mu = 0$). (iii) Given μ and ν, it is not necessary that they be singular or absolutely continuous to one another.

Example 1.39. 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 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 singular to Lebesgue measure. A p.m. on the line with density (eg., $N(0,1)$) is absolutely continuous to $\mathbf{m}$. In fact $N(0,1)$ and $\mathbf{m}$ 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).

As explained above, a p.m on the line with atoms is singular (w.r.t $\mathbf{m}$). This raises the natural question of whether every p.m. with a continuous CDF is absolutely continuous to Lebesgue measure? Surprisingly, the answer is No!

Example 1.40 (Cantor measure). Let K be the middle-thirds Cantor set. Consider the canonical probability space $([0,1],\mathscr{B},\mathbf{m})$ and the random variable $X(\omega) = \sum_{k=1}^{\infty}\frac{2X_k(\omega)}{3^k}$, where $X_k(\omega)$ is the k^{th} binary digit of ω (i.e., $\omega = \sum_{k=1}^{\infty}\frac{X_k(\omega)}{2^k}$). Then X is measurable (why?). Let $\mu := \mathbf{m}X^{-1}$ be the pushforward.

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 has zero Lebsgue measure. Thus μ has not atoms and must have a continuous CDF. Since $\mu(K) = 1$ but $\mathbf{m}(K) = 0$, we also see that $\mu \perp \mathbf{m}$.

Exercise 1.41 (Alternate construction of Cantor measure). Let $K_1 = [0,1/3] \cup [2/3,1]$, $K_2 = [0,1/9]\cup[2/9,3/9]\cup[6/9,7/9]\cup[8/9,1]$, etc., be the decreasing sequence of compact sets whose intersection is K. Observe that K_n is a union of 2^n intervals each of length 3^{-n}. Let μ_n be the p.m. which is the "renormalized Lebesgue measure" on K_n. That is, $\mu_n(A) := 3^n 2^{-n}\mathbf{m}(A \cap K_n)$. Then each μ_n is a Borel p.m. Show that $\mu_n \overset{d}{\to} \mu$, the Cantor measure.

Example 1.42 (Bernoulli convolutions). We generalize the previous example. For any $\lambda > 1$, define $X_\lambda : [0,1] \to \mathbb{R}$ by $X(\omega) = \sum_{k=1}^{\infty}\lambda^{-k}X_k(\omega)$. Let $\mu_\lambda = \mathbf{m}X_\lambda^{-1}$ (did you check that X_λ is measurable?). For $\lambda = 3$, 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 p.m. on $\frac{1}{2}K := \{x/2 : x \in K\}$) and hence is singular.

Exercise 1.43. For any $\lambda > 2$, show that μ_λ is singular w.r.t. Lebesgue measure.

For $\lambda = 2$, it is easy to see that μ_λ is just the Lebesgue measue on $[0,1/2]$. Hence, one might expect that μ_λ is absolutely continuous to Lebesgue measure for $1 < \lambda < 2$. This is false! Paul Erdős showed that μ_λ is singular to Lebesgue measure whenever λ is a Pisot-Vijayaraghavan number, i.e., if λ 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.

Theorem 1.44 (Radon Nikodym theorem). *Suppose μ and ν are two measures on $(\Omega,\mathscr{F})$. Then $\mu \ll \nu$ if and only if there exists a non-negative measurable function $f : \Omega \to [0,\infty]$ such that $\mu(A) = \int_A f(x)d\nu(x)$ for all $A \in \mathscr{F}$.*

Remark 1.45. Then, f is called the *density of μ with respect to ν*. Note that the statement of the theorem does not make sense because we have not defined what $\int_A f(x)d\nu(x)$ means! That will come next class, and then, one of the two implications of the theorem, namely, "if μ has a density w.r.t. μ, then $\mu \ll \nu$" would become obvious. The converse statement, called the Radon-Nikodym theorem is non-trivial and will be proved in the measure theory class.

1.11. Expectation

Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. We define *Expectation* or *Lebesgue integration on measure space* in three steps.

1. If X can be written as $X = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ for some $A_i \in \mathscr{F}$, we say that X is a *simple r.v.*. We define its *expectation* to be $\mathbf{E}[X] := \sum_{i=1}^{n} c_i \mathbf{P}(A_i)$.
2. If $X \ge 0$ is a r.v., we define $\mathbf{E}[X] := \sup\{\mathbf{E}[S] : S \le X \text{ is a simple, nonnegative r.v.}\}$. Then, $0 \le \mathbf{E}X \le \infty$.
3. If X is any r.v. (real-valued!), let $X_+ := X\mathbf{1}_{X \ge 0}$ and $X_- := -X\mathbf{1}_{X<0}$ so that $X = X_+ - X_-$ (also observe that $X_+ + X_- = |X|$). If both $\mathbf{E}[X_+]$ and $\mathbf{E}[X_-]$ are finite, we say that X is integrable (or that $\mathbf{E}[X]$ exists) and define $\mathbf{E}[X] := \mathbf{E}[X_+] - \mathbf{E}[X_-]$.

Naturally, there are some arguments needed to complete these steps.

1. In the first step, one should check that $\mathbf{E}[X]$ is well-defined, as a simple r.v. can be represented as $\sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ in many ways. It helps to note that there is a unique way to write it in this form with A_k p.w disjoint. Finite additivity of $\mathbf{P}$ is used here.
2. In addition, check that the expectation defined in step 1 has the properties of *positivity* ($X \ge 0$ implies $\mathbf{E}[X] \ge 0$) and *linearity* ($\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y]$).
3. In step 2, again we would like to check positivity and linearity. It is clear that $\mathbf{E}[\alpha X] = \alpha \mathbf{E}[X]$ if $X \ge 0$ is a r.v and α is a non-negative real number (why?). One can also easily see that $\mathbf{E}[X+Y] \ge \mathbf{E}[X] + \mathbf{E}[Y]$ using the definition. To show that $\mathbf{E}[X+Y] \ge \mathbf{E}[X] + \mathbf{E}[Y]$, one proves using countable additivity of $\mathbf{P}$ -

   Monotone convergence theorem (provisional version). If S_n are non-negative simple r.v.s that increase to X, then $\mathbf{E}[S_n]$ increases to $\mathbf{E}[X]$.

   From this, linearity follows, since $S_n \uparrow X$ and $T_n \uparrow Y$ implies that $S_n + T_n \uparrow X + Y$. One point to check is that there do exist simple r.v S_n, T_n that increase to X, Y. For example, we can take $S_n(\omega) = \sum_{k=0}^{2^{2n}} \frac{k}{2^n} \mathbf{1}_{X(\omega) \in [k2^{-n}, (k+1)2^{-n})}$.

   An additional remark: It is convenient to allow a r.v. to take the value $+\infty$ but adopt the convention that $0 \cdot \infty = 0$ (infinite value on a set of zero probability does not matter).
4. In step 3, one assumes that both $\mathbf{E}[X_+]$ and $\mathbf{E}[X_-]$ are finite, which is equivalent to assuming that $\mathbf{E}[|X|] < \infty$. In other words, we deal with "absolutely integrable r.v.s" (no "conditionally convergent" stuff for us).

Let us say "$X = Y$ a.s" or "$X < Y$ a.s" etc., to mean that $\mathbf{P}(X = Y) = 1$, $P(X < Y) = 1$ etc. We may also use a.e. (almost everywhere) or w.p.1 (with probability one) in place of a.s (almost surely). To summarize, we end up with an expectation operator that has the following properties.

1. *Linearity:* X, Y integrable imples $\alpha X + \beta Y$ is also integrable and $\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y]$.

(2) *Positivity:* $X \geq 0$ implies $\mathbf{E}[X] \geq 0$. Further, if $X \geq 0$ and $\mathbf{P}(X = 0) < 1$, then $\mathbf{E}[X] > 0$. As a consequence, whenever $X \leq Y$ and $\mathbf{E}[X], \mathbf{E}[Y]$ exist, then $\mathbf{E}[X] \leq \mathbf{E}[Y]$ with equality if and only if $X = Y$ a.s.

(3) If X has expectation, then $|\mathbf{E}\, X| \leq \mathbf{E}|X|$.

(4) $\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$, in particular, $\mathbf{E}[1] = 1$.

1.12. Limit theorems for Expectation

Theorem 1.46 (Monotone convergence theorem (MCT))**.** *Suppose X_n, X are non-negative r.v.s and $X_n \uparrow X$ a.s. Then $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$. (valid even when $\mathbf{E}[X] = +\infty$).*

Theorem 1.47 (Fatou's lemma)**.** *Let X_n be non-negative r.v.s. Then $\mathbf{E}[\liminf X_n] \leq \liminf \mathbf{E}[X_n]$.*

Theorem 1.48 (Dominated convergence theorem (DCT))**.** *Let $|X_n| \leq Y$ where Y is a non-negative r.v. with $\mathbf{E}[Y] < \infty$. If $X_n \to X$ a.s., then, $\mathbf{E}[|X - n - X|] \to 0$ and hence we also get $\mathbf{E}[X_n] \to \mathbf{E}[X]$.*

Assuming MCT, the other two follow easily. For example, to prove Fatou's lemma, just define $Y_n = \inf_{n \geq k} X_n$ and observe that Y_ks increase to $\liminf X_n$ a.s and hence by MCT $\mathbf{E}[Y_k] \to \mathbf{E}[\liminf X_n]$. Since $X_n \geq Y_n$ for each n, we get $\liminf \mathbf{E}[X_n] \geq \liminf \mathbf{E}[Y_n] = \mathbf{E}[\liminf X_n]$.

To prove DCT, first note that $|X_n| \leq Y$ and $|X| \leq Y$ a.s. Consider the sequence of non-negative r.v.s $2Y - |X_n - X|$ that converges to $2Y$ a.s. Then, apply Fatou's lemma to get

$$\mathbf{E}[2Y] = \mathbf{E}[\liminf(2Y - |X_n - X|)] \leq \liminf \mathbf{E}[2Y - |X_n - X|] = \mathbf{E}[2Y] - \limsup \mathbf{E}[|X_n - X|].$$

Thus $\limsup \mathbf{E}[|X_n - X|] = 0$. Further, $|\mathbf{E}[X_n] - \mathbf{E}[X]| \leq \mathbf{E}[|X_n - X|] \to 0$.

1.13. Lebesgue integral versus Riemann integral

Consider the probability space $([0,1], \overline{\mathscr{B}}, \mathbf{m})$ (note that this is the Lebesgue σ-algebra, not Borel!) and a function $f : [0,1] \to \mathbb{R}$. Let

$$U_n := \frac{1}{2^n} \sum_{k=0}^{2^n-1} \max_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x), \qquad U_n := \frac{1}{2^n} \sum_{k=0}^{2^n-1} \min_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x)$$

be the upper and lower Riemann sums. Then, $L_n \leq U_n$ and U_n decrease with n while L_n increase. *If* $\lim U_n = \lim L_n$, we say that f is Riemann integrable and this common limit is defined to be the Riemann integral of f. The question of which functions are indeed Riemann integrable is answered precisely by

Lebesgue's theorem on Riemann integrals: *A bounded function f is Riemann integrable if and only if the set of discontinuity points has zero Lebesgue outer measure.*

Next consider the Lebesgue integral $\mathbf{E}[f]$. For this we need f to be Lebesgue measurable in the first place. Clearly any bounded and measurable function is integrable (why?). Plus, if f is continuous a.e., then f is measurable (why?). Thus, Riemann integrable functions are also Lebesgue integrable (but not conversely). What about the values of the two kinds of integrals? Define

$$g_n(x) := \sum_{k=0}^{2^n-1} \left(\max_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x) \right) \mathbf{1}_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}}, \qquad h_n(x) := \sum_{k=0}^{2^n-1} \left(\min_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x) \right) \mathbf{1}_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}}$$

so that $\mathbf{E}[g_n] = U_n$ and $\mathbf{E}[h_n] = L_n$. Further, $g_n(x) \downarrow f(x)$ and $h_n(x) \uparrow f(x)$ at all continuity points of f. By MCT, $\mathbf{E}[g_n]$ and $\mathbf{E}[h_n]$ converge to $\mathbf{E}[f]$, while by the assumed Riemann integrability L_n and U_n converge to $\int_0^1 f(x)dx$ (Riemann integral). Thus we must have $\mathbf{E}[f] = \int_0^1 f(x)dx$.

In short, when a function is Riemann integrable, it is also Lebesgue integrable, and the integrals agree. But there are functions that are measurable but not a.e. continuous. For example, consider the indicator function of a totally disconnected set of positive Lebesgue measure (like a Cantor set where an α middle portion is deleted at each stage, with α sufficiently small). Then at each point of the set, the indicator function is discontinuous. Thus, Lebesgue integral is more powerful than Riemann integral.

1.14. Lebesgue spaces:

Fix $(\Omega, \mathscr{F}, \mathbf{P})$. For $p \geq 1$, define $\|X\|_p := \mathbf{E}[|X|^p]^{\frac{1}{p}}$ for those r.v.s for which this number is finite. Then $\|tX\|_p = t\|X\|_p$ for $t > 0$, and Minkowski's inequality gives $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$ for any X and Y. However, $\|X\|_p = 0$ does not imply $X = 0$ but only that $X = 0$ a.s. Thus, $\|\cdot\|_p$ is a pseudo norm.

If we introduce the equivalence $X \sim Y$ if $\mathbf{P}(X = Y) = 1$, then for $p \geq 1$ only $\|\cdot\|_p$ becomes a genuine norm on the set of equivalence classes of r.v.s for which this quantity is finite (the L^p norm of an equivalence class is just $\|X\|_p$ for any X in the equivalence class). It is known as "L^p-space". With this norm (and the corresponding metric $\|X - Y\|_p$, the space L^p becomes a normed vector space. A non-trivial fact (proof left to measure theory class) is that L^p is a complete under this metric. A normed vector space which is complete under the induced metric is called a *Banach space* and L^p spaces are the prime examples.

The most important are the cases $p = 1, 2, \infty$. In these cases, (we just write X in place of $[X]$)

$$\|X-Y\|_1 := \mathbf{E}[|X-Y|] \qquad \|X-Y\|_2 := \sqrt{\mathbf{E}[|X-Y|^2]} \qquad \|X-Y\|_\infty := \inf\{t : \mathbf{P}(|X-Y| > t) = 0\}.$$

Exercise 1.49. For $p = 1, 2, \infty$, check that $\|X - Y\|_p$ is a metric on the space $L^p := \{[X] : \|X\|_p < \infty\}$ (here $[X]$ denotes the equivalence class of X under the above equivalence relation).

Especially special is the case $p = 2$, in which case the norm comes from an inner product $\langle [X], [Y] \rangle := \mathbf{E}[XY]$. L^2 is a complete inner product space, also known as a *Hilbert space*. For $p \neq 2$, the L^p norm does not come from an inner product as $\|\cdot\|_p$ does *not* satisfy the polarization identity $\|X+Y\|_p^2 + \|X-Y\|_p^2 = 2\|X\|_p^2 + 2\|Y\|_p^2$.

1.15. Some inequalities for expectations

The following inequalities are very useful. We start with the very general, but intuitively easy to understand Jensen's inequality. For this we recall two basic facts about convex functions on $\mathbb{R}$.

Let $\phi : (a, b) \to \mathbb{R}$ be a convex function. Then, (i) ϕ is continuous. (ii) Given any $u \in \mathbb{R}$, there is a line in the plane passing through the point $(u, \phi(u))$ such that the line lies below the graph of ϕ. If ϕ is strictly convex, then the only place where the line and the graph of ϕ meet, is at the point $(u,\phi(u))$. Proofs for these facts may be found in many books, eg., Rudin's "Real and Complex Analysis" (ch. 3).

Lemma 1.50 (Jensen's inequality). *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Let X be a r.v on some probability space. Assume that X and $\phi(X)$ both have expectations. Then, $\phi(\mathbf{E}X) \le \mathbf{E}[\phi(X)]$. The same assertion holds if ϕ is a convex function on some interval (a,b) and X takes values in (a,b) a.s.*

PROOF. Let $\mathbf{E}[X] = a$. Let $y = m(x-a)+\phi(a)$ be the 'supporting line' through $(a,\phi(a))$. Since the line lies below the graph of ϕ, we have $m(X-a)+\phi(a) \le \phi(X)$, a.s. Take expectations to get $\phi(a) \le E[\phi(X)]$. ■

Lemma 1.51. *(a) [Cauchy-Schwarz inequality] If X,Y are r.v.s on a probability space, then $\mathbf{E}[XY]^2 \le \mathbf{E}[X^2]\mathbf{E}[Y^2]$.*

(b) [Hölder's inequality] If X,Y are r.v.s on a probability space, then for any $p,q \ge 1$ satisfying $p^{-1}+q^{-1}=1$, we have $\|XY\|_1 \le \|X\|_p\|Y\|_q$.

PROOF. Cauchy-Schwarz is a special case of Hölder with $p=q=2$.

The proof of Hölder inequality follows by applying the inequality $a^p/p+b^q/q \ge ab$ for $a,b \ge 0$ to $a = |X|/\|X\|_p$ and $b = Y/\|Y\|_q$ and taking expectations. The inequality $a^p/p+b^q/q \ge ab$ is evident by noticing that the rectangle $[0,a]\times[0,b]$ (with area ab) is contained in the union of the region$\{(x,y) \,:\, 0 \le x \le a,\ 0 \le y \le x^{p-1}\}$ (with area a^p/p) and the region $\{(x,y) \,:\, 0 \le y \le b,\ 0 \le x \le y^{q-1}\}$ (with area b^q/q) simply 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 $(p-1)(q-1) = 1$. ■

Lemma 1.52 (Minkowski's inequality). *For any $p \ge 1$, we have $\|X+Y\|_p \le \|X\|_p + \|Y\|_p$.*

PROOF. For the important cases of $p = 1,2,\infty$, we know how to check this (for $p = 2$, use Cauchy-Schwarz). For general p, one can get it by applying Hölder to an appropriate pair of functions. We omit details (we might not use them, actually). ■

1.16. Change of variables

Lemma 1.53. **??** *Let $T : (\Omega_1,\mathscr{F}_1,\mathbf{P}) \to (\Omega_2,\mathscr{F}_2,\mathbf{Q})$ be measurable and $\mathbf{Q} = \mathbf{P}T^{-1}$. If X is an integrable r.v. on Ω_2, then $X \circ T$ is an integrable r.v. on Ω_1 and $\mathbf{E_P}[X \circ T] = \mathbf{E_Q}[X]$.*

PROOF. For a simple r.v., $X = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$, where $A_i \in \mathscr{F}_2$, it is easy to see that $X \circ T = \sum_{i=1}^n c_i \mathbf{1}_{T^{-1}A_i}$ and by definition $\mathbf{E_P}[X \circ T] = \sum_{i=1}^n c_i \mathbf{P}\{T^{-1}A_i\} = \sum_{i=1}^n c_i \mathbf{Q}\{A_i\}$ which is precisely $\mathbf{E_Q}[X]$. Use MCT to get to positive r.v.s and then to general integrable r.v.s. ■

Corollary 1.54. *Let X_i, $i \le n$, be random variables on a common probability space. Then for any Borel measurable $f : \mathbb{R}^n \to \mathbb{R}$, the value of $\mathbf{E}[f(X_1,\ldots,X_n)]$ (if it exists) depends only on the joint distribution of $X_1,\ldots X_n$.*

Remark 1.55. The change of variable result shows the irrelevance of the underlying probability space to much of what we do. That is, in any particular situation, all our questions may be about a finite or infinite collection of random variables X_i. Then, the answers depend only on the joint distribution of these random variables and not any other details of the underlying probability space. For instance, we can unambiguously talk of the expected value of $\text{Exp}(\lambda)$ distribution when we mean the expected value of a r.v having Exp(λ) distribution and defined on some probability space.

Density: Let ν be a measure on $(\Omega,\mathscr{F})$ and $X:\Omega\to[0,\infty]$ a r.v. Then set $\mu(A):=\int_A X d\nu$. Clearly, μ is a measure, as countable additivity follows from MCT. Observe that $\mu \ll \nu$. If two given measures μ and ν are related in this way by a r.v. X, then we say that X is the *density* or *Radon-Nikodym derivative* of μ w.r.t ν and sometimes write $X = \frac{d\mu}{d\nu}$. If it exists Radon-Nikodym derivative is unique (up to sets of ν-measure zero). The Radon-Nikodym theorem asserts that whenever μ,ν are σ-finite measures with $\mu \ll \nu$, the Radon Nikodym derivative does exist. When μ is a p.m on $\mathbb{R}^d$ and $\nu = \mathbf{m}$, we just refer to X as the pdf (probability density function) of μ. We also abuse language to say that a r.v. has density if its distribution has density w.r.t Lebesgue measure.

Exercise 1.56. Let X be a non-negative r.v on $(\Omega,\mathscr{F},\mathbf{P})$ and let $\mathbf{Q}(A) = \frac{1}{\mathbf{E_P}[X]}\int_A X d\mathbf{P}$. Then, $\mathbf{Q}$ is a p.m and for any non-negative r.v. Y, we have $\mathbf{E_Q}[Y] = \mathbf{E_P}[XY]$. The same holds if Y is real valued and assumed to be integrable w.r.t $\mathbf{Q}$ (or YX is assumed to be integrable w.r.t $\mathbf{P}$).

It is useful to know how densities transform under a smooth change of variables. It is an easy corollary of the change of variable formula and well-known substitution rules for computing integrals.

Corollary 1.57. *Suppose $X = (X_1,\dots,X_n)$ has density $f(\mathbf{x})$ on $\mathbb{R}^n$. Let $T:\mathbb{R}^n\to\mathbb{R}^n$ be injective and continuously differentiable. Write $U = T\circ X$. Then, U has density g which is given by $g(\mathbf{u}) = f(T^{-1}\mathbf{u})|\det(J[T^{-1}](\mathbf{u}))|$, where $[JT^{-1}]$ is the Jacobian of the inverse map T^{-1}.*

More generally, if we can write $\mathbb{R}^n = A_0\cup\ldots\cup A_n$, where A_i are pairwise disjoint, $\mathbf{P}(X\in A_0) = 0$ and such that $T_i := T|_{A_i}$ are one-one for $i = 1,2,\ldots n$, then, $g(\mathbf{u}) = \sum_{i=1}^n f(T_i^{-1}\mathbf{u})|\det(J[T_i^{-1}](\mathbf{u}))|$ where the i^{th} summand is understood to vanish if $\mathbf{u}$ is not in the range of T_i.

PROOF. **Step 1** Change of Lebesgue measure under T: For $A\in\mathscr{B}(\mathbb{R}^n)$, let $\mu(A) := \int_A|\det(J[T^{-1}](\mathbf{u}))|dm(u)$. Then, as remarked earlier, μ is a Borel measure. For sufficiently nice sets, like rectangles $[a_1,b_1]\times\ldots\times[a_n,b_n]$, we know from Calculus class that $\mu(A) = \mathbf{m}(T^{-1}(A))$. Since rectangles generate the Borel sigma-algebra, and μ and $\mathbf{m}\circ T^{-1}$ agree on rectangles, by the $\pi-\lambda$ theorem we get $\mathbf{m}\circ T^{-1} = \mu$. Thus, $\mathbf{m}\circ T^{-1}$ is a measure with density given by $|\det(J[T^{-1}](\cdot))|$.
Step 2 Let B be a Borel set and consider

$$\begin{aligned}\mathbf{P}(U\in B) &= \mathbf{P}(X\in T^{-1}B) = \int f(x)\mathbf{1}_{T^{-1}B}(x)d\mathbf{m}(x)\\ &= \int f(T^{-1}u)\mathbf{1}_{T^{-1}B}(T^{-1}u)d\mu(u) = \int f(T^{-1}u)\mathbf{1}_B(u)d\mu(u)\end{aligned}$$

where the first equality on the second line is by the change of variable formula of Lemma **??**. Apply exercise 1.56 and recall that μ has density $|\det(J[T^{-1}](\mathbf{u}))|dm(u)$ to get

$$\mathbf{P}(U\in B) = \int f(T^{-1}u)\mathbf{1}_B(u)d\mu(u) = \int f(T^{-1}u)\mathbf{1}_B(u)|\det(J[T^{-1}](\mathbf{u}))|dm(u)$$

which shows that U has density $f(T^{-1}u)|\det(J[T^{-1}](\mathbf{u}))|$.

To prove the second part, we do the same, except that in the first step, (using $\mathbf{P}(X \in A_0) = 0$, since $\mathbf{m}(A_0) = 0$ and X has density) $\mathbf{P}(U \in B) = \cup_{i=1}^{n} \mathbf{P}(X \in T_i^{-1}B) = \sum_{i=1}^{n} \int f(x) \mathbf{1}_{T_i^{-1}B}(x) d\mathbf{m}(x)$. The rest follows as before. ■

1.17. Distribution of the sum, product etc.

Suppose we know the joint distribution of $X = (X_1, \ldots, X_n)$. Then we can find the distribution of any function of X because $\mathbf{P}(f(X) \in A) = \mathbf{P}(X \in f^{-1}(A))$. When X has a density, one can get simple formulas for the density of the sum, product etc., that are quite useful.

In the examples that follow, let us assume that the density is continuous. This is only for convenience, and so that we can invoke theorems like $\iint f = \int \left(\int f(x,y) dy \right) dx$. Analogous theorems for Lebesgue integral will come later (*Fubini's theorem*)...

Example 1.58. Suppose (X,Y) has density $f(x,y)$ on $\mathbb{R}^2$. What is the distribution of X? Of $X+Y$? Of X/Y? We leave you to see that X has density $g(x) = \int_{\mathbb{R}} f(x,y) dy$. Assume that f is continuous so that the integrals involved are also Riemann integrals and you may use well known facts like $\iint f = \int \left(\int f(x,y) dy \right) dx$. The condition of continuity is unnatural and the result is true if we only assume that $f \in L^1$ (w.r.t. Lebesgue measure on the plane). The right to write Lebesgue integrals in the plane as iterated integrals will be given to us by *Fubini's theorem* later.

Suppose (X,Y) has density $f(x,y)$ on $\mathbb{R}^2$.

1. X has density $f_1(x) = \int_{\mathbb{R}} f(x,y) dy$ and Y has density $f_2(y) = \int_{\mathbb{R}} f(x,y) dx$. This is because, for any $a < b$, we have
$$\mathbf{P}(X \in [a,b]) = \mathbf{P}((X,Y) \in [a,b] \times \mathbb{R}) = \int\limits_{[a,b] \times \mathbb{R}} f(x,y) dx dy = \int\limits_{[a,b]} \left(\int\limits_{\mathbb{R}} f(x,y) dy \right) dx.$$
This shows that the density of X is indeed f_1.
2. Density of X^2 is $\left(f_1(\sqrt{x}) + f_1(-\sqrt{x}) \right) / 2\sqrt{x}$ for $x > 0$. Here we notice that T is one-one on $\{x > 0\}$ and $\{x < 0\}$ (and $\{x = 0\}$ has zero measure under f), so the second statement in the proposition is used.
3. The density of $X+Y$ is $g(t) = \int_{\mathbb{R}} f(t-v, v) dx$. To see this, let $U = X+Y$ and $V = Y$. Then the transformation is $T(x,y) = (x+y, y)$. Clearly $T^{-1}(u,v) = (u-v, v)$ whose Jacobian determinant is 1. Hence by corollary 1.57, we see that (U,V) has the density $g(u,v) = f(u-v, v)$. Now the density of U can be obtained like before as $h(u) = \int g(u,v) dv = \int f(u-v, v) dv$.
4. To get the density of XY, we define $(U,V) = (XY, Y)$ so that for $v \neq 0$, we have $T^{-1}(u,v) = (u/v, v)$ which has Jacobian determinant v^{-1}.

We claim that $X+Y$ has the density $g(t) = \int_{\mathbb{R}} f(t-v, v) dx$.

Exercise 1.59. (1) Suppose (X,Y) has a continuous density $f(x,y)$. Find the density of X/Y. Apply to the case when (X,Y) has the *standard bivariate normal distribution* with density $f(x,y) = (2\pi)^{-1} \exp\{-\frac{x^2+y^2}{2}\}$.

(2) Find the distribution of $X+Y$ if (X,Y) has the standard bivariate normal distribution.

(3) Let $U = \min\{X,Y\}$ and $V = \max\{X,Y\}$. Find the density of (U,V).

1.18. Mean, variance, moments

Given a r.v. or a random vector, expectations of various functions of the r.v give a lot of information about the distribution of the r.v. For example,

Proposition 1.60. *The numbers* $\mathbf{E}[f(X)]$ *as* f *varies over* $C_b(\mathbb{R})$ *determine the distribution of* X.

PROOF. Given any $x \in \mathbb{R}^n$, we can recover $F(x) = \mathbf{E}[\mathbf{1}_{A_x}]$, where $A_x = (-\infty, x_1] \times \ldots \times (-\infty, x_n]$ as follows. For any $\delta > 0$, let $f(y) = \min\{1, \delta^{-1} d(y, A^c_{x+\delta\mathbf{1}})\}$, where d is the L_∞ metric on $\mathbb{R}^n$. Then, $f \in C_b(\mathbb{R})$, $f(y) = 1$ if $y \in A_x$, $f(y) = 0$ if $y \not\in A_{x+\delta\mathbf{1}}$ and $0 \le f \le 1$. Therefore, $F(x) \le \mathbf{E}[f \circ X] \le F(x + \delta\mathbf{1})$. Let $\delta \downarrow 0$, invoke right continuity of F to recover $F(x)$. ■

Much smaller sub-classes of functions are also sufficient to determine the distribution of X.

Exercise 1.61. Show that the values $\mathbf{E}[f \circ X]$ as f varies over the class of all smooth (infinitely differentiable), compactly supported functions determine the distribution of X.

Expectations of certain functionals of random variables are important enough to have their own names.

Definition 1.62. Let X be a r.v. Then, $\mathbf{E}[X]$ (if it exists) is called the *mean* or *expected value* of X. $\text{Var}(X) := \mathbf{E}\left[(X - \mathbf{E}X)^2\right]$ is called the *variance* of X, and its square root is called the *standard deviation* of X. The standard deviation measures the spread in the values of X or one way of measuring the uncertainty in predicting X. For any $p > 0$, if it exists, $\mathbf{E}[X^p]$ is called the p^{th}*-moment* of X. The function ψ defined as $\psi(\lambda) := \mathbf{E}[e^{\lambda X}]$ is called the *moment generating function* of X. Note that the m.g.f of a non-negative r.v. exists for all $\lambda < 0$. It may exist for some $\lambda > 0$ also. A similar looking object is the *characteristic function* of X, define by $\phi(\lambda) := \mathbf{E}[e^{i\lambda X}] := \mathbf{E}[\cos(\lambda X)] + i\mathbf{E}[\sin(\lambda X)]$. This exists for all $\lambda \in \mathbb{R}$.

For two random variables X, Y on the same probability space, we define their *covariance* to be $\text{Cov}(X,Y) := \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. The *correlation coefficient* is measured by $\frac{\text{Cov(X,Y)}}{\sqrt{\text{Var(X)Var(Y)}}}$. The correlation coefficient lies in $[-1,1]$ and measures the association between X and Y. A correlation of 1 implies $X = Y$ a.s while a correlation of -1 implies $X = -Y$ a.s. Covariance and correlation depend only on the joint distribution of X and Y.

Exercise 1.63. (i) Express the mean, variance, moments of $aX + b$ in terms of the same quantities for X.

(ii) Show that $\text{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$.

(iii) Compute mean, variance and moments of the Normal, exponential and other distributions defined in section 1.7.

Example 1.64 (The exponential distribution). Let $X \sim \text{Exp}(\lambda)$. Then, $\mathbf{E}[X^k] = \int x^k d\mu(x)$ where μ is the p.m on $\mathbb{R}$ with density $\lambda e^{-\lambda x}$ (for $x > 0$). Thus, $\mathbf{E}[X^k] = \int x^k \lambda e^{-\lambda x} dx = \lambda^{-k} k!$. In particular, the mean is λ, the variance is $2\lambda^2 - (\lambda)^2 = \lambda^2$. In case of the normal distribution, check that the even moments are given by $\mathbf{E}[X^{2k}] = \prod_{j=1}^k (2j-1)$.

Remark 1.65 (Moment problem). Given a sequence of numbers $(\alpha_k)_{k\ge 0}$, is there a p.m μ on $\mathbb{R}$ whose k^{th} moment is α_k? If so, is it unique?

This is an extremely interesting question and its solution involves a rich interplay of several aspects of classical analysis (orthogonal polynomials, tridiagonal matrices, functional analysis, spectral theory etc). Note that there are are some non-trivial conditions for (α_k) to be the moment sequence of a p.m. μ. For example, $\alpha_0 = 1$, $\alpha_2 \ge \alpha_1^2$ etc. In the homework you were asked to show that $(\!(\alpha_{i+j})\!)_{i,j\le n}$ should be a n.n.d. matrix for every n. The non-trivial answer is that these conditions are also sufficient!

Note that like proposition 1.60, the uniqueness question is asking whether $\mathbf{E}[f \circ X]$, as f varies over the space of polynomials, is sufficient to determine the distribution of X. However, uniqueness is not true in general. In other words, one can find two p.m μ and ν on $\mathbb{R}$ which have the same sequence of moments!

CHAPTER 2

Independent random variables

2.1. Product measures

Definition 2.1. Let μ_i be measures on $(\Omega_i, \mathscr{F}_i)$, $1 \le i \le n$. Let $\mathscr{F} = \mathscr{F}_1 \otimes \ldots \otimes \mathscr{F}_n$ be the sigma algebra of subsets of $\Omega := \Omega_1 \times \ldots \times \Omega_n$ generated by all "rectangles" $A_1 \times \ldots \times A_n$ with $A_i \in \mathscr{F}_i$. Then, the measure μ on $(\Omega, \mathscr{F})$ such that $\mu(A_1 \times \ldots \times A_n) = \prod_{i=1}^n \mu_i(A_i)$ whenever $A_i \in \mathscr{F}_i$ is called a *product measure* and denoted $\mu = \mu_1 \otimes \ldots \otimes \mu_n$.

The existence of product measures follows along the lines of the Caratheodary construction starting with the π-system of rectangles. We skip details, *but in the cases that we ever use, we shall show existence by a much neater method* in Proposition 2.8. Uniqueness of product measure follows from the $\pi - \lambda$ theorem because rectangles form a π-system that generate the σ-algebra $\mathscr{F}_1 \otimes \ldots \otimes \mathscr{F}_n$.

Example 2.2. Let $\mathscr{B}_d, \mathbf{m}_d$ denote the Borel sigma algebra and Lebesgue measure on $\mathbb{R}^d$. Then, $\mathscr{B}_d = \mathscr{B}_1 \otimes \ldots \otimes \mathscr{B}_1$ and $\mathbf{m}_d = \mathbf{m}_1 \otimes \ldots \otimes \mathbf{m}_1$. The first statement is clear (in fact $\mathscr{B}_{d+d'} = \mathscr{B}_d \otimes \mathscr{B}_{d'}$). Regarding $\mathbf{m}_d$, by definition, it is the unique measure for which $\mathbf{m}_d(A_1 \times \ldots \times A_n)$ equals $\prod_{i=1}^n \mathbf{m}_1(A_i)$ for all intervals A_i. To show that it is the d-fold product of $\mathbf{m}_1$, we must show that the same holds for any Borel sets A_i.

Fix intervals $A_2, \ldots, A_n$ and let $S := \{A_1 \in \mathscr{B}_1 : \mathbf{m}_d(A_1 \times \ldots \times A_n) = \prod_{i=1}^n \mathbf{m}_1(A_i)\}$. Then, S contains all intervals (in particular the π-system of semi-closed intervals) and by properties of measures, it is easy to check that S is a λ-system. By the $\pi - \lambda$ theorem, we get $S = \mathscr{B}_1$ and thus, $\mathbf{m}_d(A_1 \times \ldots \times A_n) = \prod_{i=1}^n \mathbf{m}_1(A_i)$ for all $A_1 \in \mathscr{B}_1$ and any intervals $A_2, \ldots, A_n$. Continuing the same argument, we get that $\mathbf{m}_d(A_1 \times \ldots \times A_n) = \prod_{i=1}^n \mathbf{m}_1(A_i)$ for all $A_i \in \mathscr{B}_1$.

The product measure property is defined in terms of sets. As always, it may be written for measurable functions and we then get the following theorem.

Theorem 2.3 (Fubini's theorem). *Let $\mu = \mu_1 \otimes \mu_2$ be a product measure on $\Omega_1 \times \Omega_2$ with the product σ-algebra. If $f : \Omega \to \mathbb{R}_+$ is either a non-negative r.v. or integrable w.r.t μ, then,*

(1) *For every $x \in \Omega_1$, the function $y \to f(x,y)$ is $\mathscr{F}_2$-measurable, and the function $x \to \int f(x,y) d\mu_2(y)$ is $\mathscr{F}_1$-measurable. The same holds with x and y interchanged.*

(2) $\int\limits_{\Omega} f(z) d\mu(z) = \int\limits_{\Omega_1} \left(\int\limits_{\Omega_2} f(x,y) d\mu_2(y) \right) d\mu_1(x) = \int\limits_{\Omega_2} \left(\int\limits_{\Omega_1} f(x,y) d\mu_1(x) \right) d\mu_2(y).$

PROOF. Skipped. Attend measure theory class. ■

Needless to day (*self:* then why am I saying this?) all this goes through for finite products of σ-finite measures.

Infinite product measures: Given $(\Omega_i, \mathscr{F}_i, \mu_i)$, $i = 1, 2, \ldots$, let $\Omega := \Omega_1 \times \Omega_2 \times \ldots$ and let $\mathscr{F}$ be the sigma algebra generated by all finite dimensional cylinders $A_1 \times \ldots \times A_n \times \Omega_{n+1} \times \Omega_{n+2} \ldots$ with $A_i \in \mathscr{F}_i$. Does there exist a "product measure" μ on $\mathscr{F}$?

For concreteness take all $(\Omega_i, \mathscr{F}_i, \mu_i) = (\mathbb{R}, \mathscr{B}, \nu)$. What measure should the product measure μ give to the set $A \times \mathbb{R} \times \mathbb{R} \times \ldots$? If $\nu(\mathbb{R}) > 1$, it is only reasonable to set $\mu(A \times \mathbb{R} \times \mathbb{R} \times \ldots)$ to infinity, and if $\nu(\mathbb{R}) < 1$, it is reasonable to set it to 0. But then all cylinders will have zero measure or infinite measure!! If $\nu(\mathbb{R}) = 1$, at least this problem does not arise. We shall show that it is indeed possible to make sense of infinite products of Thus, the only case when we can talk reasonably about infinite products of measures is for probability measures.

2.2. Independence

Definition 2.4. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space. Let $\mathscr{G}_1, \ldots, \mathscr{G}_k$ be sub-sigma algebras of $\mathscr{F}$. We say that $\mathscr{G}_i$ are *independent* if for every $A_1 \in \mathscr{G}_1, \ldots, A_k \in \mathscr{G}_k$, we have $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_k)$.

Random variables $X_1, \ldots, X_n$ on $\mathscr{F}$ are said to be independent if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent. This is equivalent to saying that $\mathbf{P}(X_i \in A_i \; i \le k) = \prod_{i=1}^{k} \mathbf{P}(X_i \in A_i)$ for any $A_i \in \mathscr{B}(\mathbb{R})$.

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 saying that $\mathbf{P}(A_{j_1} \cap \ldots \cap A_{j_\ell}) = \mathbf{P}(A_{j_1}) \ldots \mathbf{P}(A_{j_\ell})$ for any $1 \le j_1 < j_2 < \ldots < j_\ell \le k$.

In all these cases, an infinite number of objects (sigma algebras or random variables or events) are said to be independent if every finite number of them are independent.

Some remarks are in order.

(1) As usual, to check independence, it would be convenient if we need check the condition in the definition only for a sufficiently large class of sets. However, if $\mathscr{G}_i = \sigma(S_i)$, and for every $A_1 \in S_1, \ldots, A_k \in S_k$ if we have $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_k)$, we *cannot* conclude that $\mathscr{G}_i$ are independent! If S_i are π-systems, this is indeed true (see below).

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

Lemma 2.5. *If S_i are π-systems and $\mathscr{G}_i = \sigma(S_i)$ and for every $A_1 \in S_1, \ldots, A_k \in S_k$ if we have $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_k)$, then $\mathscr{G}_i$ are independent.*

PROOF. Fix $A_2 \in S_2, \ldots, A_k \in S_k$ and set $\mathscr{F}_1 := \{B \in \mathscr{G}_1 \; : \; \mathbf{P}(B \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(B)\mathbf{P}(A_2) \ldots \mathbf{P}(A_k)\}$. Then $\mathscr{F}_1 \supset S_1$ by assumption and it is easy to check that $\mathscr{F}_1$ is a λ-system. By the π-λ theorem, it follows that $\mathscr{F}_1 = \mathscr{G}_1$ and we get the assumptions of the lemma for $\mathscr{G}_1, S_2, \ldots, S_k$. Repeating the argument for S_2, S_3 etc., we get independence of $\mathscr{G}_1, \ldots, \mathscr{G}_k$. ■

Corollary 2.6. (1) *Random variables $X_1, \ldots, X_k$ are independent if and only if $\mathbf{P}(X_1 \le t_1, \ldots, X_k \le t_k) = \prod_{j=1}^{k} \mathbf{P}(X_j \le t_j)$.*

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

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

PROOF. (1) The sets $(-\infty,t]$ form a π-system that generates $\mathscr{B}(\mathbb{R})$. (2) For $j \le k$, let S_j be the collection of finite intersections of sets A_i, $i \in I_j$. Then S_j are π-systems and $\sigma(S_j) = \mathscr{F}_j$. (3) Follows from (2) by considering $\mathscr{G}_{i,j} := \sigma(X_{i,j})$ and observing that $f_i(X_{i,1},\dots,X_{i,k}) \in \sigma(\mathscr{G}_{i,1} \cup \dots \cup \mathscr{G}_{i,n_i})$. ■

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

Lemma 2.7. (1) *Sigma algebras $\mathscr{G}_1,\dots,\mathscr{G}_k$ are independent if and only if for every bounded $\mathscr{G}_i$-measurable functions X_i, $1 \le i \le k$, we have, $\mathbf{E}[X_1 \dots X_k] = \prod_{i=1}^k \mathbf{E}[X_i]$.*

(2) *In particular, random variables $Z_1,\dots,Z_k$ (Z_i is an n_i dimensional random vector) are independent if and only if $\mathbf{E}[\prod_{i=1}^k f_i(Z_i)] = \prod_{i=1}^k \mathbf{E}[f_i(Z_i)]$ 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 $X_2,\dots,X_k$ 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 $\mathscr{G}_i$ are independent. If X_i are $\mathscr{G}_i$ measurable then it is clear that X_i are independent and hence $\mathbf{P}(X_1,\dots,X_k)^{-1} = \mathbf{P}X_1^{-1} \otimes \dots \otimes \mathbf{P}X_k^{-1}$. Denote $\mu_i := \mathbf{P}X_i^{-1}$ and apply Fubini's theorem (and change of variables) to get

$$\begin{aligned}
\mathbf{E}[X_1 \dots X_k] &\overset{\text{c.o.v}}{=} \int_{\mathbb{R}^k} \prod_{i=1}^k x_i d(\mu_1 \otimes \dots \otimes \mu_k)(x_1,\dots,x_k) \\
&\overset{\text{Fub}}{=} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \prod_{i=1}^k x_i d\mu_1(x_1) \dots d\mu_k(x_k) \\
&= \prod_{i=1}^k \int_{\mathbb{R}} u d\mu_i(u) \overset{\text{c.o.v}}{=} \prod_{i=1}^k \mathbf{E}[X_i].
\end{aligned}$$

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

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

■

2.3. Independent sequences of random variables

First we make the observation that product measures and independence are closely related concepts. For example,

An observation: The independence of random variables $X_1,\dots,X_k$ is precisely the same as saying that $\mathbf{P} \circ X^{-1}$ is the product measure $\mathbf{P}X_1^{-1} \otimes \dots \otimes \mathbf{P}X_k^{-1}$, where $X = (X_1,\dots,X_k)$.

Consider the following questions. Henceforth, we write $\mathbb{R}^\infty$ for the countable product space $\mathbb{R}\times\mathbb{R}\times\dots$ and $\mathscr{B}(\mathbb{R}^\infty)$ for the cylinder σ-algebra generated by all finite dimensional cylinders $A_1\times\dots\times A_n\times\mathbb{R}\times\mathbb{R}\times\dots$ with $A_i\in\mathscr{B}(\mathbb{R})$. This notation is justified, becaue the cylinder σ-algebra is also the Borel σ-algebra on $\mathbb{R}^\infty$ with the product topology.

Question 1: Given $\mu_i\in\mathscr{P}(\mathbb{R})$, $i\ge 1$, does there exist a probability space with independent random variables X_i having distributions μ_i?

Question 2: Given $\mu_i\in\mathscr{P}(\mathbb{R})$, $i\ge 1$, does there exist a p.m μ on $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ such that $\mu(A_1\times\dots\times A_n\times\mathbb{R}\times\mathbb{R}\times\dots)=\prod_{i=1}^n\mu_i(A_i)$?

Observation: The above two questions are equivalent. For, suppose we answer the first question by finding an $(\Omega,\mathscr{F},\mathbf{P})$ with *independent* random variables $X_i:\Omega\to\mathbb{R}$ such that $X_i\sim\mu_i$ for all i. Then, $X:\Omega\to\mathbb{R}^\infty$ defined by $X(\omega)=(X_1(\omega),X_2(\omega),\dots)$ is measurable w.r.t the relevant σ-algebras (why?). Then, let $\mu:=\mathbf{P}X^{-1}$ be the pushforward p.m on $\mathbb{R}^\infty$. Clearly

$$\begin{aligned}\mu(A_1\times\dots\times A_n\times\mathbb{R}\times\mathbb{R}\times\dots) &= \mathbf{P}(X_1\in A_1,\dots,X_n\in A_n)\\ &= \prod_{i=1}^n\mathbf{P}(X_i\in A_i)=\prod_{i=1}^n\mu_i(A_i).\end{aligned}$$

Thus μ is the product measure required by the second question.

Conversely, if we could construct the product measure on $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$, then we could take $\Omega=\mathbb{R}^\infty$, $\mathscr{F}=\mathscr{B}(\mathbb{R}^\infty)$ and X_i to be the i^{th} co-ordinate random variable. Then you may check that they satisfy the requirements of the first question.

The two questions are thus equivalent, but what is the answer?! It is 'yes', of course or we would not make heavy weather about it.

Proposition 2.8 (Daniell). *Let $\mu_i\in\mathscr{P}(\mathbb{R})$, $i\ge 1$, be Borel p.m on $\mathbb{R}$. Then, there exist a probability space with* independent *random variables $X_1,X_2,\dots$ such that $X_i\sim\mu_i$.*

PROOF. We arrive at the construction in three stages.

(1) **Independent Bernoullis:** Consider $([0,1],\mathscr{B},\mathbf{m})$ and 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. For definiteness, we may always take the infinite binary expansion. Then by an earlier homework exercise, $X_1,X_2,\dots$ are independent Bernoulli(1/2) random variables.

(2) **Independent uniforms:** Note that as a consequence, 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 $U_1:=X_1/2+X_3/2^3+X_5/2^5+\dots$, $U_2:=X_2/2+X_6/2^2+\dots$, etc. Clearly, U_i are i.i.d. U[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 $\mu_1,\mu_2,\dots$. 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. ■

Sometimes in books one finds construction of uncountable product measures too. It has no use. But a very natural question at this point is to go beyond independence. We just state the following theorem which generalizes the previous proposition.

Theorem 2.9 (Kolmogorov's existence theorem). *For each $n \geq 1$ and each $1 \leq i_1 < i_2 < \ldots < i_n$, let $\mu_{i_1,\ldots,i_n}$ be a Borel p.m on $\mathbb{R}^n$. Then there exists a unique probability measure μ on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ such that*

$$\mu(A_1 \times \ldots \times A_n \times \mathbb{R} \times \mathbb{R} \times \ldots) = \mu_{i_1,\ldots,i_n}(A_1 \times \ldots \times A_n) \text{ for all } n \geq 1 \text{ and all } A_i \in \mathscr{B}(\mathbb{R}),$$

if and only if the given family of probability measures satisfy the consistency condition

$$\mu_{i_1,\ldots,i_n}(A_1 \times \ldots \times A_{n-1} \times \mathbb{R}) = \mu_{i_1,\ldots,i_{n-1}}(A_1 \times \ldots \times A_{n-1})$$

for any $A_k \in \mathscr{B}(\mathbb{R})$ and for any $1 \leq i_1 < i_2 < \ldots < i_n$ and any $n \geq 1$.

2.4. Some probability estimates

Lemma 2.10 (Borel Cantelli lemmas). *Let A_n be events on a common probability space.*

(1) *If $\sum_n \mathbf{P}(A_n) < \infty$, then $\mathbf{P}(A_n \text{ i.o}) = 0$.*
(2) *If A_n are independent and $\sum_n \mathbf{P}(A_n) = \infty$, then $\mathbf{P}(A_n \text{ i.o}) = 1$.*

PROOF. (1) For any N, $\mathbf{P}\left(\cup_{n=N}^{\infty} A_n\right) \le \sum_{n=N}^{\infty} \mathbf{P}(A_n)$ which goes to zero as $N \to \infty$. Hence $\mathbf{P}(\limsup A_n) = 0$.
(2) For any $N < M$, $\mathbf{P}(\cup_{n=N}^{M} A_n) = 1 - \prod_{n=N}^{M} \mathbf{P}(A_n^c)$. Since $\sum_n \mathbf{P}(A_n) = \infty$, it follows that $\prod_{n=N}^{M}(1 - \mathbf{P}(A_n)) \le \prod_{n=N}^{M} e^{-\mathbf{P}(A_n)} \to 0$, for any fixed N as $M \to \infty$. Hence $\mathbf{P}\left(\cup_{n=N}^{\infty} A_n\right) = 1$ for all N, implying that $\mathbf{P}(A_n \text{ i.o}) = 1$. ■

Lemma 2.11 (First and second moment methods). *Let $X \ge 0$ be a r.v.*

(1) *(**Markov's inequality a.k.a first moment method**) For any $t > 0$, we have $\mathbf{P}(X \ge t) \le t^{-1}\mathbf{E}[X]$.*
(2) *(**Paley-Zygmund inequality a.k.a second moment method**) For any non-negative r.v. X,*

$$(i)\,\mathbf{P}(X > 0) \ge \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]}. \qquad (ii)\,\mathbf{P}(X > \alpha\mathbf{E}[X]) \ge (1-\alpha)^2 \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]}.$$

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

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

Remark 2.12. Applying these inequalities to other functions of X can give more information. For example, if X has finite variance, $\mathbf{P}(|X - \mathbf{E}[X]| \ge t) = \mathbf{P}(|X - \mathbf{E}[X]|^2 \ge t^2) \le t^{-2}\text{Var}(X)$, which is called *Chebyshev's inequality*. Higher the moments that exist, better the asymptotic tail bounds that we get. For example, if $\mathbf{E}[e^{\lambda X}] < \infty$ for some $\lambda > 0$, we get exponential tail bounds by $\mathbf{P}(X > t) = \mathbf{P}(e^{\lambda X} < e^{\lambda t}) \le e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$.

2.5. Applications of first and second moment methods

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

Application 1: Borel-Cantelli lemmas: The first Borel Cantelli lemma follows from Markov's inequality. In fact, applied to $X = \sum_{k=N}^{\infty} \mathbf{1}_{A_k}$, Markov's inequality is the same as the union bound $\mathbf{P}(A_N \cup A_{N+1} \cup \ldots) \le \sum_{k=N}^{\infty} \mathbf{P}(A_k)$ which is what gave us the first Borel-Cantelli lemma.

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

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

Apply the second moment method to se that for any fixed n, as $m \to \infty$,

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

by assumption that $\sum \mathbf{P}(A_k) = \infty$. This shows that $\mathbf{P}(\cup_{k\geq n} A_k) = 1$ for any n and hence $\mathbf{P}(\limsup A_n) = 1$.

Note that this proof used independence only to claim that $\mathbf{P}(A_k \cap A_\ell) = \mathbf{P}(A_k)\mathbf{P}(A_\ell)$. Therefore the second Borel-Cantelli lemma holds for *pairwise independent* events too!

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

Theorem 2.13. *Let T_n denote the number of days till each book is picked at least once. Then T_n is "concentrated around $n \log n$ in a window of size n" by which we mean that for any sequence $\theta_n \to \infty$, we have*

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

Remark 2.14. In the following proof and many other places, we shall have occasion to make use of the elementary estimate

$$1 - x \leq e^{-x} \quad \forall x, \qquad 1 - x \geq e^{-x-x^2} \quad \forall |x| < \frac{1}{2}. \tag{2.1}$$

The first inequality follows by expanding e^{-x} while the second follows by expanding $\log(1-x) = -x - x^2/2 - x^3/3 - \ldots$ (valid for $|x| < 1$).

PROOF. Fix an integer $t \geq 1$ and let $X_{t,k}$ be the indicator that the k^{th} book is not picked up on the first t days. Then, $\mathbf{P}(T_n > t) = \mathbf{P}(S_{t,n} \geq 1)$ where $S_{t,n} = X_{t,1} + \ldots + X_{t,n}$. As $\mathbf{E}[X_{t,k}] = (1 - 1/n)^k$ and $\mathbf{E}[X_{t,k}X_{t,\ell}] = (1 - 2/n)^k$ for $k \neq \ell$, we also compute that thefirst two moments of $S_{t,n}$ and use (2.1) to get

$$ne^{-\frac{t}{n} - \frac{t}{n^2}} \leq \mathbf{E}[S_{t,n}] = n\left(1 - \frac{1}{n}\right)^t \leq ne^{-\frac{t}{n}}. \tag{2.2}$$

$$\mathbf{E}[S_{t,n}^2] = n\left(1 - \frac{1}{n}\right)^t + n(n-1)\left(1 - \frac{2}{n}\right)^t \leq ne^{-\frac{t}{n}} + n(n-1)e^{-\frac{2t}{n}}. \tag{2.3}$$

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

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

$$\mathbf{P}(T_n > n\log n + n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \leq \mathbf{E}[S_{t,n}] \leq ne^{-\frac{n\log n + n\theta_n}{n}} \leq e^{-\theta_n} = o(1). \tag{2.4}$$

On the other hand, taking $t < n\log n - n\theta_n$ (where we take $\theta_n < \log n$, of course!), we now apply the second moment method. For any $n \geq 2$, by using (2.3) we get $\mathbf{E}[S_{t,n}^2] \leq e^{\theta_n} + e^{2\theta_n}$. The first inequality in (2.2) gives $\mathbf{E}[S_{t,n}] \geq e^{\theta_n - \frac{\log n - \theta_n}{n}}$. Thus,

$$\mathbf{P}(T_n > n\log n - n\theta_n) = \mathbf{P}(S_{t,n} \geq 1) \geq \frac{\mathbf{E}[S_{t,n}]^2}{\mathbf{E}[S_{t,n}^2]} \geq \frac{e^{2\theta_n - 2\frac{\log n - \theta_n}{n}}}{e^{\theta_n} + e^{2\theta_n}} = 1 - o(1) \tag{2.5}$$

as $n \to \infty$. From (2.4) and (2.5), we get the sharp bounds

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

Application 3: Branching processes: Consider a Galton-Watson branching process with offsprings that are i.i.d ξ. Let Z_n be the number of offsprings in the n^{th} generation. Take $Z_0 = 1$.

Theorem 2.15. (1) *If $m < 1$, then w.p.1, the branching process dies out. That is $\mathbf{P}(Z_n = 0$ for all large $n) = 1$.*

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

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

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

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

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

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

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

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

Remark 2.16. The fundamental result of branching processes also asserts the a.s extinction for the critical case $m = 1$. We omit this for now.

Application 4: How many prime divisors does a number typically have? For a natural number k, let $\nu(k)$ be the number of (distinct) prime divisors of n. What is the typical size of $\nu(n)$ as compared to n? We have to add the word typical, because if p is a prime number then $\nu(p) = 1$ whereas $\nu(2 \times 3 \times \ldots \times p) = p$. Thus there are arbitrarily large numbers with $\nu = 1$ and also numbers for which ν is as large as we wish. To give meaning to "typical", we draw a number at random and look at its ν-value. As there is no natural way to pick one number at random, the usual way of making precise what we mean by a "typical number" is as follows.

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

Theorem 2.17 (Hardy, Ramanujan). *With the above setting, for any $\delta > 0$, as $n \to \infty$ we have*

$$\mu_n\left\{k \in [n] : \Big|\frac{\nu(k)}{\log\log n} - 1\Big| > \delta\right\} \to 0. \tag{2.6}$$

PROOF. (**Turan**). Fix n and for any prime p define $X_p : [n] \to \mathbb{R}$ by $X_p(k) = \mathbf{1}_{p|k}$. Then, $\nu(k) = \sum\limits_{p\leq k} X_p(k)$. We define $\psi(k) := \sum\limits_{p\leq \sqrt[4]{k}} X_p(k)$. Then, $\psi(k) \leq \nu(k) \leq \psi(k)+4$ since there can be at most four primes larger than $\sqrt[4]{k}$ that divide k. From this, it is clearly enough to show (2.6) for ψ in place of ν (why?).

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

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

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

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

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

The second one is obvious, while the first one is not hard, (see exercise 2.18 below)). Thus, we get $\mathbf{E}_n[\psi] = \log\log n + O(n^{-\frac{3}{4}})$ and $\mathrm{Var}_n[\psi] = \log\log n + O(1)$. Thus, by

Chebyshev's inequality,

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

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

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

Exercise 2.18. $\sum\limits_{p\le\sqrt[4]{n}}\frac{1}{p}\sim\log\log n$

2.6. Weak law of large numbers

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

Theorem 2.19 (Kolmogorov). *Let $X_1,X_2\ldots$ be i.i.d random variables. If* $\mathbf{E}[|X_1|]<\infty$*, then for any $\delta>0$, as $n\to\infty$, we have*

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

In language to be introduced later, we shall say that S_n/n converges to zero in probability *and write* $\frac{S_n}{n}\overset{P}{\to}\mathbf{E}[X_1]$

PROOF. **Step 1:** First assume that X_i have finite variance σ^2. Without loss of generality take $\mathbf{E}[X_1]=0$ (or else replace X_i by $X_i-\mathbf{E}[X_1]$. Then, $\mu=\mathbf{E}[X_1]$. Then, by the first moment method (Chebyshev's inequality), $\mathbf{P}(|n^{-1}S_n|>\delta)\le n^{-2}\delta^{-2}\mathrm{Var}(S_n)$. By the independence of X_is, we see that $\mathrm{Var}(S_n)=n\sigma^2$. Thus, $\mathbf{P}(|\frac{S_n}{n}|>\delta)\le\frac{\sigma^2}{n\delta^2}$ 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 $Y_k:=X_k\mathbf{1}_{|X_k|\le A_n}$ and $Z_k:=X_k\mathbf{1}_{|X_k|>A_n}$ for some A_n to be chosen later. Then, Y_i are i.i.d, with some mean $\mu_n:=\mathbf{E}[Y_1]=-\mathbf{E}[Z_1]$ that depends on A_n and goes to zero as $A_n\to 0$. We shall choose A_n going to infinity, so that for large enough n, we do have $|\mu_n|<\delta$ (for an arbitrary fixed $\delta>0$).

$|Y_1|\le A_n$, hence $\mathrm{Var}(Y_1)\le\mathbf{E}[Y_1^2]\le A_n\mathbf{E}[|X_1|]$. By the Chebyshev bound that we used in step 1,

$$\mathbf{P}\left(\left|\frac{S_n^Y}{n}-\mu_n\right|>\delta\right)\le\frac{\mathrm{Var}(Y_1)}{n\delta^2}\le\frac{A_n\mathbf{E}[|X_1|]}{n\delta^2}.\tag{2.7}$$

Further, if n is large, then $|\mu_n|<\delta$ and then

$$\mathbf{P}\left(\left|\frac{S_n^Z}{n}+\mu_n\right|>\delta\right)\le\mathbf{P}\left(S_n^Z\ne 0\right)\le n\mathbf{P}(Z_1\ne 0)=n\mathbf{P}(|X_1|>A_n).\tag{2.8}$$

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

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

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

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

which gives the desired conclusion. ■

2.7. Applications of weak law of large numbers

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

Application 1: Bernstein's proof of Wierstrass' approximation theorem.

Theorem 2.20. *The set of polynomials is dense in the space of continuous functions (with the sup-norm metric) on an interval of the line.*

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

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

$$\begin{aligned}\left|\mathbf{E}_p\left[f\left(\frac{S_n}{n}\right)\right] - f(p)\right| &\le \mathbf{E}_p\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\right] \\ &= \mathbf{E}_p\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\mathbf{1}_{\left|\frac{S_n}{n}-p\right|\le\delta}\right] + \mathbf{E}_p\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\mathbf{1}_{\left|\frac{S_n}{n}-p\right|>\delta}\right] \\ &\le \omega_f(\delta) + 2\|f\|\mathbf{P}_p\left(\left|\frac{S_n}{n} - p\right| > \delta\right) \qquad (2.9)\end{aligned}$$

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

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

Plugging this into (2.9) and recalling that $Q_{f,n}(p) = \mathbf{E}_p\left[f\left(\frac{S_n}{n}\right)\right]$, we get

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

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

Remark 2.21. It is possible t write the proof without invoking WLLN. In fact, we did not use WLLN, but the Chebyshev bound. The main point is that the Binomial(n,p) probability measure puts almost all its mass between $np(1-\delta)$ and $np(1+\delta)$. Nevertheless, WLLN makes it transparent why this is so.

Application 2: Monte Carlo method for evaluating integrals. Consider a continuous function $f : [a,b] \to \mathbb{R}$ whose integral we would like to compute. Quite often, the form of the function may be sufficiently complicated that we cannot analytically compute it, but is explicit enough that we can numerically evaluate (on a computer) $f(x)$ for any specified x. Here is how one can evaluate the integral by use of random numbers.

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

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

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

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

Exercise 2.22. 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.

Application 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 $\overline{X_n}$ where X_i are i.i.d Ber(p). The accuracy of the estimate is measured by its mean-squared deviation $\sqrt{\text{Var}(\overline{X_n})} = \sqrt{p(1-p)}n^{-\frac{1}{2}}$. 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...

2.8. Modes of convergence

Definition 2.23. We say that $X_n \xrightarrow{P} X$ ("X_n converges to X in probability") if for any $\delta > 0$, $\mathbf{P}(|X_n - X| > \delta) \to 0$ as $n \to \infty$. Recall that we say that $X_n \xrightarrow{a.s.} X$ if $\mathbf{P}(\omega : \lim X_n(\omega) = X(\omega)) = 1$.

2.8.1. Almost sure and in probability.

Are they really different? Usually looking at Bernoulli random variables elucidates the matter.

Example 2.24. Suppose A_n are events in a probability space. Then one can see that

$$\text{(a) } \mathbf{1}_{A_n} \xrightarrow{P} 0 \Longleftrightarrow \lim_{n\to\infty} \mathbf{P}(A_n) = 0. \qquad \text{(b) } \mathbf{1}_{A_n} \xrightarrow{a.s.} 0 \Longleftrightarrow \mathbf{P}(\limsup_{n\to\infty} A_n) = 0.$$

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

Lemma 2.25. *Suppose X_n, X are r.v. on the same probability space. Then,*

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

PROOF. Note that analogous to the example,

$$\text{(a) } X_n \xrightarrow{P} X \quad \Longleftrightarrow \quad \forall \delta > 0,\ \lim_{n\to\infty} \mathbf{P}(|X_n - X| > \delta) = 0.$$

$$\text{(b) } X_n \xrightarrow{a.s.} X \quad \Longleftrightarrow \quad \forall \delta > 0,\ \mathbf{P}(\limsup_{n\to\infty} |X_n - X| > \delta) = 0.$$

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

Exercise 2.26.
1. If $X_n \xrightarrow{P} X$, show that $X_{n_k} \xrightarrow{a.s.} X$ for some subsequence.
2. Show that $X_n \xrightarrow{a.s.} X$ 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 $X_n Y_n \xrightarrow{P} XY$.

2.8.2. 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 2.27. *Suppose X_n, X are r.v. 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 $\delta > 0$

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

we see that $\limsup \mathbf{P}(X_n \le t) \le \mathbf{P}(X \le t+\delta)$ and $\liminf \mathbf{P}(X_n \le t) \ge \mathbf{P}(X \le t-\delta)$ for any $\delta > 0$. Taking $\delta \downarrow 0$ and letting t be a continuity point of the cdf of$X$, we immediately get $\lim \mathbf{P}(X_n \le t) = \mathbf{P}(X \le t)$. Thus, $X_n \xrightarrow{d} X$.

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

Exercise 2.28. (1) Give an example to show that convergence in distribution does not imply convergence in probability.

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

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

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

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

and thus $X_n \xrightarrow{P} X$. The converse is not true as the following example shows.

Example 2.29. Let $X_n = 2^n$ w.p $1/n$ and $X_n = 0$ w.p $1 - 1/n$. Then, $X_n \xrightarrow{P} 0$ but $\mathbf{E}[X_n^p] = n^{-1} 2^{np}$ for all n, and hence X_n does not go to zero in L^p (for any $p > 0$).

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

As example 2.29 shows, the converse is not true in full generality either. What goes wrong in this 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.

2.9. Uniform integrability

Definition 2.30. A family $\{X_i\}_{i \in I}$ of random variables is said to be *uniformly integrable* if given any $\epsilon > 0$, there exists A large enough so that $\mathbf{E}[|X_i| \mathbf{1}_{|X_i| > A}] < \epsilon$ for all $i \in I$.

Example 2.31. A finite set of integrable r.v.s is uniformly integrable. More interestingly, an L^p-bounded family with $p > 1$ is u.i. For, if $\mathbf{E}[|X_i|^p] \le M$ for all $i \in I$ for some $M > 0$, then $\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>t}] \le t^{-(p-1)}M$ which goes to zero as $t \to \infty$. Thus, given $\epsilon > 0$, one can choose t large so that $\sup_{i\in I}\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>t}] < \epsilon$.

This fails for $p = 1$ as the example 2.29 shows a family of L^1 bounded random variables that are not u.i. However, a u.i family must be bounded in L^1. To see this find $A > 0$ so that $\mathbf{E}[|X_i|\mathbf{1}_{|X_i|>A}] < 1$ for all i. Then, for any $i \in I$, we get $\mathbf{E}[|X_i|] = \mathbf{E}[|X_i|\mathbf{1}_{|X_i|<A}] + \mathbf{E}[|X_i|\mathbf{1}_{|X_i|\ge A}] \le A + 1$.

Exercise 2.32. If $\{X_i\}_{i\in I}$ and $\{Y_j\}_{j\in J}$ are both u.i, then $\{X_i + Y_j\}_{(i,j)\in I\times J}$ is u.i. What about the family of products, $\{X_iY_j\}_{(i,j)\in I\times J}$?

Lemma 2.33. *Suppose X_n, X are r.v. on the same probability space. Then, the following are equivalent.*

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

PROOF. If $Y_n = X_n - X$, then $X_n \xrightarrow{L^1} X$ iff $Y_n \xrightarrow{L^1} 0$, while $X_n \xrightarrow{P} X$ iff $Y_n \xrightarrow{P} 0$ and by the first part of exercise 2.32, $\{X_n\}$ 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 $Y_n \xrightarrow{L^1} 0$. Then we showed that $Y_n \xrightarrow{P} 0$. To show that $\{Y_n\}$ is u.i, let $\epsilon > 0$ and fix N_ϵ so that $\mathbf{E}[|Y_n|] < \epsilon$ for all $n \ge N_\epsilon$. Then, pick $A > 1$ so large that $\mathbf{E}[|Y_k|\mathbf{1}_{|Y_k|>A}] \le \epsilon$ for all $k \le N$. With the same A and any $k \ge N_\epsilon$, we get $\mathbf{E}[|Y_k|\mathbf{1}_{|Y_k|>A}] \le A^{-1}\mathbf{E}[|Y_k|] < \epsilon$ since $A > 1$ and $\mathbf{E}[|Y_k|] < \epsilon$. Thus we have found one A which works for all Y_k. Hence $\{Y_k\}$ is u.i.

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

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

For all $t \in [0, A]$, by assumption $\mathbf{P}(|Y_k| > t) \to 0$, while we also have $\mathbf{P}(|Y_k| > t) \le 1$ for all k and 1 is integrable on $[0, A]$. Hence, by DCT the first term goes to 0 as $k \to \infty$. Thus $\limsup \mathbf{E}[|Y_k|] \le \epsilon$ for any ϵ and it follows that $Y_k \xrightarrow{L^1} 0$. ■

Corollary 2.34. *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{L^1} X$ if and only if $\{X_n\}$ is u.i.*

To deduce convergence in mean from a.s convergence, we have so far always invoked DCT. As shown by Lemma 2.33 and corollary 2.34, uniform integrability is the sharp condition, so it must be weaker than the assumption in DCT. Indeed, if $\{X_n\}$ are dominated by an integrable Y, then whatever A works for Y in the u.i condition will work for the whole family $\{X_n\}$. Thus a dominated family is u.i., while the converse is false.

Remark 2.35. Like tightness of measures, uniform integrability is also related to a compactness question. On the space $L^1(\mu)$, apart from the usual topology coming from the norm, there is another one called *weak topology* (where $f_n \to f$ if and only if $\int f_n g d\mu \to \int f g d\mu$ for all $g \in L^\infty(\mu)$). The *Dunford-Pettis theorem* asserts that pre-compact subsets of $L^1(\mu)$ in this weak topology are precisely uniformly integrable subsets of $L^1(\mu)$! A similar question can be asked in L^p for $p > 1$ where weak topology means that $f_n \to f$ if and only if $\int f_n g d\mu \to \int f g d\mu$ for all $g \in L^q(\mu)$ where $q^{-1} + p^{-1} = 1$. Another part of Dunford-Pettis theorem asserts that pre-compact subsets of $L^p(\mu)$ in this weak topology are precisely those that are bounded in the $L^p(\mu)$ norm.

2.10. Strong law of large numbers

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

Theorem 2.36 (Kolmogorov's SLLN). *Let X_n be i.i.d with $\mathbf{E}[|X_1|] < \infty$. Then, as $n \to \infty$, we have $\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}[X_1]$.*

The proof of this theorem is somewhat complicated. First of all, we should ask if WLLN implies SLLN? From Lemma 2.27 we see that this can be done if $\mathbf{P}\left(|n^{-1}S_n - \mathbf{E}[X_1]| > \delta\right)$ is summable, for every $\delta > 0$. Even assuming finite variance $\mathrm{Var}(X_1) = \sigma^2$, Chebyshev's inequality only gives a bound of $\sigma^2\delta^{-2}n^{-1}$ for this probability and this is not summable. Since this is at the borderline of summability, if we assume that p^{th} moment exists for some $p > 2$, we may expect to carry out this proof. Suppose we assume that $\alpha_4 := \mathbf{E}[X_1^4] < \infty$ (of course 4 is not the smallest number bigger than 2, but how do we compute $\mathbf{E}[|S_n|^p]$ in terms of moments of X_1 unless p is an even integer?). Then, we may compute that (assume $\mathbf{E}[X_1] = 0$ wlog)

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

Thus $\mathbf{P}\left(|n^{-1}S_n| > \delta\right) \le n^{-4}\delta^{-4}\mathbf{E}[S_n^4] = O(n^{-2})$ which is summable, and by Lemma 2.27 we get the following weaker form of SLLN.

Theorem 2.37. *Let X_n be i.i.d with $\mathbf{E}[|X_1|^4] < \infty$. Then, $\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}[X_1]$ as $n \to \infty$.*

Now we return to the serious question of proving the strong law under first moment assumptions. The presentation of the following proof is adapted from a blog article of Terence Tao.

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

$$\frac{S_{N_1}}{N_2} \le \frac{S_n}{n} \le \frac{S_{N_2}}{N_1} \quad \text{if } N_1 \le n \le N_2. \tag{2.10}$$

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

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

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{S_{n_j}}{n_j} - \mathbf{E}[X_1]\right| > 2\delta\right) &\le \mathbf{P}\left(\left|\frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1]\right| + \left|\frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1]\right| > 2\delta\right) \\
&\le \mathbf{P}\left(\left|\frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1]\right| > \delta\right) + \mathbf{P}\left(\left|\frac{S_{n_j}^Z}{n_j} - \mathbf{E}[Z_1]\right| > \delta\right) \\
&\le \mathbf{P}\left(\left|\frac{S_{n_j}^Y}{n_j} - \mathbf{E}[Y_1]\right| > \delta\right) + \mathbf{P}\left(\frac{S_{n_j}^Z}{n_j} \ne 0\right). \qquad (2.11)
\end{aligned}$$

We shall show that both terms in (2.11) are summable over j. The first term can be bounded by Chebyshev's inequality

$$\text{(2.12)} \qquad \mathbf{P}\left(\,\Big|\,\frac{S^Y_{n_j}}{n_j} - \mathbf{E}[Y_1]\,\Big| > \delta\right) \le \frac{1}{\delta^2 n_j}\mathbf{E}[Y_1^2] = \frac{1}{\delta^2 n_j}\mathbf{E}[X_1^2 \mathbf{1}_{X_1 \le n_j}].$$

while the second term is bounded by the union bound

$$\text{(2.13)} \qquad \mathbf{P}\left(\frac{S^Z_{n_j}}{n_j} \neq 0\right) \le n_j \mathbf{P}(X_1 > n_j).$$

The right hand sides of (2.12) and (2.13) are both summable. To see this, observe that for any positive x, there is a unique k such that $n_k < x \le n_{k+1}$, and then

$$\text{(2.14)} \qquad (a)\ \sum_{j=1}^{\infty} \frac{1}{n_j} x^2 \mathbf{1}_{x \le n_j} \le x^2 \sum_{j=k+1}^{\infty} \frac{1}{\lambda^j} \le C_\lambda x. \qquad (b)\ \sum_{j=1}^{\infty} n_j \mathbf{1}_{x > n_j} \le \sum_{j=1}^{k} \lambda^j \le C_\lambda x.$$

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

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

As $\mathbf{E}[X_1] < \infty$, the probabilities on the left hand side of (2.12) and (2.13) are summable in j, and hence it also follows that $\mathbf{P}\left(\,\big|\,\frac{S_{n_j}}{n_j} - \mathbf{E}[X_1]\,\big| > 2\delta\right)$ is summable. This happens for every $\delta > 0$ and hence Lemma 2.27 implies that $\frac{S_{n_j}}{n_j} \overset{a.s.}{\to} \mathbf{E}[X_1]$ a.s. This proves the claim.

Step 3: Fix $\lambda > 1$. Then, for any n, find k such that $\lambda^k < n \le \lambda^{k+1}$, and then, from (2.10) we get

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

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

2.11. Kolmogorov's zero-one law

We saw that in strong law the limit of $n^{-1}S_n$ turned out to be constant, while a priori, it could well have been random. This is a reflection of the following more general and surprising fact.

Definition 2.39. Let $\mathscr{F}_n$ be sub-sigma algebras of $\mathscr{F}$. Then the tail σ-algebra of the sequence $\mathscr{F}_n$ is defined to be $\mathscr{T} := \cap_n \sigma(\cup_{k \ge n} \mathscr{F}_k)$. For a sequence of random variables $X_1, X_2, \ldots$, the tail sigma algebra is the tail of the sequence $\sigma(X_n)$.

We also say that a σ-algebra is trivial (w.r.t a probability measure) if $\mathbf{P}(A)$ equals 0 or 1 for every A in the sig-algebra.

Theorem 2.40 (Kolmogorov's zero-one law). *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space.*

1. *If $\mathscr{F}_n$ is a sequence of independent sub-sigma algebras of $\mathscr{F}$, then the tail sig-algebra is trivial.*

(2) *If X_n are independent random variables, and A is a tail event, then* $\mathbf{P}(A)$ *is* 0 *or* 1 *for every* $A \in \mathcal{T}$.

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

Exercise 2.41. Let X_i be independent random variables. Which of the following random variables must necessarily be constant almost surely? $\limsup X_n$, $\liminf X_n$, $\limsup n^{-1}S_n$, $\liminf S_n$.

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

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

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

In pathbreaking work, it was proved by 1980s that $\mathbf{P}_p(A) = 0$ if $p \le \frac{1}{2}$ and $\mathbf{P}_p(A) = 1$ if $p > \frac{1}{2}$.

The same problem can be considered on $\mathbb{Z}^3$, keeping each edge with probability p and deleting it with probability $1-p$, independently of all other edges. It is again known (and not too difficult to show) that there is some number $p_c \in (0,1)$ such that $\mathbf{P}_p(A) = 0$ if $p < p_c$ and $\mathbf{P}_p(A) = 1$ if $p > p_c$. The value of p_c is not known, and more importantly, it is not known whether $\mathbf{P}_{p_c}(A)$ is 0 or 1!

2.12. The law of iterated logarithm

If $a_n \uparrow \infty$, then the reasoning in the previous section applies and $\limsup a_n^{-1}S_n$ is constant a.s. This motivates the following natural question.

Question: Let X_i be i.i.d random variables taking values ± 1 with equal probability. Find a_n so that $\limsup \frac{S_n}{a_n} = 1$ a.s.

The question is about the growth rate of sums of random independent ± 1s. We know that $n^{-1}S_n \overset{a.s.}{\to} 0$ by the SLLN, hence, $a_n = n$ is "too much". What about n^α. Applying Hoeffding's inequality (proved in the next section), we see that $\mathbf{P}(n^{-\alpha}S_n > t) \le \exp\{-\frac{1}{2}t^2n^{2\alpha-1}\}$. If $\alpha > \frac{1}{2}$, this is a summable sequence for any $t > 0$, and therefore $\mathbf{P}(n^{-\alpha}S_n > t \text{ i.o.}) = 0$. That is $\limsup n^{-\alpha}S_n \overset{a.s.}{\to} 0$ for $\alpha > \frac{1}{2}$. What about $\alpha = \frac{1}{2}$? One can show that $\limsup n^{-\frac{1}{2}}S_n = +\infty$ a.s, which means that $\sqrt{n}$ is too slow compared to S_n. So the right answer is larger than $\sqrt{n}$ but smaller than $n^{\frac{1}{2}+\epsilon}$ for any $\epsilon > 0$. The sharp answer, due to Khinchine is a crown jewel of probability theory!

Result 2.42 (Khinchine's law of iterated logarithm). Let X_i be i.i.d with zero mean and finite variance $\sigma^2 = 1$ (without loss of generality). Then,

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

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

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

Exercise 2.43. Let X_i be i.i.d random variables taking values ± 1 with equal probability. Show that $\limsup\limits_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} \le 1$, almost surely.

2.13. Hoeffding's inequality

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

Lemma 2.44. (Hoeffding's inequality). *Let $X_1,\dots,X_n$ be independent, and assume that $|X_k| \le d_k$ w.p.1. For simplicity assume that $\mathbf{E}[X_k] = 0$. Then, for any $n \ge 1$ and any $t > 0$,*

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

Remark 2.45. The boundedness assumption on X_ks is essential. That $\mathbf{E}[X_k] = 0$ is for convenience. If we remove that assumption, note that $Y_k = X_k - \mathbf{E}[X_k]$ satisfy the assumptions of the theorem, except that we can only say that $|Y_k| \le 2d_k$ (because $|X_k| \le d_k$ implies that $|\mathbf{E}[X_k]| \le d_k$ and hence $|X_k - \mathbf{E}[X_k]| \le 2d_k$). Thus, applying the result to Y_ks, we get

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

PROOF. Without loss of generality, take $\mathbf{E}[X_k] = 0$. Now, if $|X| \le d$ w.p.1, and $\mathbf{E}[X] = 0$, by convexity of exponential on $[-1,1]$, we write for any $\lambda > 0$

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

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

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

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

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

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

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

(1) **(Large deviation regime)** Take $t = n\delta$ to get

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

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

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

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

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

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

2.14. Random series with independent terms

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

Example 2.46. Let a_n be i.i.d with finite mean. Important examples are $a_n \sim N(0,1)$ or $a_n = \pm 1$ with equal probability. Then, define $f(z) = \sum_n a_n z^n$. What is the radius of convergence of this series? From the formula for radius of convergence

$R = \left(\limsup_{n\to\infty} |a_n|^{\frac{1}{n}}\right)^{-1}$, it is easy to find that the radius of convergence is exactly 1 (a.s.) **[Exercise]**. Thus we get a random analytic function on the unit disk.

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

Theorem 2.47 (Khinchine). *Let X_n be independent random variables with finite second moment. Assume that $\mathbf{E}[X_n] = 0$ for all n and that $\sum_n Var(X_n) < \infty$.*

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

$$\mathbf{P}(|S_n - S_N| > \delta \text{ for some } n \ge N) = \lim_{m\to\infty} \mathbf{P}(|S_n - S_N| > \delta \text{ for some } N \le n \le N+m). \tag{2.15}$$

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

Invoking 2.50, we get

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

The right hand side goes to zero as $N \to \infty$. Thus, from (2.15), we conclude that for any $\delta > 0$,

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

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

Remark 2.48. What to do if the assumptions are not exactly satisfied? First, suppose that $\sum_n \text{Var}(X_n) < \infty$ but $\mathbf{E}[X_n]$ may not be zero. Then, we can write $\sum X_n = \sum (X_n - \mathbf{E}[X_n]) + \sum \mathbf{E}[X_n]$. The first series on the right satisfies the assumptions of Theorem thm:convergenceofrandomseries and hence converges a.s. Therefore, $\sum X_n$ will then converge a.s if the deterministic series $\sum_n \mathbf{E}[X_n]$ converges and conversely, if $\sum_n \mathbf{E}[X_n]$ does not converge, then $\sum X_n$ diverges a.s.

Next, suppose we drop the finite variance condition too. Now X_n are arbitrary independent random variables. We reduce to the previous case by truncation. Suppose we could find some $A > 0$ such that $\mathbf{P}(|X_n| > A)$ is summable. Then set $Y_n = X_n \mathbf{1}_{|X_n| > A}$. By Borel-Cantelli, almost surely, $X_n = Y_n$ for all but finitely many n and hence $\sum X_n$ converges if and only if $\sum Y_n$ converges. Note that Y_n has finite variance. If $\sum_n \mathbf{E}[Y_n]$ converges and $\sum_n \text{Var}(Y_n) < \infty$, then it follows from the argument in the previous paragraph and Theorem 2.47 that $\sum Y_n$ converges a.s. Thus we have proved

Lemma 2.49 (Kolmogorov's three series theorem - part 1). *Suppose X_n are independent random variables. Suppose for some $A > 0$, the following hold with $Y_n := X_n \mathbf{1}_{|X_n| \le A}$.*

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

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

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

2.15. Kolmogorov's maximal inequality

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

Lemma 2.50 (Kolmogorov's maximal inequality). *Let X_n be independent random variables with finite variance and $\mathbf{E}[X_n] = 0$ for all n. Then, $\mathbf{P}(\max_{k\le n}|S_k| > t) \le t^{-2}\sum_{k=1}^n Var(X_k)$.*

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

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

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

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

$$\begin{aligned}\mathbf{E}[S_n^2] &= \mathbf{E}\left[(S_\tau + (S_n - S_\tau))^2\right]\\ &= \mathbf{E}[S_\tau^2] + \mathbf{E}\left[(S_n - S_\tau)^2\right] - 2\mathbf{E}[S_\tau(S_n - S_\tau)]\\ &\ge \mathbf{E}[S_\tau^2] - 2\mathbf{E}[S_\tau(S_n - S_\tau)]. \qquad (2.16)\end{aligned}$$

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

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

In the second line we used the fact that $S_k\mathbf{1}_{\tau=k}$ depends on $X_1, \ldots, X_k$ only, while $S_n - S_k$ depends only on $X_{k+1}, \ldots, X_n$. Putting this result into (2.16), we get the $\mathbf{E}[S_n^2] \ge \mathbf{E}[S_\tau^2]$ which together with Chebyshev's gives us

$$\mathbf{P}(\max_{k\le n} S_k > t) \le t^{-2}\mathbf{E}[S_n^2]. \qquad \blacksquare$$

2.16. Central limit theorem - statement, heuristics and discussion

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

Theorem 2.51. *Let X_n be i.i.d with mean μ and finite variance σ^2.Then, $\frac{S_n-n\mu}{\sigma\sqrt{n}}$ converges in distribution to $N(0,1)$.*

Informally, letting χ denote a standard Normal variable, we may write $S_n \approx n\mu + \sigma\sqrt{n}\chi$. This means, the distribution of S_n is hardly dependent on the distribution of X_1 that we started with, except for the two parameter of mean and variance. This is a statement about a remarkable symmetry!

Heuristics: Why should one expect such a statement to be true? Without losing generality, let us take $\mu = 0$ and $\sigma^2 = 1$. As $\mathbf{E}\left[\left(\frac{S_n}{\sqrt{n}}\right)^2\right] = 1$ is bounded, we see that $n^{-\frac{1}{2}}S_n$ is tight, and hence has weakly convergent subsequences. Let us make a leap of faith and suppose that $\frac{S_n}{\sqrt{n}}$ converges in distribution. To what? Let Y be a random variable with the limiting distribution. Then, $(2n)^{-\frac{1}{2}}S_{2n} \xrightarrow{d} Y$ 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, by an earlier exercise, we also get

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

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

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

Thus we must have $Y_1 + Y_2 \overset{d}{=} \sqrt{2}Y$. Therefore, if $\psi(t)$ denotes the characteristic function of Y, then

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

Similarly, for any $k \geq 1$, we can prove that $Y_1 + \ldots Y_k \overset{d}{=} \sqrt{k}Y$, where Y_i are i.i.d copies of Y and hence $\psi(t) = \psi(tk^{-1/2})^k$. From this, by standard methods, one can deduce that $\psi(t) = e^{-at^2}$ for some $a > 0$ (**exercise**). By uniqueness of characteristic functions, $Y \sim N(0,2a)$. Since we expect $\mathbf{E}[Y^2] = 1$, we must get $N(0,1)$.

It is an instructive exercise to prove the CLT by hand for specific distributions. For example, suppose X_i are i.i.d exp(1) so that $\mathbf{E}[X_1] = 1$ and $\text{Var}(X_1) = 1$. Then $S_n \sim \Gamma(n,1)$ and hence $\frac{S_n-n}{\sqrt{n}}$ has density

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

by elementary calculations. By an earlier exercise convergence of densities implies convergence in distribution and thus we get CLT for sums of exponential random variables.

Exercise 2.52. Prove the CLT for $X_1 \sim \text{Ber}(p)$. Note that this also implies CLT for $X_1 \sim \text{Bin}(k,p)$.

2.17. Central limit theorem - Proof using characteristic functions

We shall use characteristic functions to prove the CLT. To make the main idea of the proof transparent, we first prove a restricted version assuming third moments. Once the idea is clear, we prove a much more general version later which will also give Theorem 2.51. We shall need the following fact.

Exercise 2.53. Let z_n be complex numbers such that $nz_n \to z$. Then, $(1+z_n)^n \to e^z$.

Theorem 2.54. *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 $N(0,1)$.*

Proof. By Lévy's continuity theorem, it suffices to show that the characteristic functions of $n^{-\frac{1}{2}}S_n$ converge to the of $N(0,1)$. Note that

$$\psi_n(t) := \mathbf{E}\left[e^{itS_n/\sqrt{n}}\right] = \psi\left(\frac{t}{\sqrt{n}}\right)^n$$

where ψ is the c.f of X_1. Use Taylor expansion

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

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

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

Clearly, $|R_n(t)| \le Cn^{-3/2}$ for a constant C (that depends on t but not n). Hence $nR_n(t) \to 0$ and by Exercise 2.53 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 c.f of $N(0,1)$. ■

2.18. CLT 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.

Theorem 2.55 (Lindeberg Feller CLT). *Suppose $X_{n,k}$, $k \le n$, $n \ge 1$, are random variables. We assume that*

(1) *For each n, the random variables $X_{n,1},\dots,X_{n,n}$ are defined on the same probability space, are* independent *and have finite second moments.*
(2) $\mathbf{E}[X_{n,k}]=0$ *and* $\sum_{k=1}^n \mathbf{E}[X_{n,k}^2] \to \sigma^2$, *as* $n \to \infty$.
(3) *For any* $\delta > 0$, *we have* $\sum_{k=1}^n \mathbf{E}[X_{n,k}^2 \mathbf{1}_{|X_{n,k}|>\delta}] \to 0$ *as* $n \to \infty$.

Corollary 2.56. *Let X_n be i.i.d, having zero mean and unit variance. Then, $\frac{S_n}{\sqrt{n}}$ converges in distribution to $N(0,1)$.*

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

Now we prove the Lindeberg-Feller CLT. As in the previous section, we need a fact comparing a product to an exponential.

Exercise 2.57. If z_k, w_k are complex numbers with absolute value bounded by θ, then $\left|\prod_{k=1}^n z_k - \prod_{k=1}^n w_k\right| \le \theta^{n-1}\sum_{k=1}^n |z_k - w_k|$.

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

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

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

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

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

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

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

$$|R_{n,k}(t)| \le t^2\mathbf{E}\left[\mathbf{1}_{|X_{n,k}|>\delta}X_{n,k}^2\right] + \frac{|t|^3\delta}{6}\mathbf{E}\left[X_{n,k}^2\right]. \tag{2.17}$$

We want to apply Exercise 2.57 to $z_k = \mathbf{E}\left[e^{itX_{n,k}}\right]$ and $w_k = 1 - \frac{1}{2}\sigma_{n,k}^2 t^2$. Clearly $|z_k| \le 1$ by properties of c.f. If we prove that $\max\limits_{k\le n}\sigma_{n,k}^2 \to 0$, then it will follow that $|w_k| \le 1$ and hence with $\theta = 1$ in Exercise 2.57, we get

$$\begin{aligned} \limsup_{n\to\infty} \left|\prod_{k=1}^n \mathbf{E}\left[e^{itX_{n,k}}\right] - \prod_{k=1}^n \left(1 - \frac{1}{2}\sigma_{n,k}^2 t^2\right)\right| &\le \limsup_{n\to\infty} \sum_{k=1}^n |R_{n,k}(t)| \\ &\le \frac{1}{6}|t|^3\sigma^2\delta \quad \text{(by 2.17)} \end{aligned}$$

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

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

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

$$\lim_{n\to\infty}\prod_{k=1}^{n}\mathbf{E}\left[e^{itX_{n,k}}\right] = \lim_{n\to\infty}\prod_{k=1}^{n}\left(1-\frac{1}{2}\sigma_{n,k}^2 t^2\right). \tag{2.18}$$

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

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

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

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

From 2.18 and Lévy's continuity theorem, we get $\sum_{k=1}^{n}X_{n,k}\xrightarrow{d} N(0,\sigma^2)$. ■

2.19. Limits of sums of 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 location (by $n\mathbf{E}[X_1]$) and scaling (by $\sqrt{n}$). What if we drop the assumption of second moments? Let us first consider the case of Cauchy random variables to see that such results may be expected in general.

Example 2.58. Let X_i be i.i.d Cauchy(1), with density $\frac{1}{\pi(1+x^2)}$. Then, one can check that $\frac{S_n}{n}$ has exactly the same Cauchy distribution! Thus, to get distributional convergence, we just write $\frac{S_n}{n}\xrightarrow{d} C_1$. If X_i were i.i.d with density $\frac{a}{\pi(a^2+(x-b)^2)}$ (which can be denoted $C_{a,b}$ with $a>0$, $b\in\mathbb{R}$), then $\frac{X_i-b}{a}$ are i.i.d C_1, and hence, we get

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

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

This raises the following questions.

(1) For general i.i.d sequences, how are the location and scaling parameter determined, so that $b_n^{-1}(S_n-a_n)$ converges in distribution to a non-trivial measure on the line?
(2) What are the possible limiting distributions?
(3) What are the *domains of attraction* for each possible limiting distribution, e.g., for what distributions on X_1 do we get $b_n^{-1}(S_n-a_n)\xrightarrow{d} C_1$?

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

2.20. Poisson convergence for rare events

Theorem 2.59. *Let $A_{n,k}$ be events in a probability space. Assume that*

(1) *For each n, the events $A_{n,1},\dots,A_{n,n}$ are independent.*
(2) $\sum\limits_{k=1}^{n} \mathbf{P}(A_{n,k}) \to \lambda \in (0,\infty)$ *as* $n \to \infty$.
(3) $\max\limits_{k\le n} \mathbf{P}(A_{n,k}) \to 0$ *as* $n \to \infty$.

Then $\sum_{k=1}^{n} \mathbf{1}_{A_{n,k}} \xrightarrow{d}$ *Pois*(λ).

Proof. Let $p_{n,k} = \mathbf{P}(A_{n,k})$ and $X_n = \sum_{k=1}^{n} \mathbf{1}_{A_{n,k}}$. By assumption (3), for large enough n, we have $p_{n,k} \le 1/2$ and hence

$$e^{-p_{n,k}-p_{n,k}^2} \le 1 - p_{n,k} \le e^{-p_{n,k}}. \tag{2.19}$$

Thus, for any fixed $\ell \ge 0$,

$$\begin{aligned}\mathbf{P}(X_n = \ell) &= \sum_{S\subseteq[n]:|S|=\ell} \mathbf{P}\left(\bigcap_{j\in S} A_j \bigcap_{j\notin S} A_j^c\right)\\ &= \sum_{S\subseteq[n]:|S|=\ell}\prod_{j\notin S}(1-p_{n,j})\prod_{j\in S} p_{n,j}\\ &= \prod_{j=1}^{n}(1-p_{n,j}) \sum_{S\subseteq[n]:|S|=\ell}\prod_{j\in S}\frac{p_{n,j}}{1-p_{n,j}}. \qquad (2.20)\end{aligned}$$

By assumption $\sum_k p_{n,k} \to \lambda$. Together with the third assumption, this implies that $\sum_k p_{n,k}^2 \to 0$. Thus, using 2.19 we see that

$$\prod_{j=1}^{n}(1-p_{n,j}) \longrightarrow e^{-\lambda}. \tag{2.21}$$

Let $q_{n,j} = p_{n,j}/(1-p_{n,j})$. The second factor in 2.20 is

$$\begin{aligned}\sum_{S\subseteq[n]:|S|=\ell}\prod_{j\in S} q_{n,j} &= \frac{1}{k!}\sum_{\substack{j_1,\dots j_\ell\\ \text{distinct}}}\prod_{i=1}^{\ell} q_{n,j_i}\\ &= \frac{1}{\ell!}\left(\sum_j q_{n,j}\right)^{\ell} - \frac{1}{\ell!}\sum_{\substack{j_1,\dots j_\ell\\ \text{not distinct}}}\prod_{i=1}^{\ell} q_{n,j_i}.\end{aligned}$$

The first term converges to $\lambda^\ell/\ell!$. To show that the second term goes to zero, divide it into cases where $j_a = j_b$ with $a,b \le \ell$ being chosen in one of $\binom{\ell}{2}$ ways. Thus we get

$$\sum_{\substack{j_1,\dots j_\ell\\ \text{not distinct}}}\prod_{i=1}^{\ell} q_{n,j_i} \le \binom{\ell}{2}\left(\sum_{j=1}^{n} q_j\right)^{\ell-1}\max_j q_{n,j} \to 0$$

because like $p_{n,j}$ we also have $\sum_j q_{n,j} \to \lambda$ and $\max_j q_{n,j} \to 0$. Put this together with 2.20 and 2.21 to conclude that

$$\mathbf{P}(X_n = \ell) \to e^{-\lambda}\frac{\lambda^\ell}{\ell!}.$$

Thus $X_n \xrightarrow{d} \text{Pois}(\lambda)$. ■

Exercise 2.60. Use characteristic functions to give an alternate proof of Theorem 2.59.

As a corollary, $\text{Bin}(n, p_n) \to \text{Pois}(\lambda)$ if $n \to \infty$ and $p_n \to 0$ in such a way that $np_n \to \lambda$. Contrast this with the Binomial convergence to Normal (after location and scale change) if $n \to \infty$ but p is held fixed.

CHAPTER 3

Brownian motion

In this chapter we introduce a very important probability measure, called *Wiener measure* on the space $C[0,\infty)$ of continuous functions on $[0,\infty)$. We shall barely touch the surface of this very deep subject. A $C[0,\infty)$-valued random variable whose distribution is the Wiener measure, is called *Brownian motion*. First we recall a few basic facts about the space of continuous functions.

3.1. Brownian motion and Winer measure

Let $C[0,1]$ and $C[0,\infty)$ be the space of real-valued continuous functions on $[0,1]$ and $[0,\infty$, respectively. On $C[0,1]$ the sup-norm $\|f-g\|_{\sup}$ defines a metric. On $C[0,\infty)$ a metric may be defined by setting

$$d(f,g)=\sum_{n=1}^{\infty}\frac{1}{2^n}\frac{\|f-g\|_{\sup[0,n]}}{1+\|f-g\|_{\sup[0,n]}}.$$

It is easy to see that $f_n \to f$ in this metric if and only if f_n converges to f uniformly on compact subsets of $\mathbb{R}$. The metric itself does not matter to us, but the induced topology does, and so does the fact that this topology can be induced by a metric that makes the space complete and separable. In this section, we denote the corresponding Borel σ-algebras by $\mathscr{B}_1$ and $\mathscr{B}_\infty$. Recall that a cylinder set is a set of the form

$$C=\{f\ :\ f(t_1)\in A_1,\dots,f(t_k)\in A_k\} \tag{3.1}$$

for some $t_1<t_2<\ldots<t_k$, $A_j\in\mathscr{B}(\mathbb{R})$ and some $k\ge 1$. Here is a simple exercise.

Exercise 3.1. Show that $\mathscr{B}_1$ and $\mathscr{B}_\infty$ are generated by finite dimensional cylinder sets (where we restrict $t_k\le 1$ in case of $C[0,1]$).

As a consequence of the exercise, if two Borel probability measures agree on all cylinder sets, then they are equal. We define Wiener measure by specifying its probabilities on cylinder sets. Let ϕ_{σ^2} denote the density function of $N(0,\sigma^2)$.

Definition 3.2. *Wiener measure* μ *is a probability measure on* $(C[0,\infty),\mathscr{B}_\infty)$ *such that*

$$\mu(C)=\int\limits_{A_1}\ldots\int\limits_{A_k}\phi_{t_1}(x_1)\phi_{t_2-t_1}(x_2)\ldots\phi_{t_k-t_{k-1}}(x_k)dx_1\ldots dx_k$$

for every cylinder set $C=\{f\ :\ f(t_1)\in A_1,\dots,f(t_k)\in A_k\}$. Clearly, if Wiener measure exists, it is unique.

We now define Brownian motion. Let us introduce the term *Stochastic process* to indicate a collection of random variables indexed by an arbitrary set - usually, the indexing set is an interval of the real line or of integers, and the indexing variable may have the interpretation of 'time'.

Definition 3.3. Let $(\Omega,\mathscr{F},\mathbf{P})$ be any probability space. A stochastic process $(B_t)_{t\ge 0}$ indexed by $t \ge 0$ is called *Brownian motion* if

1. $B_0 = 0$ w.p.1.
2. (**Finite dimensional distributions**). For any $k \ge 1$ and any $0 \le t_1 < t_2 < \ldots < t_k$, the random variables $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_k} - B_{t_{k-1}}$ are independent, and for any $s < t$ we have $B_t - B_s \sim N(0, t-s)$.
3. (**Continuity of sample paths**). For a.e. $\omega \in \Omega$, the function $t \to B_t^\omega$ is a continuous.

In both definitions, we may restrict to $t \in [0,1]$, and the corresponding measure on $C[0,1]$ is also called Wiener measure and the corresponding stochastic process $(B_t)_{0\le t\le 1}$ is also called Brownian motion.

The following exercise shows that Brownian motion and Wiener measure are two faces of the same coin (just like a normal random variable and the normal distribution).

Exercise 3.4.

1. Suppose $(B_t)_{t\ge 0}$ is a Brownian motion on some probability space $(\Omega,\mathscr{F},\mathbf{P})$. Define a map $B : \Omega \to C[0,\infty)$ by setting $B(\omega)$ to be the function whose value at t is given by $B_t(\omega)$. Show that B is a measurable function, and the induced measure $\mathbf{P}B^{-1}$ on $(C[0,\infty),\mathscr{B}_\infty)$ is the Wiener measure.
2. Conversely, suppose the Wiener measure μ exists. Let $\Omega = C[0,\infty)$, $\mathscr{F} = \mathscr{B}_\infty$ and $\mathbf{P} = \mu$. For each $t \ge 0$, define the r.v $B_t : \Omega \to C[0,\infty)$ by $B_t^\omega = \omega(t)$ for $\omega \in C[0,\infty)$. Then, show that the collection $(B_t)_{t\ge 0}$ is a Brownian motion.

The exercise shows that if the Wiener measure exists, then Brownian motion exists, and conversely. But it is not at all clear that either Brownian motion or Wiener measure exists.

3.2. Some continuity properties of Brownian paths - Negative results

We construct Brownian motion in the next section. Now, assuming the existence, we shall see some very basic properties of Brownian paths. We can ask many questions about the sample paths. We just address some basic questions about the continuity of the sample paths.

Brownian paths are quite different from the 'nice functions' that we encounter regularly. For example, *almost surely*, there is no interval on which Brownian motion is increasing (or decreasing)! To see this, fix an interval $[a,b]$ and points $a = t_0 < t_1 < \ldots < t_k = b$. If W was increasing on $[a,b]$, we must have $W(t_i) - W(t_{i-1}) \ge 0$ for $i = 1,2\ldots,k$. These are independent events that have probability 1/2 each, whence the probability for W to be increasing on $[a,b]$ is at most 2^{-k}. As k is arbitrary, the probability is 0. Take intersection over all rational $a < b$ to see that *almost surely*, there is no interval on which Brownian motion is increasing.

Warm up question: Fix $t_0 \in [0,1]$. What is the probability that W is differentiable at t_0?

Answer: Let $I_n = [t_0+2^{-n}, t_0+2^{-n+1}]$. If $W'(t_0)$ exists, then $\lim_{n\to\infty} 2^n \sum_{k>n} \Delta W(I_k) = W'(t_0)$. Whether the limit on the left exists, is a tail event of the random variables $\Delta W(I_n)$ which are independent, by properties of W. By Kolmogorov's law, the event that this limit exists has probability 0 or 1. If the probability was 1, we would have

$2^n(W(t_0+2^{-n})-W(t_0)) \overset{a.s.}{\to} W'(t_0)$, and hence also in distribution. But for each n, observe that $2^n(W(t_0+2^{-n})-W(t_0)) \sim N(0,2^n)$ which cannot converge in distribution. Hence the probability that W is differentiable at t_0 is zero!

Remark 3.5. Is the event $A_{t_0} := \{W \text{ is differentiable at } t_0\}$ measurable? We haven't shown this! Instead what we showed was that it is contained in a measurable set of Wiener measure zero. In other words, if we complete the Borel σ-algebra on $C[0,1]$ w.r.t Wiener measure (or complete whichever probability space we are working in), then under the completion A_{t_0} is measurable and has zero measure. This will be the case for the other events that we consider.

For any $t \in [0,1]$, we have shown that W is a.s. not differentiable at t. Can we claim that W is nowhere differentiable a.s? No! $\mathbf{P}(A_t) = 0$ and hence $\mathbf{P}(\bigcup_{t\in\mathbb{Q}} A_t) = 0$ but we cannot say anything about uncountable unions. For example,$\mathbf{P}(W_t = 1) = 0$ for any t. But $\mathbf{P}(W_t = 1 \text{ for some t}) > 0$ because $\{W_t = 1 \text{ for some t}\} \supset \{W_1 > 1\}$ and $\mathbf{P}(W_1 > 1) > 0$.

Notation: For $f \in C[0,1]$ and an interval $I = [a,b] \subseteq [0,1]$, let $\Delta f(I) := |f(b)-f(a)|$.

Theorem 3.6 (Paley, Wiener and Zygmund). *Almost surely, W is nowhere differentiable.*

PROOF. (Dvoretsky, Erdös and Kakutani). Fix $M < \infty$. For any $n \ge 1$, consider the event $A_n^{(M)} := \{\exists 0 \le j \le 2^n - 3 : |\Delta W(I_{n,p})| \le M2^{-n} \text{ for } p = j, j+1, j+2\}$. $\Delta W(I_{n,p})$ has $N(0,2^{-n})$ distribution, hence $\mathbf{P}(|\Delta W(I_{n,p})| \le M2^{-n}) = \mathbf{P}(|\chi| \le 2^{-n/2}) \le 2^{-n/2}$. By independence of $\Delta W(I_{n,j})$ fo rdistinct j, we get

$$\mathbf{P}(A_n^{(M)}) \le (2^n - 2)\left(2^{-n/2}\right)^3 \le 2^{-n/2}.$$

Therefore $\mathbf{P}(A_n^{(M)} \text{ i.o.}) = 0$.

Let $f \in C[0,1]$. Suppose f is differentiable at some point t_0 with $|f'(t_0)| \le M/2$. Then, for some $\delta > 0$, we have $\Delta f([a,b]) \le |f(b)-f(t)| + |f(t)-f(a)| \le M|b-a|$ for all $a,b \in [t_0-\delta, t_0+\delta]$. In particular, for large n so that $2^{-n+2} < \delta$, there will be three consecutive dyadic intervals $I_{n,j}, I_{n,j+1}, I_{n,j+2}$ that are contained in $[t_0-\delta, t_0+\delta]$. and for each of $p = j, j+1, j+2$ we have $\Delta f(I_{n,p}) \le M2^{-n}$.

Thus the event that W is differentiable somewhere with a derivative less than M (in absolute value), is contained in $\{A_n^{(M)} \text{ i.o.}\}$ which has probability zero. Since this is true for every M, taking union over integer $M \ge 1$, we get the statement of the theorem. ■

Exercise 3.7. For $f \in C[0,1]$, we say that t_0 is an α-Hölder continuity point of f if $\limsup_{s\to t} \frac{|f(s)-f(t)}{|s-t|^\alpha} < \infty$. Show that *almost surely*, Brownian motion has no α-Hölder continuity point for any $\alpha > \frac{1}{2}$.

We next show that Brownian motion is everywhere α-Hölder continuous for any $\alpha < \frac{1}{2}$. Hence together with the above exercie, we have made almost the best possible statement. Except, these statements do not answer what happens for $\alpha = \frac{1}{2}$.

3.3. Some continuity properties of Brownian paths - Positive results

To investigate Hölder continuity for $\alpha < 1/2$, we shall need the following lemma which is very similar to Kolmogorov's maximal inequality (the difference is that there we only assumed second moments, but here we assume that the X_i are normal and hence we get a stronger conclusion).

Lemma 3.8. *Let $X_1,\dots,X_n$ be i.i.d $N(0,\sigma^2)$. Then,* $\mathbf{P}\left(\max\limits_{k\le n} S_k \ge x\right) \le e^{-\frac{x^2}{2\sigma^2 n}}$.

PROOF. Let $\tau = \min\{k : S_k \ge x\}$ and set $\tau = n$ if there is no such k. Then, $\mathbf{P}(\max_{k\le n} S_k \ge x) = \mathbf{P}(S_\tau \ge x) \le e^{-\theta x}\mathbf{E}\left[e^{\theta S_\tau}\right]$ for any $\theta > 0$. Recall that for $\mathbf{E}\left[e^{\theta N(0,b)}\right] = e^{\theta^2 b^2/2}$. Thus, as in the proof of Kolmogorov's inequality

$$\begin{aligned} e^{\frac{1}{2}\theta^2\sigma^2 n} &= \mathbf{E}\left[e^{\theta S_n}\right] = \sum_{k=1}^{n}\mathbf{E}\left[\mathbf{1}_{\tau=k}e^{\theta S_k}e^{\theta(S_n-S_k)}\right] = \sum_{k=1}^{n}\mathbf{E}\left[\mathbf{1}_{\tau=k}e^{\theta S_k}\right]\mathbf{E}\left[e^{\theta(S_n-S_k)}\right] \\ &= \sum_{k=1}^{n}\mathbf{E}\left[\mathbf{1}_{\tau=k}e^{\theta S_k}\right]e^{\frac{1}{2}\theta^2(n-k)\sigma^2} \ge \sum_{k=1}^{n}\mathbf{E}\left[\mathbf{1}_{\tau=k}e^{\theta S_k}\right] = \mathbf{E}\left[e^{\theta S_\tau}\right]. \end{aligned}$$

Thus, $\mathbf{P}(S_\tau > x) \le e^{-\theta x}e^{\frac{1}{2}\theta^2\sigma^2 n}$. Set $\theta = \frac{x}{n\sigma^2}$ to get the desired inequality. ■

Corollary 3.9. *Let W be Brownian motion. Then, for any $a < b$, we have*

$$\mathbf{P}\left(\max_{s,t\in[a,b]} |W_t - W_s| > x\right) \le 2e^{-\frac{x^2}{8(b-a)}}$$

PROOF. Fix $n \ge 1$ and divide $[a,b]$ into n intervals of equal length. Let $X_i = W(a+k/n) - W(a+(k-1)/n)$ for $k = 1,2,\dots,n$. Then X_i are i.i.d $N(0,(b-a)/n)$. By the lemma, $\mathbf{P}(A_n) \le \exp\{-x^2/2(b-a)\}$ where $A_n := \{\max_{k\le n}|W(a+\frac{k}{n}) - W(a)| > x\}$. Observe that A_n are increasing and therefore $\mathbf{P}(\bigcup A_n) = \lim \mathbf{P}(A_n) \le \exp\{-\frac{x^2}{2(b-a)}\}$.

Evidently $\{\max_{t\in[a,b]} W_t - W_a > x\} \subseteq \bigcup_n A_n$ and hence $\mathbf{P}\left(\max_{t\in[a,b]} W_t - W_a > x\right) \le \exp\{-\frac{x^2}{2(b-a)}\}$. Putting absolute values on $W_t - W_s$ increases the probability by a factor of 2. Further, if $|W_t - W_s| > x$ for some s,t, then $|W_t - W_a| \ge \frac{x}{2}$ or $|W_s - W_a| \ge \frac{x}{2}$. Hence, we get the inequality in the statement of the corollary. ■

Theorem 3.10 (Paley, Wiener and Zygmund). *For any $\alpha < \frac{1}{2}$, almost surely, W is α-Hölder continuous on* $[0,1]$.

PROOF. Let $\Delta^* f(I) = \max\{|f(t) - f(s)| : t,s \in I\}$. By the corollary, $\mathbf{P}(\Delta^* W(I) > x) \le 2\exp\{-x^2|I|^{-1}/8\}$. Fix $n \ge 1$ and apply this to $I_{n,j}$, $j \le 2^n - 1$ to get

$$\mathbf{P}\left(\max_{j\le 2^n-1}\Delta^*(I_{n,j}) \ge 2^{-n\alpha}\right) \le 2^n \exp\left\{-c2^{n(1-2\alpha)}\right\}$$

which is summable. By Borel-Cantelli lemma, we conclude that there is some (almost surely finite) random constant A such that $\Delta^*(I_{n,j}) \le A2^{-n\alpha}$ for all dyadic intervals $I_{n,j}$.

Now consider any $s < t$. Pick the unique n such that $2^{-n-2} < t - s \le 2^{-n-1}$. Then, for some j, the dyadic interval $I_{n,j}$ contains both s and t. Hence $|W_t - W_s| \le \Delta^*(I_{n-1,j}) \le A2^{-n\alpha} \le A'|s-t|^\alpha$ whre $A' := A2^{2\alpha}$. Thus, W is a.s α-Hölder continuous. ■

3.4. Lévy's construction of Brownian motion

Let $(\Omega,\mathscr{F},\mathbf{P})$ be any probability space with i.i.d $N(0,1)$ random variables $\chi_1,\chi_2,\dots$ defined on it. The idea is to construct a sequence of random functions on $[0,1]$ whose finite dimensional distributions agree with that of Brownian motion at more and more points.

Step 1: A sequence of piecewise linear random functions: Let $W_1(t) := t\chi_1$ for $t \in [0,1]$. Clearly $W_0(1) - W_0(0) \sim N(0,1)$ as required for BM, but for any other t, $W_0(t) \sim N(0,t^2)$ whereas we want it to be $N(0,t)$ distribution..

Next, define

$$F_0(t) := \begin{cases} \frac{1}{2}\chi_2 & \text{if } t = \frac{1}{2}. \\ 0 & \text{if } t = 0 \text{ or } 1. \\ \text{linear in between.} \end{cases} \qquad \text{and} \qquad W_1 := W_0 + F_0 = \begin{cases} 0 & \text{if } t = 0. \\ \frac{1}{2}\chi_1 + \frac{1}{2}\chi_2 & \text{if } t = \frac{1}{2}. \\ \chi_1 & \text{if } t = 1. \\ \text{linear in between.} \end{cases}$$

$W_1(1) - W_1(1/2) = \frac{1}{2}\chi_1 - \frac{1}{2}\chi_2$ and $W_1(1/2) - W_1(0) = \frac{1}{2}\chi_1 + \frac{1}{2}\chi_2$ are clearly i.i.d $N(0,\frac{1}{2})$.

Proceeding inductively, suppose after n steps we have defined functions $W_0, W_1, \ldots, W_n$ such that for any $k \le n$,

(1) W_k is linear on each dyadic interval $[j2^{-k},(j+1)2^{-k}]$ for $j = 0,1,\ldots,2^k - 1$.
(2) The 2^k r.v.s $W_k((j+1)2^{-k}) - W_k(j2^{-k})$ for $j = 0,1,\ldots,2^k - 1$ are i.i.d $N(0,2^{-k})$.
(3) If $t = j2^{-k}$, then $W_\ell(t) = W_k(t)$ for any $\ell > k$ (and $\ell \le n$).
(4) W_k is defined using only χ_j, $j \le 2^k - 1$.

Then define (for some c_n to be chosen shortly)

$$F_n(t) = \begin{cases} c_n \chi_{j+2^n} & \text{if } t = \frac{2j+1}{2^{n+1}},\ 0 \le j \le 2^n - 1. \\ 0 & \text{if } t = \frac{2j}{2^{n+1}},\ 0 \le j \le 2^n. \\ \text{linear in between.} \end{cases} \qquad \text{and} \qquad W_{n+1} := W_n + F_n.$$

Does W_{n+1} satisfy the four properties listed above? The property (1) is evident by definition. To see (3), since F_n vanishes on dyadics of the form $j2^{-n}$, it is clear that $W_{n+1} = W_n$ on these points. Equally easy is (4), since we use 2^n fresh normal variables (and we had used $2^n - 1$ previously).

This leaves us to check (2). For ease of notation, for a function f and an interval $I = [a,b]$ denote the increment of f over I by let $\Delta f(I) := f(b) - f(a)$.

Let $0 \le j \le 2^n - 1$ and consider the dyadic interval $I_j = [j2^{-n},(j+1)2^{-n}]$ which gets broken into two intervals, $L_j = [(2j)2^{-n-1},(2j+1)2^{-n-1}]$ and $R_j = [(2j+1)2^{-n-1},(2j+2)2^{-n-1}]$ at level $n+1$. W_n is linear on I_j and hence, $\Delta W_n(L_j) = \Delta W_n(R_j) = \frac{1}{2}\Delta W_n(I_j)$. On the other hand, $\Delta F_n(L_j) = c_n \chi_{j+2^n} = -\Delta F_n(R_j)$. Thus, the increments of W_{n+1} on L_j and R_j are given by

$$\Delta W_{n+1}(L_j) = \frac{1}{2}\Delta W_n(I_j) + c_n\chi_{j+2^n} \qquad \text{and} \qquad \Delta W_{n+1}(R_j) = \frac{1}{2}\Delta W_n(I_j) - c_n\chi_{j+2^n}.$$

Inductively, we know that $\Delta W_n(I_j)$, $j \le 2^n - 1$ are i.i.d $N(0,2^{-n})$ and independent of χ_{j+2^n}, $j \ge 0$. Therefore, $\Delta W_{n+1}(L_j)$, $\Delta W_n(R_j)$, $j \le 2^n - 1$ are jointly normal (being linear combinations of independent normals). The means are clearly zero. To find the covariance, observe that for distinct j, these random variables are independent. Further,

$$\text{Cov}(\Delta W_{n+1}(L_j), \Delta W_{n+1}(R_j)) = \frac{1}{4}\text{Var}(\Delta W_n(I_j)) - c_n^2 = 2^{-n-2} - c_n^2.$$

Thus, if we choose $c_n := 2^{-(n+2)/2}$, then the covariance is zero, and hence $\Delta W_{n+1}(L_j)$, $\Delta W_{n+1}(R_j)$, $j \le 2^n - 1$ are independent. Also,

$$\text{Var}(\Delta W_{n+1}(L_j)) = \frac{1}{4}\text{Var}(\Delta W_n(I_j)) + c_n^2 = \frac{1}{4}2^{-n} + 2^{-n-2} = 2^{-n-1}.$$

Thus, W_{n+1} satisfies (2).

Step 2: The limiting random function: We found an infinite sequence of functions $W_1, W_2, \ldots$ satisfying (1)-(4) above. We want to show that almost surely, the sequence of functions W_n is uniformly convergent. For this, we need the fact that $\mathbf{P}(|\chi_1| > t) \le e^{-t^2/2}$ for any $t > 0$.

Let $\|\cdot\|$ denote the sup-norm on $[0,1]$. Clearly $\|F_m\| = c_m \sup\{|\chi_{j+2^m}| \, : \, 0 \le j \le 2^m - 1\}$. Hence

$$\mathbf{P}\left(\|F_m\| > \sqrt{c_m}\right) \le 2^m \mathbf{P}\left(|\chi| > \frac{1}{\sqrt{c_m}}\right) \le 2^m e^{-c_m^2/2} = 2^m \exp\left\{-2^{\frac{m-1}{2}}\right\}$$

which is clearly summable in m. Thus, by the Borel-Cantelli lemma, we deduce that almost surely, $\|F_m\| \le \sqrt{c_m}$ for all but finitely many m. Then, we can write $\|F_m\| \le A\sqrt{c_m}$ for all m, where A is a random constant (finite a.s.).

Thus, $\sum_m \|F_m\| < \infty$ a.s. Since $\|W_n - W_m\| \le \sum\limits_{k=n}^{m-1} \|F_k\|$, it follows that W_n is almost surely a Cauchy sequence in $C[0,1]$, and hence converges uniformly to a (random) continuous function W.

Step 3: Properties of W: We claim that W has all the properties required of Brownian motion. Since $W_n(0) = 0$ for all n, we also have $W(0) = 0$. We have already shown that $t \to W(t)$ is a continuous function a.s. (since W is the uniform limit of continuous functions). It remains to check the finite dimensional dstributions. If $s < t$ are dyadic rationals, say, $s = k2^{-n}$ and $t = \ell 2^{-n}$, then denoting $I_{n,j} = [j2^{-n}, (j+1)2^{-n}]$, we get that $W(s) = W_n(s) = \sum_{j=0}^{k-1} \Delta W_n(I_{n,j})$ and $W(t) - W(s) = \sum_{j=k}^{\ell-1} \Delta W_n(I_{n,j})$. As $\Delta W_n(I_{n,j})$ are i.i.d $N(0, 2^{-n})$ we get that $W(s)$ and $W(t) - W(s)$ are independent $N(0,s)$ and $N(0, t-s)$ respectively.

Now, suppose $s < t$ are not necessarily dyadics. We can pick s_n, t_n that are dyadics and converge to s, t respectively. Since $W(s_n) \sim N(0, s_n)$ and $W(t) - W(s) \sim N(0, t_n - s_n)$ are independent, by general facts about a.s convergence and weak convergence, we see that $W(s) \sim N(0,s)$ and $W(t) - W(s) \sim N(0, t-s)$ and the two are independent. The case of more than two intervals is dealt similarly. Thus we have proved the existence of Brownian motion on the time interval $[0,1]$.

Step 4: Extending to $[0,\infty)$: From a countable collection of $N(0,1)$ variables, we were able to construct a BM W on $[0,1]$. By subdividing the collection of Gaussians into a disjoint countable collection of countable subsets, we can construct i.i.d Brownian motions $B_1, B_2, \ldots$. Then, define for any $t \ge 0$,

$$B(t) = \sum_{j=1}^{\lfloor t \rfloor - 1} B_j(1) + B_{\lfloor t \rfloor}(t - \lfloor t \rfloor).$$

B is just a concatenation of $B_1, B_2, \ldots$. It is not difficult to check that B is a Brownian motion on $[0,\infty)$.

Theorem 3.11 (Wiener). *Brownian motion exists. Equivalenty Wiener measure exists.*

APPENDIX A

Characteristic functions as tool for studying weak convergence

Defintions and basic properties

Definition A.1. Let μ be a probability measure on $\mathbb{R}$. The function $\psi_\mu : \mathbb{R}^d \to \mathbb{R}$ define by $\psi_\mu(t) := \int_{\mathbb{R}} e^{itx} d\mu(x)$ is called the *characteristic function* or the *Fourier transform* of μ. If X is a random variable on a probability space, we sometimes say "characteristic function of X" to mean the c.f of its distribution. We also write $\hat{\mu}$ instead of ψ_μ.

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

The following lemma gives some basic properties of a c.f.

Lemma A.2. *Let $\mu \in \mathscr{P}(\mathbb{R})$. Then, $\hat{\mu}$ is a uniformly continuous function on $\mathbb{R}$ with $|\hat{\mu}(t)| \leq 1$ for all t with $\hat{\mu}(0) = 1$. (equality may be attained elsewhere too).*

PROOF. Clearly $\hat{\mu}(0) = 1$ and $|\hat{\mu}(t)| \leq 1$. U ■

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

(A) It transforms well under certain operations of measures, such as shifting a scaling and under convolutions.
(B) The c.f. determines the measure.
(C) $\hat{\mu}_n(t) \to \hat{\mu}(t)$ pointwise, if and only if $\mu_n \xrightarrow{d} \mu$.
(D) There exist necessary and sufficient conditions for a function $\psi : \mathbb{R} \to \mathbb{C}$ to be the c.f o f a measure. Because of this and part (B), sometimes one defines a measure by its characteristic function.

(A) Transformation rules

Theorem A.3. *Let X, Y be random variables.*

(1) *For any $a, b \in \mathbb{R}$, we have $\psi_{aX+b}(t) = e^{ibt}\psi_X(at)$.*
(2) *If X, Y are independent, then $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$.*

PROOF. (1) $\psi_{aX+b}(t) = \mathbf{E}[e^{it(aX+b)}] = \mathbf{E}[e^{itaX}]e^{ibt} = e^{ibt}\psi_X(at)$.
(2) $\psi_{X+Y}(t) = \mathbf{E}[e^{it(X+Y)}] = \mathbf{E}[e^{itX}e^{itY}] = \mathbf{E}[e^{itX}]\mathbf{E}[e^{itY}] = \psi_X(t)\psi_Y(t)$.

■

Examples.

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

(2) If $X \sim \text{Exp}(\lambda)$, then $\psi_X(t) = \int_0^\infty \lambda e^{-\lambda x} e^{itx} dx = \frac{1}{\lambda - it}$. If $Y \sim \text{Gamma}(\nu,\lambda)$, then if ν is an integer, then $Y \overset{d}{=} X_1 + \ldots + X_n$ where X_k are i.i.d Exp(λ). Therefore, $\psi_Y(t) = \frac{1}{(\lambda - it)^\nu}$.

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

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

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?! I gave a rigorous proof that the stuff inside brackets is indeed equal to 1, in class using contour integration, which will not be repeated here. The final concusion is that $N(\mu,\sigma^2)$ has c.f $e^{it\mu - \frac{\sigma^2 t^2}{2}}$.

(B) Inversion formulas

Theorem A.4. *If $\hat{\mu} = \hat{\nu}$, then $\mu = \nu$.*

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

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

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 $a < b$,

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

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

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

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

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

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

Corollary A.5 (Fourier inversion formula). *Let $\mu \in \mathscr{P}(\mathbb{R})$.*

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

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

(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 (A.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

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

Thus, we get the first statement of the corollary.

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

Next, note that for any $\sigma>0$, we have $|f_\sigma(\alpha)|\le C$ for all α where $C=\int|\hat{\mu}$. 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 A.4 tells us that

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

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

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

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

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

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

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

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

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

(C) Continuity theorem

Theorem A.6. *Let $\mu_n, \mu \in \mathscr{P}(\mathbb{R})$.*

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

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

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

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

Let μ_{n_k} be any subsequence that converges in distribution, say to ν. By tightness, $nu \in \mathscr{P}(\mathbb{R})$. Therefore, by part (a), $\hat{\mu}_{n_k} \to \hat{\nu}$ 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 $\mu_n \xrightarrow{d} \mu$ (because, if not, then there is some subsequence $\{n_k\}$ and some $\epsilon > 0$ such that the Lévy distance between μ_{n_k} and μ is at least ϵ. By tightness, $\mu_{\mathbf{n}_k}$ must have a subsequence that converges to some p.m ν which cannot be equal to μ contradicting what we have shown!).

It remains to show tightness. From Lemma A.7 below, as $n \to \infty$,

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

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

Thus, given $\epsilon > 0$, we can find $\delta > 0$ such that $\limsup_{n\to\infty}\mu_n\left([-2/\delta, 2/\delta]^c\right) < \epsilon$. This means that for some finite N, we have $\mu_n\left([-2/\delta, 2/\delta]^c\right) < \epsilon$ for all $n \ge N$. Now, find $A > 2/\delta$ such that for any $n \le N$, we get $\mu_n\left([-2/\delta, 2/\delta]^c\right) < \epsilon$. Thus, for any $\epsilon > 0$, we have produced an $A > 0$ so that $\mu_n\left([-A, A]^c\right) < \epsilon$ for all n. This is the definition of tightness. ■

Lemma A.7. *Let $\mu \in \mathscr{P}(\mathbb{R})$. Then, for any $\delta > 0$, we have*

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

PROOF. We write

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

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

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