18.675: Theory of Probability

Lecturer: Professor Nike Sun

Notes by: Andrew Lin

Fall 2019

Introduction

The course website for this class can be found at the URL in [10]. That page notably contains a summary of topics covered, as well as some additional references and brief lecture notes from other similar classes.

18.675 is the new numbering for 18.175; the material covered will be similar to previous years. The primary goal of the class is to cover **basic objects and laws in probability** in a formal mathematical framework, so some background in analysis and probability is strongly recommended. (In particular, if we are missing 18.100, we should talk to Professor Sun after class.) The main readings for this class will come from chapters 1–4 of our textbook [4], and readings for each chapter are included in the above summary of topics. There is some overlap with 18.125 (measure theory), particularly near the beginning, but most of the content will be significantly different.

Grades are calculated mostly from homework (25%), exam 1 (35%), and exam 2 (35%). Class participation is the remaining 5% of the grade; while this is a graduate class, we are expected to attend. This is a large class that will probably get smaller over time, so attendance may be just checked manually. Alternatively, a short quiz may be given occasionally in class which won't be graded and is mostly to check whether we're all following the material (and also to have a formal record of who showed up). Realistically, attending class is highly correlated with doing well, so we should definitely do so!

Remark 1. *Notes for this class have been edited to include some details not covered in class.*

1 September 4, 2019

We'll begin the class with some **measure theory**, which gives us a formal language for probability. On its own, measure theory is a bit dry, so we'll try to motivate why it's an important thing to study, starting with some elementary examples of probability and random variables. We should be familiar with these concepts from 18.600 or some other similar course:

- Consider a p-biased coin, where each flip comes up heads with probability p and tails with probability $(1-p)$. Then the probability that six independent p-biased coin flips come up (H, H, T, H, T, H) in that order is just the product of the probabilities, which is $p^4(1-p)^2$. And since there are $\binom{6}{4}$ ways to pick 4 heads among 6 coin tosses, the probability of having four heads when we flip six p-coins is $\binom{6}{4}p^4(1-p)^2$.
- Suppose U_n is a random variable distributed **uniformly** over the finite set $\{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}, 1\}$,and V_n is an **inde-**

pendent copy of U_n. Then for any $1 \le i, j \le n$, we have statements of independence like

$$\mathbb{P}\left(U_n = \frac{j}{n}, V_n = \frac{k}{n}\right) = \mathbb{P}\left(U_n = \frac{j}{n}\right)\mathbb{P}\left(V_n = \frac{k}{n}\right) = \frac{1}{n^2}.$$

But trying to compute probabilities like this becomes more complicated if the set of allowed states is infinite. For example, we might want to write down a formal definition for a random variable U which is uniform on $[0, 1]$, but we can't have a positive probability that U is equal to any particular value. And we might also want to ask if U_n converges to U as $n \to \infty$ in some sense (since we seem to be populating $[0, 1]$ uniformly).

We'll now see an example that illustrates that this kind of problem is trickier than it might initially seem:

Problem 2

Alice and Bob play a game with an infinite sequence of boxes. Alice puts a real number in each box, and Bob can reveal the number she put in every box except for one (which he can choose). "Show" that Bob can then guess the unseen number with 90 percent probability.

"Solution". Let S be the set of infinite sequences of real numbers, which can also be denoted $\mathbb{R}^{\mathbb{N}}$. For two sequences $x, y \in S$, say that $x \sim y$ (in other words, x and y are **equivalent**) if they eventually agree, so $x_n = y_n$ for all $n > N$ for some finite index N. This is an equivalence relation on S, so there is a quotient space $S/\sim$ of equivalence classes. Pick one representative z from each equivalence class, so that for any sequence $x \in S$, there is a unique representative sequence $z = z(x)$ such that z is equivalent to x. And since z and x are equivalent by definition, we can let $n(x)$ be the first index for which x and z agree for all indices $n(x)$ and onward.

We can now describe Bob's strategy for guessing the unseen number. He splits the boxes into ten rows, where row k contains the boxes originally in spots congruent to k mod 10. We now have ten real-valued sequences $x^{(1)}, \cdots, x^{(10)}$ corresponding to our ten rows – Bob picks one of these rows, uniformly at random (say row 10 without loss of generality). He then reveals the numbers in all nine other rows and computes

$$N = \max\left\{n(x^{(1)}), n(x^{(2)}), \cdots, n(x^{(9)})\right\} + 1.$$

(In other words, he looks at the first nine rows, and he finds the first index where each row agrees with its representative sequence. Then he finds the place N after which all rows agree.) If Bob then reveals every box in row 10 except for box N, he has enough information to deduce $z^{(10)}$, the representative sequence for $x^{(10)}$. Bob can then guess that the last box contains the representative sequence's Nth element. $z_N^{(10)}$.

But $x_N^{(10)}$ only disagrees with $z_N^{(10)}$ if the value of $n(x^{(10)})$ was the largest out of all of the $n(x^{(i)})$s. Because the row Bob chose was selected uniformly at random, and there's at most a 10 percent chance that he got the largest of the $n(x^{(i)})$s, his chance of guessing the correct number is at least 90 percent. □

This is a strange example – we can clearly get the probability to be arbitrarily close to 1 by replacing 10 with a larger number – so we've violated some laws of probability to get to this point. But we've constructed the problem in a way such that we don't really know where we've messed up! That's why we'll start with the axioms and use measure theory to help us understand more clearly where we stand (and the issue will end up having to do with **measurability**).

Example 3

We'll start with a seemingly basic problem, defining the uniform random variable $U \sim \text{Unif}[0, 1]$ from above.

One way to formalize this random variable is to construct a function (which we call a **measure**) μ on subsets of $[0, 1]$, such that for any subset A, $\mu(A)$ is the probability that $U \in A$. For example, for any $0 \le a \le b \le 1$, it's natural

to want

$$\mu([a,b]) = b - a$$

if we require U to be uniform. We also want to have some form of **additivity**, meaning that if $A, B \subseteq [0,1]$ are disjoint, then their disjoint union $A \sqcup B$ should have measure

$$\mu(A \sqcup B) = \mu(A) + \mu(B).$$

By induction, this means that for any positive integer n,

$$\mu\left(\bigsqcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i).$$

Generalizing this further, if we take the limiting disjoint union $\bigsqcup_{i=1}^{n} A_i \to \bigsqcup_{i=1}^{\infty} A_i$, it's also natural that we may want **countable additivity**

$$\mu\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Remark 4. *On the other hand, note that $\mu([0,1]) = 1$, but $[0,1]$ is an uncountable union of points, and each point has measure $\mu(\{p\}) = \mu([p,p]) = 0$. So we can't ask our measure to have* **uncountable addivity***, because 1 is not the sum of a bunch of zeros.*

So to summarize, our measure $\mu(A) = \mathbb{P}(U \in A)$ (which is meant to define the uniform random variable U) should have the following properties:

- It should respect lengths of intervals, meaning that$\mu([a,b]) = b - a$ for all $0 \le a \le b \le 1$.
- It should satisfy countable additivity (as defined above).
- It should also be **translation-invariant**, meaning that if $A \subseteq [0,1]$, and we define $A_x = A + x$ to be A shifted rightward by x (mod 1), then $\mu(A) = \mu(A_x)$.

These properties may not seem like much, so we might ask if this is enough to characterize U – in other words, if we're given any subset A, we want to use these three equations to determine $\mu(A)$. Let's try a slightly complicated example to see this in action:

Example 5

Consider the Cantor set $C = \bigcap_{i=1}^{\infty} C_n$, where the C_n are iteratively defined as

$$C_0 = [0,1], \quad C_{n+1} = C_n \setminus (\text{"middle third" of each interval of } C_n).$$

(Because the C_is are nested, their intersection is indeed well-defined.) To find the measure of C, we know that $\mu(C_0) = 1$, and we also know that C_n is the disjoint union of C_{n+1} and the middle third of C_n, so

$$\mu(C_{n+1}) = \mu(C_n) - \mu(\text{middle third of } C_n) = \frac{2}{3}\mu(C_n).$$

This means that $\mu(C_n) = \left(\frac{2}{3}\right)^n$ for all n, and taking the limit, the measure of the Cantor set is $\mu(C) = 0$. So this seems like encouraging evidence towards an answer of "yes, we can determine $\mu(A)$ in general," but it turns out the Cantor set is actually pretty well-behaved compared to some other subsets of $[0,1]$. In fact, we've actually been asking the wrong question – this function μ isn't even well-defined for all subsets:

Proposition 6

There is no function $\mu : \mathcal{P}([0,1]) \to [0,1]$ that satisfies all three of the properties we want in a measure.

Proof (the Vitali construction). Define an equivalence relation on the real numbers by setting $x \sim y$ if $x - y \in \mathbb{Q}$. This partitions the real line into equivalence classes (cosets of the form $[x] = x + \mathbb{Q}$), where each one is a shift of $\mathbb{Q}$ by some real number. Similarly to Problem 2, we pick a representative z from each equivalence class, and for simplicity, we can pick all z to be in $[0,1]$ (since a number is always in the same class as its fractional part).

Let A be the set of all representatives, which is a subset of $[0,1]$ (this is sometimes called the **Vitali set**). Since A_q is A translated by q units, every real number in $[0,1]$ is in A_q for some rational number q. Now for any rational $q \in [0,1]$, we have $A_q \subset [0,2]$, so we can write

$$A_q = (A_q \cap [0,1]) \sqcup (A_q \cap [1,2]) \implies \boxed{\mu(A_q)} = \mu(A_q \cap [0,1]) + \mu(A_q \cap [1,2]) = \boxed{\mu(A_q \cap [0,1]) + \mu(A_{q-1} \cap [0,1])}.$$

(Basically, think of taking the part of A_q that lies in $[1,2]$ and translating it to the left by 1 unit.) Now for any rational number $q \in [0,1)$, A_{q+n} only intersects the interval $[0,1]$ if $n = 0$ or $n = -1$. Thus, we can split up the interval $[0,1]$ into contributions from different rational numbers q and use countable additivity:

$$\begin{aligned} 1 = \mu([0,1]) &= \mu\left(\bigcup_{q \in \mathbb{Q}, q \in [0,1)} (A_{q-1} \cap [0,1]) \cup (A_q \cap [0,1]) \right) \\ &= \sum_{q \in \mathbb{Q}, q \in [0,1)} \mu(A_{q-1} \cap [0,1]) + \mu(A_q \cap [0,1]) \\ &= \sum_{q \in \mathbb{Q}, q \in [0,1)} \mu(A_q). \end{aligned}$$

But now all A_qs have equal measure by translation invariance, so $1 = \sum_{q \in \mathbb{Q}, q \in [0,1)} \mu(A)$, which is impossible because 1 cannot be the sum of countably many equal numbers. □

The way measure theory deals with this problem is to **not require that μ be defined on all subsets of our space.** As a historical note, this was a pretty surprising idea when it was first proposed, but it's really the only thing we can do if we try to write formal definitions down. (We don't really want to relax the conditions on our measure, so it's better to just not define the measure on some pathological subsets.)

Definition 7

A **measure space** is a triple $(\Omega, \mathcal{F}, \mu)$ satisfying the following axioms:

- Ω, the **state space** or **outcome space**, is a nonempty set.
- $\mathcal{F}$, a set of **measurable subsets** or **events**, is a ***σ*-field** or ***σ*-algebra** over Ω (those terms will be used interchangeably). In other words, $\mathcal{F}$ is a nonempty collection of subsets of Ω which is closed under complementation and countable union, so if $A \in \mathcal{F}$, then $A^c = \Omega \setminus A$ is also in $\mathcal{F}$, and if $A_i \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i$ is also in $\mathcal{F}$.
- μ, the **measure**, is a map $\mathcal{F} \to [0, \infty]$ which is not infinite everywhere and is countably additive. In other words, if $A_i \in \mathcal{F}$ are disjoint, then $\mu\left(\bigsqcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$.

The goal of our next two lectures will basically be to construct a uniform measure on $[0,1]$: we'll define a measure space $([0,1], \mathcal{L}, \mu)$, where $\mathcal{L}$ will be defined later, and similarly this will also allow us to define the measure space

$(\mathbb{R}, \mathcal{L}_\mathbb{R}, \mu)$.

Definition 8

A measure μ is a **probability measure** (which we will denote $\mathbb{P}$) if $\mu(\Omega) = 1$.

Just working with the definitions, we can gather a few immediate consequences about measure spaces:

Proposition 9

Given a measure space $(\Omega, \mathcal{F}, \mu)$, the following must hold:

1. $\varnothing$ and Ω are in $\mathcal{F}$,
2. $\mu(\varnothing) = 0$,
3. (continuity from below) if $A_i \uparrow A$ – that is, $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{i=1}^{\infty} A_i$ – then $\mu(A_i) \uparrow \mu(A)$,
4. (continuity from above) if $A_i \downarrow A$ – that is, $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \bigcap_{i=1}^{\infty} A_i$ – and also $\mu(A_1) < \infty$, then $\mu(A_i) \downarrow \mu(A)$,
5. for any Ω, $\mathcal{F} = \{\varnothing, \Omega\}$ is a valid σ-field, and so is $\mathcal{P}(\Omega)$.

Proof. For (1), because $\mathcal{F}$ is nonempty, there is some event A in $\mathcal{F}$. Then $A^c \in \mathcal{F}$, so $A \cup A^c = \Omega$ must be in $\mathcal{F}$, and so must $\Omega^c = \varnothing$. For (2), there is some A such that $\mu(A) < \infty$, so by additivity $\mu(A \cup \varnothing) = \mu(A) = \mu(A) + \mu(\varnothing)$, meaning $\mu(\varnothing) = 0$.

For (3), we use countable additivity on $A = \bigsqcup_{i=1}^{\infty} B_i$, where $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus A_2$, and so on. Then

$$\mu(A) = \sum_{n=1}^{\infty} \mu(B_n),$$

and since $\mu(A_n) = \sum_{i=1}^{n} \mu(B_i)$ are the partial sums of our infinite sum, $\mu(A_i)$ indeed converges to $\mu(A)$. Similarly, for (4), define $B_1 = \varnothing$, $B_2 = A_1 \setminus A_2$, $B_3 = A_1 \setminus A_3$, and so on. All B_is have finite measure since $\mu(A_1)$ is finite, and $B_1 \subseteq B_2 \subseteq \cdots \subseteq A_1 \setminus A$, so we can apply (3) on the B_is to find that $\mu(A_1 \setminus A_i) \uparrow \mu(A_1 \setminus A)$, so $\mu(A_i) \downarrow \mu(A)$ by additivity. (The assumption that $\mu(A_1) < \infty$ is necessary – a counterexample with $\mu(A_1) = \infty$ is $A_i = [i, \infty)$.)

Finally, the two examples in (5) are both σ-fields because they satisfy both closure axioms (complementation and countable union). □

If Ω is a finite set, and we wanted to define a probability space in 18.600, we would define a probability mass function $\mathbb{P}(\omega) = a_\omega$ for every $\omega \in \Omega$, such that $\sum_{\omega \in \Omega} a_\omega = 1$. In our new notation, $\mathcal{F} = \mathcal{P}(\Omega)$ is the set of all possible events, and $\mathbb{P}(A) = \sum_{w \in A} a_\omega$ for A. (We can check that this is a valid probability space from the axioms.) But our discussion above shows that if we have a set like $\Omega = [0, 1]$ or $\mathbb{R}$, and we take $\mathcal{F} = \mathcal{P}(\Omega)$, then we do have a valid σ-field, but the Vitali set shows that we can't define a measure μ on it. So next week, we'll basically ask how to restrict our set $\mathcal{F}$ to get a useful measure.

Definition 10

The **Borel σ-field** $\mathcal{B}_\mathbb{R}$ is the smallest σ-field over $\mathbb{R}$ that contains $\mathcal{I}$, the set of all open intervals on $\mathbb{R}$.

Importantly, whenever we see the word "smallest" in a definition, we should ask whether we have a well-defined object (since there can be multiple "minimal" objects in general). But in this case, if $\{\mathcal{F}_\alpha : \alpha \in I\}$ is a collection of σ-fields over Ω, then the intersection of the $\mathcal{F}_\alpha$s is also a σ-field. (Indeed, if a set A is the intersection of some

σ-fields, then both A and A^c are in all of the σ-fields, so A^c is in the intersection. The same argument works for countable union.) So we can just let $\mathcal{B}_{\mathbb{R}}$ be the intersection of all σ-fields, and thus the Borel σ-field is well-defined.

2 September 9, 2019

Fact 11

The next two lectures in this class were given by Professor Subhabrata Sen.

Our central question today is how to define the uniform measure on $[0,1]$. Basically, we will define a function μ on (some) subsets of $[0,1]$ with the requirements that

- $\mu((a,b]) = b - a$ (choosing half-open intervals is just a matter of convention),
- given disjoint subsets $\{A_i : i \geq 1\}$, we have $\mu\left(\bigsqcup_i A_i\right) = \sum_i \mu(A_i)$, and
- if we define $A_x = (A + x \bmod 1)$, then $\mu(A_x) = \mu(A)$ (translation invariance).

Vitali's construction tells us that we can't do this with all subsets of $[0,1]$, or in other words that there is no function $\mu : 2^{[0,1]} \to [0,1]$ that satisfies all three conditions above. (Here 2^A is the powerset of A.) So we have to compromise and try to define μ on just a subcollection of the subsets of $[0,1]$. Last time, we defined a **measure space** $(\Omega, \mathcal{F}, \mu)$, consisting of a nonempty state space, a σ-field, and a measure. We also defined the **Borel σ-field** to be the smallest σ-field containing the open intervals (which are the sets where we really want the measure to be defined in a particular way). Notationally, this can be written as

$$\mathcal{B} = \sigma(\mathcal{I}), \quad \mathcal{I} = \{(a,b] \cap \mathbb{R} : a \leq b\}.$$

With this, we will try to construct the **Lebesgue measure** on $[0,1]$. Our hope is that because we already know how to assign a measure to intervals, we can keep assigning measures to subsets of $[0,1]$ by building them from intervals. Unfortunately, this is not rigorous, because there's no closed form for an arbitrary element of $\mathcal{B}_{\mathbb{R}}$, and in fact it's not true that every set in the Borel σ-algebra can be written as a countable union and/or intersection of open intervals!

Instead, we do define $\mu((a,b]) = b-a$, and we do define $\mu\left(\bigsqcup_i (a_i,b_i]\right) = \sum_i (b_i - a_i)$ for unions of disjoint intervals $(a_i, b_i]$, but we're not quite done. Instead, we then need to know whether there is a **consistent extension** of our function μ to all sets in $\mathcal{B}_{\mathbb{R}}$. Since this idea will come up again later in the course, we'll make a slight generalization, defining the function $F(x) = \mu((0,x])$. (So we're primarily working with the case $F(x) = x$, but this works for a general measure μ.)

Remark 12. *Remember that measures in general do not need to be translation-invariant (it's not part of Definition 7) – we just want translation invariance when we are trying to define a uniform measure. In particular, translation invariance will not hold if $F(x)$ is nonlinear.*

From the properties of measures we've been studying, we can notice the following properties of F:

- For any $x_1 > x_2$, we have $\mu((0,x_1]) \geq \mu((0,x_2]) \implies F(x_1) \geq F(x_2)$, so F must be **monotone**.
- If $x_n \downarrow x$, then the intervals $(0,x_n] \downarrow (0,x]$, meaning $\mu((0,x_n]) \downarrow \mu((0,x])$ by continuity from above. Thus, we also have $F(x_n) \downarrow F(x)$ – in other words, F is **right-continuous**.

We'll see soon why the function F is not required to be left-continuous, and we'll give functions $F(x)$ of this form a name:

Definition 13

A **Stieltjes measure function** on $\mathbb{R}$ is a function $F : \mathbb{R} \to \mathbb{R}$ which is non-decreasing and right-continuous.

Theorem 14 (Lebesgue-Stieltjes)

For any Stieltjes measure function F, there is a unique measure $\mu = \mu_F$ on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ satisfying $\mu((a, b]) = F(b) - F(a)$ for all $a, b \in \mathbb{R}$.

We'll do the proof in two steps (noting that we start off with the definition of μ on the set of open intervals $(a, b]$:

1. Show that there exists a unique extension from $\mu : \mathcal{I} \to [0, \infty]$ to $\mu : \mathcal{A} \to [0, \infty]$, where

$$\mathcal{A} = \{A \subseteq \mathbb{R} : A \text{ is a finite disjoint union of sets in } \mathcal{I}.\}$$

   (In other words, we define μ for any finite disjoint union of intervals.)

2. Show that we can extend our measure from being defined on $\mathcal{A}$ to being defined on $\mathcal{B}_\mathbb{R}$.

Remark 15. *Notice that $\mathcal{A}$ is closed under complements and* **finite** *unions (exercise), while the Borel σ-algebra is closed under complements and* **countable** *unions. So by using $\mathcal{A}$ as an intermediate step, we're not quite at the σ-algebra yet, but we're kind of close.*

Proof of step 1. The construction itself is pretty intuitive: given any subset $A \in \mathcal{A}$, we can write it as $A = \bigsqcup_{i=1}^n E_j$ for some intervals E_js in $\mathcal{A}$, so we define

$$\mu(A) = \sum_{j=1}^n \mu(E_j)$$

by additivity. But there are many different ways to write an element of $\mathcal{A}$ as a finite disjoint union of intervals, so we need to check that

$$A = \bigsqcup_{j=1}^n E_j = \bigsqcup_{k=1}^m F_k \implies \sum_{j=1}^n \mu(E_j) \stackrel{?}{=} \sum_{k=1}^m \mu(F_k).$$

Indeed, because all E_js and F_ks are half-open intervals, we can take the common refinement of the partitions and notice that

$$\sum_{j=1}^n \mu(E_j) = \sum_{j=1}^n \mu(E_j \cap A) = \sum_{j=1}^n \mu\left(E_j \cap \left(\bigsqcup_{k=1}^m F_k\right)\right) = \sum_{j=1}^n \sum_{k=1}^m \mu(E_j \cap F_k).$$

Importantly, the last equality $\mu(E_j \cap \bigsqcup_{k=1}^m F_k) = \sum_{k=1}^m \mu(E_j \cap F_k)$ comes not from additivity (because we don't know that yet) but from the fact that the F_ks form a disjoint union of A and thus must cover E_j. This can only be done with a finite set of half-open intervals if they line up back-to-back, and then the lengths will indeed add up to the total length of E_j. But then starting this calculation from $\sum_k \mu(F_k)$ also yields this result (after swapping the order of summation, which is okay because we have finite sums), so the two ways of computing the measure are the same, and $\mu(A)$ is consistently defined. Thus we have indeed extended μ from $\mathcal{I}$ to $\mathcal{A}$. □

We will now record a few properties of the measure which will be useful for next lecture:

Proposition 16

Our measure $\mu = \mu_F$ has the following properties on $\mathcal{A}$:

1. μ is monotone and finitely additive over $\mathcal{A}$, meaning that for any **disjoint** $A_1, \cdots, A_n \in \mathcal{A}$, $\mu\left(\bigsqcup_i A_i\right) = \sum_{i=1}^n \mu(A_i)$.
2. μ is finitely subadditive over $\mathcal{A}$, meaning that for **any** $A_1, \cdots, A_n \in \mathcal{A}$, $\mu\left(\bigcup_i A_i\right) \leq \sum_{i=1}^n \mu(A_i)$.
3. μ is countably subadditive over $\mathcal{I}$: if $A, A_i \in \mathcal{I}$ (where i ranges over the positive integers) and $A \subseteq \bigcup_i A_i$, then $\mu(A) \leq \sum_i \mu(A_i)$.
4. μ is countably additive on $\mathcal{A}$.

Property (4) is really what we care about most, because we want countable additivity on the final space $\mathcal{B}_{\mathbb{R}}$ that we're defining μ_F on.

Proof. It suffices to show (1) and (2) for $n = 2$ by induction. For (1), write $A_1 = E_1 \sqcup \cdots \sqcup E_x$ and $A_2 = F_1 \sqcup \cdots \sqcup F_y$, where all E_is and F_is are half-open intervals. By definition of μ on $\mathcal{A}$, we then have

$$\mu(A_1 \sqcup A_2) = \mu(E_1 \sqcup \cdots \sqcup E_x \sqcup F_1 \sqcup \cdots \sqcup F_y) = \sum_{i=1}^{x} \mu(E_i) + \sum_{i=1}^{y} \mu(F_i)$$

because the E_is and F_is are all disjoint intervals, and then the two sums on the right-hand side are $\mu(A_1)$ and $\mu(A_2)$, proving additivity. Monotonicity then also follows because all terms here are nonnegative (lengths of intervals), so if $A \subset B$, then $\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A)$. For (2), write $A_1 \cup A_2$ as the disjoint union of A_1 and $A_2 \setminus A_1$, so then by additivity and monotonicity from (1) and using the fact that $A_2 \setminus A_1 \subseteq A_2$,

$$\mu(A_1 \cup A_2) = \mu(A_1 \cup (A_2 \setminus A_1)) = \mu(A_1) + \mu(A_2 \setminus A_1) \leq \mu(A_1) + \mu(A_2).$$

For (3), we're working over $\mathcal{I}$, so write $A = (a, b]$ and $A_i = (a_i, b_i]$, and we're given that $(a, b] \subseteq \bigcup_{i=1}^{\infty}(a_i, b_i]$. (By definition, we have $\mu(A) = F(b) - F(a)$ and $\mu(A_i) = F(b_i) - F(a_i)$.) Since F is right-continuous, there exist real numbers $x > a, c_i > b_i$ such that $F(x) - F(a) \leq \varepsilon$ and $F(c_i) - F(b_i) \leq \frac{\varepsilon}{2^i}$. (Intuitively, this now lets us work with A as a closed interval and the A_is as open intervals.) We still have a covering

$$[x, b] \subseteq \bigcup_{i=1}^{\infty}(a_i, c_i)$$

because we've only made A smaller and the A_is bigger. Since $[x, b]$ is compact, and the right hand side is an open covering, there exists a finite subcover by the Heine-Borel theorem, which we will denote (with some abuse of notation)

$$[x, b] \subseteq \bigcup_{i=1}^{n}(a_i, c_i)$$

Now by (1) and (2), because μ is finitely subadditive and monotone,

$$\begin{aligned}\mu((a,b]) &\le \mu((a,x]) + \sum_{i=1}^{n} \mu((a_i,c_i]) \\ &= (F(x)-F(a)) + \sum_{i=1}^{n}(F(c_i)-F(a_i)) \\ &= (F(x)-F(a)) + \sum_{i=1}^{n}(F(c_i)-F(b_i)) + \sum_{i=1}^{n}(F(b_i)-F(a_i)) \qquad \le 2\varepsilon + \sum_{i=1}^{\infty}(F(b_i)-F(a_i))\end{aligned}$$

(one ε factor comes from $F(x)-F(a)$, and the other comes from the sum of the geometric series $\sum \frac{\varepsilon}{2^i}$). But now taking $\varepsilon \to 0$, we indeed get countable subadditivity

$$\mu((a,b]) \le \sum_{i=1}^{\infty} \mu((a_i,b_i])$$

as desired. Finally, for (4), we are trying to prove that given $A, A_i \in \mathcal{A}$, where $A = \bigsqcup_{i=1}^{\infty} A_i$, we have $\mu(A) = \sum_{i=1}^{\infty} \mu(A_i)$. Write

$$A = \bigsqcup_{i=1}^{n} A_i \sqcup \left(\bigsqcup_{i=n+1}^{\infty} A_i \right) = \bigsqcup_{i=1}^{n} A_i \sqcup C_{n+1},$$

where $C_{n+1} = \bigsqcup_{i=n+1}^{\infty} A_i$ is actually a finite disjoint union of intervals as well (because it is the intervals of A, with the intervals in $A_1, \cdots, A_n$ removed). Thus, we can apply (1) to say that

$$\mu(A) = \sum_{i=1}^{n} \mu(A_i) + \mu(C_{n+1}) \ge \sum_{i=1}^{n} \mu(A_i).$$

Taking n to infinity, we find that

$$\mu(A) \ge \sum_{i=1}^{\infty} \mu(A_i),$$

which gives one direction of the inequality. To show the upper bound, without loss of generality, we can assume that all $A_i \in \mathcal{I}$, because each A_i is originally a finite disjoint union of intervals (so the whole disjoint union is a countable disjoint union of intervals). Also, because A is an element of $\mathcal{A}$, we can separately write $A = \bigsqcup_{j=1}^{n} E_j$ with all $E_j \in \mathcal{I}$. So now because A is entirely contained in $\left(\bigcup_i A_i\right)$, we may write

$$\mu(A) = \sum_{j=1}^{n} \mu(E_j) = \sum_{j=1}^{n} \mu\left(E_j \cap \left(\bigcup_i A_i \right) \right) = \sum_{j=1}^{n} \mu\left(\bigcup_i (E_j \cap A_i) \right).$$

Since E_j and A_i are both intervals, so is $E_j \cap A_i$, and thus by (3) we have

$$\mu(A) \le \sum_{j=1}^{n} \sum_{i=1}^{\infty} \mu(E_j \cap A_i) = \sum_{i=1}^{\infty} \sum_{j=1}^{n} \mu(E_j \cap A_i)$$

by swapping the order of summation (okay because we have a nonnegative sum), and finally by (1) this yields

$$\mu(A) \le \sum_{i=1}^{\infty} \mu(A \cap A_i) = \sum_{i=1}^{\infty} \mu(A_i),$$

proving the other direction of the inequality and verifying (4). □

Our next step is to extend μ from $\mathcal{A}$ to the actual σ-field $\mathcal{B}_{\mathbb{R}}$. For now, let's assume the measure we want to define is a **finite** measure (meaning that it never takes on the value of ∞). We'll show the following result next time:

Theorem 17 (Carathéodory extension theorem)
Let $\mathcal{A}$ be an algebra over Ω (meaning that it is closed under complements and finite unions), and let $\mu : \mathcal{A} \to [0, \infty)$ be a countably additive finite measure. Then there exists a unique measure on the generated σ-algebra $\mu : \sigma(\mathcal{A}) \to [0, \infty)$ which extends μ.

3 September 11, 2019

Recall that the central object we're studying is the Borel σ-field $\mathcal{B}_{\mathbb{R}}$, which is the smallest σ-algebra generated by the intervals $\mathcal{I} = \{(a, b] \cap \mathbb{R} : a \leq b\}$. Last time, we started working towards a proof that given any Stieltjes measure function (monotone, right-continuous) F, there exists a unique μ_F on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ such that $\mu((a, b]) = F(b) - F(a)$. In particular, we extended the measure from $\mathcal{I}$ to the algebra $\mathcal{A}$ consisting of finite disjoint unions of intervals in $\mathcal{I}$, and today, we'll further extend this to $\mathcal{B}_{\mathbb{R}} = \sigma(\mathcal{A})$.

Remark 18. *A quick follow-up from last lecture – if we define $F(x) = \mu((0, x])$, then it is not true that F needs to be left-continuous. After all, when $x_n \uparrow x$, $(0, x_n] \uparrow (0, x)$, while $F(x)$ uses the interval $(0, x]$ instead:*

$$\mu((0, x_n]) \uparrow \mu((0, x)) \neq \mu((0, x]) = F(x).$$

The example to keep in mind is that if we have a step function $F(x)$ with a point mass, then $(0, x)$ and $(0, x]$ differ by the mass at point x, but this is a valid Stieltjes measure function.

The extension of the measure μ_F (and thus one way to construct a uniform measure) will be done by proving the Carathéodory extension theorem, Theorem 17, today. We'll go through the proof in a slightly nonlinear way; we want to prove (1) that an extension exists and (2) that it is unique, but we'll do the latter first by introducing the π-λ theorem.

Definition 19
A **π-system** $\mathcal{P}$ on a set Ω is a nonempty collection of subsets of Ω, such that for any $A, B \in \mathcal{P}$, $A \cap B \in \mathcal{P}$.

For example, the collection of half-open intervals $\mathcal{I}$ is a π-system.

Definition 20
A **λ-system** $\mathcal{L}$ on a set Ω is a collection of subsets of Ω satisfying the following conditions:

- $\Omega \in \mathcal{L}$.
- For any $A, B \in \mathcal{L}$ where $B \subseteq A$, $A \setminus B \in \mathcal{L}$.
- If we have a sequence $A_n \in \mathcal{L}$, and $A_n \uparrow A$, then $A \in \mathcal{L}$.

All σ-algebras are both π-systems (we have closure under intersections by definition) and λ-systems (in any σ-algebra, we have Ω by taking any $A \cup A^c$, we have $A \setminus B$ because it is also $A \cap B^c$, and we have $A = \bigcup_n A_n = A$ whenever $A_n \uparrow A$ by countable union). And the converse also holds:

Proposition 21
If a λ-system $\mathcal{L}$ is also a π-system, then $\mathcal{L}$ is a σ-algebra.

Proof. Because $\mathcal{L}$ contains the whole set Ω (by definition of a λ-system), it is nonempty. And because $\Omega \in \mathcal{L}$, $\Omega \setminus A = A^c \in \mathcal{L}$ for any $A \in \mathcal{L}$, so it is also closed under complementation. Finally, for countable union, if $\mathcal{L}$ contains A_1, A_2, then it also contains $A_1 \cup A_2$ (the complement of $A_1^c \cap A_2^c$), and more generally it contains $A_1 \cup \cdots \cup A_n$ for any n, so it must also contain the limit $\bigcup_n A_n$. □

In other words, the definition of a σ-algebra can be split into the verifying the intersections part (π-system) and the rest (λ-system), and this turns out to be very helpful:

Theorem 22 (Dynkin's π-λ theorem)
If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system, and $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.

This theorem will have applications in various different contexts, so we should try to understand how it works.

Proof. Without loss of generality, we can let $\mathcal{L}$ be the smallest λ-system containing $\mathcal{P}$ (this only makes $\sigma(\mathcal{P}) \subseteq \mathcal{L}$ harder to achieve). This is well-defined, because there is at least one λ-system – the powerset of Ω – and intersections $\mathcal{L}_1 \cap \mathcal{L}_2$ of λ-systems are λ-systems as well:

- Ω is in both $\mathcal{L}_1$ and $\mathcal{L}_2$, so it's also in $\mathcal{L}_1 \cap \mathcal{L}_2$.
- If A and B are both in $\mathcal{L}_1 \cap \mathcal{L}_2$, they're both in $\mathcal{L}_1$ and $\mathcal{L}_2$, so $A \setminus B$ is in both λ-systems, meaning it's also in their intersection.
- If all A_ns (with $A_n \uparrow A$) are in $\mathcal{L}_1 \cap \mathcal{L}_2$, then they're in both λ-systems, so A is in both λ-systems and therefore the intersection as well.

Thus, the smallest $\mathcal{L}$ is the intersection of all λ-systems containing P. It suffices to show that $\mathcal{L}$ is also a π-system, because that would mean $\mathcal{L}$ is a σ-field (by Proposition 21) containing $\mathcal{P}$, so it must contain $\sigma(P)$, the smallest σ-field containing $\mathcal{P}$. In other words, we need to show that given any $A, B \in \mathcal{L}$, $A \cap B \in \mathcal{L}$. Fix any $A \in \mathcal{P}$, and define

$$\mathcal{L}_A = \{B \in \mathcal{L} : A \cap B \in \mathcal{L}\} \subseteq \mathcal{L},$$

the subset of $\mathcal{L}$ whose intersections with A are also in $\mathcal{L}$. We can can check that $\mathcal{L}_A$ is a λ-system:

- $A \cap \Omega = A \in \mathcal{L}$, so $\Omega \in \mathcal{L}_A$.
- If $B_2 \subseteq B_1$ are in $\mathcal{L}_A$, meaning $A \cap B_1, A \cap B_2 \in \mathcal{L}$, then $(A \cap B_1) \setminus (A \cap B_2) = A \cap (B_1 \setminus B_2)$ is also in $\mathcal{L}$, so $B_1 \setminus B_2$ is also in $\mathcal{L}_A$.
- If we have an increasing sequence $B_n \uparrow B \in \mathcal{L}$, and all $B_n \in \mathcal{L}_A$ (that is, $A \cap B_n \in \mathcal{L}$ for all n), then $A \cap B \in \mathcal{L}$ because $(A \cap B_n) \uparrow A \cap B$, so B is also in $\mathcal{L}_A$.

In addition, because $A \in \mathcal{P}$ and $\mathcal{P}$ is a π-system, any element $B \in \mathcal{P}$ also has $A \cap B \in \mathcal{P} \subseteq \mathcal{L}$. Thus $\mathcal{L}_A$ contains $\mathcal{P}$. But $\mathcal{L}$ is the **smallest** λ-system containing $\mathcal{P}$, so $\mathcal{L} \subseteq \mathcal{L}_A \subseteq \mathcal{L}$ and the two sets are equal. Therefore, for all $A \in \mathcal{P}$ and $B \in \mathcal{L}$, $A \cap B \in \mathcal{L}$. Now fix any $B \in \mathcal{L}$, and define

$$\mathcal{L}_B = \{A \in \mathcal{L} : A \cap B \in \mathcal{L}\}.$$

By the argument we just made, $\mathcal{P} \subset \mathcal{L}_B$ (any intersection between an element of $\mathcal{P}$ and $\mathcal{L}$ is in $\mathcal{L}$). In addition, the exact same argument as for $\mathcal{L}_A$ shows that $\mathcal{L}_B$ must be a λ-system, so $\mathcal{L}_B = \mathcal{L}$. So the intersection of any two elements of $\mathcal{L}$ is in $\mathcal{L}$, meaning that $\mathcal{L}$ is a π-system, which is what sufficed to prove our result. □

The important takeaway from this proof is that measure-theoretic proofs are difficult because we don't have a closed form for all measurable subsets. Instead, the right way is to first verify that a certain subset $\mathcal{P}$ has the property we want, and then this theorem is powerful because it can bootstrap results from $\mathcal{P}$ to its σ-algebra. In particular, we can now prove the uniqueness part of the Carathéodory extension theorem:

Proof of uniqueness for Theorem 17. Suppose that there were two distinct extensions μ_1 and μ_2 of our measure μ from $\mathcal{A}$ to $\sigma(\mathcal{A})$. Then $\mu(A) = \mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{A}$ by definition, and we can define

$$\mathcal{L} = \{B \in \sigma(A) : \mu_1(B) = \mu_2(B)\}$$

to be the subsets in $\sigma(A)$ where the extensions agree. (We want to show that $\mathcal{L}$ is the σ-algebra $\sigma(A)$ itself.) We've established that $\mathcal{A} \subseteq \mathcal{L}$, and we know that $\mathcal{A}$ is a π-system (intersections of finite collections of intervals are a finite collection of intervals). Now $\mathcal{L}$ is a λ-system, as we can verify directly:

- The whole space Ω is in $\mathcal{L}$, because an algebra always contains $A \cup A^c = \Omega$.

  Remark 23. *There are actually some small technicalities here depending on the space Ω we choose – for example, $[0, 1]$ is not a finite union of half-open intervals because of the closed lower bound, so it's easiest to define a uniform measure on the real line first and restrict to $[0, 1]$ afterward. But we're also assuming that our measure only takes on finite values, so the way we actually need to set up our construction is to make this argument on (for example) $(-n, n]$ for each positive integer n and take the union over all n.*

- If $\mu_1(A) = \mu_2(A)$, $\mu_1(B) = \mu_2(B)$, and $B \subseteq A$, then $\mu_1(A \setminus B) = \mu_2(A \setminus B)$ by additivity, so $A \setminus B$ is also in $\mathcal{L}$.
- If $\mu_1(A_n) = \mu_2(A_n)$ for all n in an increasing sequence $\{A_n\}$, we can define $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, and so on. From the previous point, $\mu_1(B_i) = \mu_2(B_i)$ for all i, so by countable additivity we have $\mu_1(\bigcup B_i) = \mu_2(\bigcup B_i)$. Since $\mu_1(\bigcup B_i) = \mu_1(\bigcup A_i) = \mu_1(A)$ (and same for μ_2), $\mu_1(A) = \mu_2(A)$ as desired.

Thus the π-λ theorem tells us that $\sigma(\mathcal{A}) = \mathcal{L}$, so $\mu_1 = \mu_2$ and the measures are the same. □

This proof technique is quite powerful – we only had to verify that $\mu_1(A) = \mu_2(A)$ for a small subset of all measurable subsets, and $\mathcal{L}$ helped us bootstrap that to a more general statement (even though there isn't an easy way to write down what an arbitrary element of $\sigma(\mathcal{A})$ even looks like). Now, we'll actually construct a measure that works:

Proof of existence for Theorem 17. We begin with a definition:

Definition 24

An **outer measure** on Ω is a map $\nu : 2^\Omega \to [0, \infty)$ which is monotone and countably subadditive.

An outer measure is basically a "crude" measure which has certain desirable characteristics of a measure but not all of them, and the example we'll use is the following:

Definition 25

Let μ be a measure on an algebra $\mathcal{A}$. Then for all $E \subseteq \Omega$, define

$$\mu^*(E) = \inf\left\{\sum_{i=1}^{\infty} \mu(A_i) : A_i \in \mathcal{A}, E \subseteq \bigcup_i A_i\right\}.$$

In other words, $\mu^*(E)$ is the minimum measure we need to cover E using sets in $\mathcal{A}$. We can verify that μ^* is indeed an outer measure:

- It is monotone, because any covering of $B \supseteq A$ also covers A, meaning that $\mu^*(B) \geq \mu^*(A)$.
- It is also countably subadditive, because the union of coverings of A_i also covers $\bigcup_i A_i$, and we're still using a countable number of sets in $\mathcal{A}$ to do so.

Lemma 26

The outer measure μ^* extends μ on $\mathcal{A}$ (the set of finite disjoint unions of half-open intervals) – in other words, $\mu^*(A) = \mu(A)$ for all $A \in \mathcal{A}$.

Proof. First of all, we know that $\mu^*(A) \leq \mu(A)$, since A covers itself. Suppose for the sake of contradiction that $\mu^*(A) < \mu(A)$. Then there is some countable union of half-open intervals that covers A with measure less than $\mu(A)$ (since the infimum over all covering is less than $\mu(A)$), and we can choose the covering so that each interval I in the covering is contained within one of the intervals of A (if I covers multiple intervals, then split it up, and then cut off the endpoints, which only decreases $\sum \mu^* A$). Since A is made up of finitely many intervals (say N of them), by the pigeonhole principle, there is some interval $I \in A$ covered by intervals $I_n \in \mathcal{I}$ such that $\sum_n \mu^*(I_n) < \mu(I)$. Pick the interval I with the largest deviation, and we know that

$$\sum_n \mu^*(I_n) = \mu(I) - \varepsilon$$

for some $\varepsilon > \frac{\mu(A) - \mu^*(A)}{N} > 0$. However, $\mu^*(I_n) = \mu(I_n)$ (this can be checked very similarly to part (3) of Proposition 16, covering a compact interval with an open covering), and then again by part (3) of Proposition 16 we have subadditivity of μ over intervals, meaning

$$\mu(I) \leq \sum_n \mu(I_n),$$

a contradiction. Thus we cannot have $\mu^*(A) < \mu(A)$, and $\mu^*(A) = \mu(A)$ as desired. □

In other words, the outer measure assigns a value to every subset of our space Ω, and it does so correctly on $\mathcal{A}$. We've already seen that doing this assignment naively won't give us a valid measure, so to refine this argument, we need to find a suitable subcollection $\mathcal{A}^* \subseteq 2^\Omega$ such that $\mathcal{A}^*$ is a σ-algebra and $\mu^*|_{\mathcal{A}^*}$ is a measure. If we can find such a set, then $\mathcal{A} \subseteq \mathcal{A}^*$ implies that $\sigma(\mathcal{A}) \subseteq \sigma(\mathcal{A}^*)$, which means we will have successfully defined our extension on $\sigma(\mathcal{A})$.

Definition 27

A subset $E \subseteq \Omega$ is **measurable** with respect to μ^* if for all $F \subseteq \Omega$, we have

$$\mu^*(F) = \mu^*(F \cap E) + \mu^*(F \cap E^c).$$

Measurability will turn out to be the "niceness" property that we want, and the set on which we will define our measure is

$$\mathcal{A}^* = \{E \subset \Omega : E \text{ is measurable with respect to } \mu^*\}.$$

First of all, we need to make sure that we are working with a large enough subcollection:

Lemma 28

With the definition of $\mathcal{A}^*$ above (and the algebra $\mathcal{A}$ in the theorem statement), $\mathcal{A} \subseteq \mathcal{A}^*$.

Proof of lemma. We wish to show that any element of $\mathcal{A}$ is measureable with respect to μ^*, or in other words that for any A in $\mathcal{A}$, we have

$$\mu^*(F) = \mu^*(F \cap A) + \mu^*(F \cap A^c).$$

One direction is clear: we have

$$\mu^*(F) \leq \mu^*(F \cap A) + \mu^*(F \cap A^c)$$

because the union of the coverings of $F \cap A$ and $F \cap A^c$ always covers the left side. Now, suppose the left side is covered by some countable union of intervals $I_1, I_2, \cdots$. We will show that we can cover the right side with at most 4ε more measure for any $\varepsilon > 0$ (which will prove that the infimum on the right-hand side is at most 4ε the infimum on the left-hand side).

Since A is a finite collection of intervals, each interval I_k in the covering of F can only have finitely many transitions between the parts contained in A and A^c, so there are a countable number of changes between A and A^c overall. Break up the intervals at those changes, which doesn't change the total μ of the covering. We can now cover $F \cap A$ and $F \cap A^c$ as follows: $F \cap A$ is a subset of $(\bigcup_{k \geq 1} I_k) \cap A$ (since the I_ks cover F), which is a countable union of intervals (each possibly closed or open). Similarly, $F \cap A^c$ is a subset of $(\bigcup I_k) \cap A^c$, which is another countable union of intervals. Extend both sides of the ith interval in $(\bigcup I_k) \cap A$ (which has countably many intervals) by $\frac{\varepsilon}{2^i}$, and similarly extend both sides of the jth interval in $(\bigcup I_k) \cap A^c$ by $\frac{\varepsilon}{2^j}$. (Also, use this extra length to turn all of the intervals into half-open ones.) We've then used 4ε more μ than in our covering of A, but we've covered both $F \cap A$ and $F \cap A^c$. Because this argument works for any covering of F,

$$\mu^*(F) + 4\varepsilon \geq \mu^*(F \cap A) + \mu^*(F \cap A^c),$$

and taking $\varepsilon \to 0$ shows the other direction of the inequality, completing the proof. □

To finish our construction, we just need to show that $\mathcal{A}^*$ is a σ-algebra and that μ^* is countably additive on $\mathcal{A}^*$. First of all, $\mathcal{A}^*$ is closed under complementation, because E and E^c are symmetric in Definition 27 (so if E satisfies the condition, so does E^c). Towards showing closure under countable union, we will first prove that it's closed under finite union. By induction, it suffices to show that if $E_1, E_2 \in \mathcal{A}^*$, then $E_1 \cup E_2 \in \mathcal{A}^*$. By definition, this is equivalent to plugging $E_1 \cup E_2$ in Definition 27, and it suffices to just show the inequality

$$\mu^*(F) \geq \mu^*(F \cap (E_1 \cup E_2)) + \mu^*(F \cap (E \cup E_2)^c),$$

because the reverse inequality is true by subadditivity of μ^* (one of the properties of outer measure). To show this, note that

$$\mu^*(F \cap (E_1 \cup E_2)) \leq \mu^*(F \cap E_1) + \mu^*((F \setminus E_1) \cap E_2)$$

by subadditivity (the two sets on the right-hand side are disjoint and their union is $F \cap (E_1 \cup E_2)$), and we also have

$$\mu^*((F \setminus E_1) \cap E_2) = \mu^*(F \cap E_1^c) - \mu^*(F \setminus (E_1 \cup E_2))$$

by the measurability of E_2 applied to the set $F \cap E_1^c = F \setminus E_1$. Adding the two equations, the blue terms cancel, so

$$\mu^*(F \cap (E_1 \cup E_2)) + \mu^*(F \setminus (E_1 \cup E_2)) \leq \mu^*(F \cap E_1) + \mu^*(F \cap E_1^c) = \mu^*(F),$$

where the last step comes from E_1 being measurable. So $\mathcal{A}^*$ is indeed closed under finite unions – to prove that $\mathcal{A}^*$ is also closed under countable unions, we need to show that for any $E_i \in \mathcal{A}^*$, we also have $\bigcup_{i=1}^{\infty} E_i \in \mathcal{A}^*$. First of all,

for any disjoint $E_1, E_2 \in \mathcal{A}^*$, by the measurability of E_1 applied to the set $F \cap (E_1 \cup E_2)$, we have (for any set F)

$$\mu^*(F \cap (E_1 \cup E_2)) = \mu^*(F \cap (E_1 \cup E_2) \cap E_1) + \mu^*(F \cap (E_1 \cup E_2) \setminus E_1) = \mu^*(F \cap E_1) + \mu^*(F \cap E_2),$$

and thus by induction we see that $\boxed{\mu^*\left(F \cap \left(\bigcup_{i=1}^{n} E_i\right)\right) = \sum_{i=1}^{n} \mu^*(F \cap E_i)}$. To now actually prove our claim, first assume without loss of generality that the E_is are disjoint – because we already have closure under complementation, we can instead use the sets $M_1 = E_1$, $M_2 = E_2 \setminus E_1 = E_2 \cup E_1^c$, and so on, without changing the overall union of the E_is. Now define

$$A_n = \bigcup_{i=1}^{n} E_i, \quad A = \bigcup_{i=1}^{\infty} E_i.$$

Since we have just proved closure for finite unions, we know that $A_n \in \mathcal{A}^*$ for all n, and therefore

$$\mu^*(F) = \mu^*(F \cap A_n) + \mu^*(F \setminus A_n)$$

by the definition of measurability. So now applying the boxed equality to the first term and monotonicity to the second term, we have

$$\mu^*(F) \geq \sum_{i=1}^{n} \mu^*(F \cap E_i) + \mu^*(F \setminus A),$$

and taking $n \to \infty$ yields

$$\mu^*(F) \geq \sum_{i=1}^{\infty} \mu^*(F \cap E_i) + \mu^*(F \setminus A) \geq \mu^*(F \cap A) + \mu^*(F \setminus A)$$

by countable subadditivity of μ^*. The reverse inequality is true by subadditivity of outer measure, so we've verified the measurability of the countable union A, meaning that $\mathcal{A}^*$ is indeed a σ-algebra.

Finally, to prove that μ^* is indeed a measure on $\mathcal{A}^*$, we need to show countable additivity, and we can do this by showing inequalities in both directions. For any measurable sets A_i, we have

$$\mu^*\left(\bigsqcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$$

by countable subadditivity of the outer measure μ^*, meaning that

$$\mu^*\left(\bigsqcup_{i=1}^{\infty} A_i\right) \geq \mu^*\left(\bigsqcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu^*(A_i),$$

where in the last step we apply measurability of A_n to the set $\bigsqcup_{i=1}^{n} A_i$, then apply measurability of A_{n-1} to the set $\bigsqcup_{i=1}^{n-1} A_i$, and so on. Taking $n \to \infty$, we obtain the other direction of inequality, proving countable additivity of μ^*. Since μ^* extends μ and is defined on the σ-algebra $\mathcal{A}^*$ containing $\mathcal{A}$, we have successfully defined a measure on $\sigma(\mathcal{A})$. □

Putting the proofs of existence and uniqueness together, we have finally extended our measure to $\sigma(\mathcal{I})$, as desired. And in particular, using the Stieltjes measure function $F(x) = x$, we have finally (after two lectures of work) successfully defined the uniform (Lebesgue) measure on $[0, 1]$.

4 September 16, 2019

Our first homework assignment will be posted later today (after lecture); it will be due in class next Wednesday. There will be four or five problem sets in this class, and this first one will have us work a little more with properties of measures.

We'll start with a restatement of the material covered in the last two lectures. We proved Theorem 14, which states that for any nondecreasing right-continuous function $F : \mathbb{R} \to \mathbb{R}$, we have a unique measure μ_F on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ satisfying $\mu((a, b]) = F(b) - F(a)$. The proof has two parts – first, we extend μ from $\mathcal{I}$ (the set of half-open intervals of the form $(a, b]$) to $\mathcal{A}$ (the set of finite disjoint union of such intervals). Then, we extend μ from $\mathcal{A}$ to its Borel σ-algebra using the Carathéodory extension theorem; uniqueness follows from the π-λ theorem, and existence comes from the construction of the outer measure.

This is a class about probability, so we'll start to introduce random variables and their properties today. We'll see soon that these variables are specific examples of **measurable functions**, and that their properties (like expected value and variance) are specific examples of **Lebesgue integrals**.

Definition 29

A **probability space** is a measure space $(\Omega, \mathcal{F}, \mu)$ such that $\mu(\Omega) = 1$.

We will often denote μ as $\mathbb{P}$ for a probability space. A way to interpret this definition is that Ω is a space of possible outcomes, $\mathcal{F}$ is the set of possible events that can be assigned measures (some sets of outcomes are not "well-behaved" enough to be assigned a probability measure), and $\mathbb{P}$ is the measure itself.

Example 30

If we have a sequence of n independent fair coin tosses, we can write down its probability space as

$$\Omega = \{\text{heads}, \text{tails}\}^n = \{0, 1\}^n, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}(A) = \frac{|A|}{|\Omega|} \quad \forall A \subseteq \Omega.$$

In this case, the set of possible events $\mathcal{F}$ can just be the full powerset of Ω because we have a finite set. And the way we've defined the space, our probability measure is uniform on the 2^n possible sequences: for any individual event (that is, any set of size $|A| = 1$), we have $\mathbb{P}(A) = \frac{1}{2^n}$.

In order to study properties of more general probability spaces, we'll need to be more abstract. From here, let $(\Omega, \mathcal{F}, \mu)$ be any measure space, in particular allowing for $\mu(\Omega) = \infty$.

Definition 31

The **indicator function** for the set $A \in \mathcal{F}$, denoted $1_A(\omega)$ or $1\{\omega \in A\}$, is defined to be

$$1_A(\omega) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \in \Omega \setminus A. \end{cases}$$

Definition 32

A **simple function** $f : \Omega \to \mathbb{R}$ is a function that can be written in the form $f = \sum_{i=1}^n c_i 1_{A_i}$, where $c_i \in \mathbb{R}$ and A_i are disjoint sets in $\mathcal{F}$.

Definition 33

A **measurable function** f is one that can be pointwise approximated by simple functions, meaning that there exists a sequence of simple functions $\{f_n\}$ such that $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ for all $\omega \in \Omega$.

(One way to phrase this is that the "nicely-behaved" functions are those which can be built up from indicator functions using pointwise limits of linear combinations.)

Definition 34

Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{G})$ be two measurable spaces. A **measurable mapping** is a function $f : \Omega \to S$ such that for every $G \in \mathcal{G}$, the preimage $f^{-1}(G)$ is an element of $\mathcal{F}$.

In other words, if we take any "well-behaved" set in S, its preimage is also well-behaved in Ω.

Proposition 35

A measurable function f on $(\Omega, \mathcal{F})$ (as defined in Definition 33) is equivalent to a measurable mapping $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ as defined in Definition 34.

Proof. This is an exercise left for our homework (we must show that being measurable in each definition also means being measurable in the other sense). □

The definition of a measurable function is more concrete, while the mapping is a bit more abstract, but the latter is particularly useful because a measure μ on the domain space Ω naturally induces a measure on the target space S:

Proposition 36

If μ is a measure on $(\Omega, \mathcal{F})$, and we have a mapping $f : (\Omega, \mathcal{F}) \to (S, \mathcal{G})$, then we have a measure ν on $(S, \mathcal{G})$ called the **pushforward measure**, defined as

$$\nu(G) = \mu(f^{-1}(G)) \quad \forall G \in \mathcal{G}.$$

This is typically written as $\nu = f_*\mu$ or $f_\#\mu$.

(Soon in this lecture, we'll connect this abstract idea to the idea of a "distribution" from undergraduate probability.)

Proposition 37

If f, g are measurable, then $af + bg$ is measurable for all $a, b \in \mathbb{R}$, and so is $f \cdot g$.

Proof sketch. Write f and g as pointwise limits of simple functions and do some bookkeeping. Alternatively, we can check using the other definition of measurability, verifying that the preimage of sets of the form $(a, b]$ (it suffices to check $(a, \infty]$) are measurable. □

Definition 38

A **random variable** is a (real-valued) measurable function defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. (We often use the variables X or Y rather than f.) More generally, an $(S, \mathcal{G})$-valued random variable is a measurable mapping $(\omega, \mathcal{F}) \to (S, \mathcal{G})$.

Random variables will usually be real-valued in this class, but we may occasionally specify a different target space.

Fact 39 (Helpful notation)
Since f is a function from Ω to $\mathbb{R}$, we will often consider the pre-image $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ of a set of real numbers B. We will denote this set $\{X \in B\}$, so for example, $\{X \geq \frac{1}{2}\} = \{\omega \in \Omega : X(\omega) \geq \frac{1}{2}\}$.

To see why this is good notation, remember that X is a mapping from some space $(\Omega, \mathcal{F}, \mathbb{P})$ to the real line $(\mathbb{R}, \mathcal{B})$, so if we take the pushforward measure $X_*\mathbb{P}$, we get the **distribution** or **law** of the random variable X, often denoted $\mathcal{L}_X$. We'll go through the symbol-pushing to make sure we understand how it behaves: $\mathcal{L}_X$ is a measure on our target space $(\mathbb{R}, \mathcal{B})$, meaning that it takes in a measurable subset $B \in \mathcal{B}$ and outputs

$$\mathcal{L}_X(B) = (X_*(\mathbb{P}))B = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}),$$

the "probability" of X mapping to B. So measure theory gives a more rigorous meaning to $\mathbb{P}(X \in B)$, which is notation that we like to use.

Example 40
Looking back at the coin-tossing example, where we have $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ a uniform measure on Ω, consider the two random variables $X(\omega) = \omega_1 = 1\{\text{first toss is head}\}$ and $Y(\omega) = \sum_{i=1}^n \omega_i = \text{number of heads}$.

We can visualize Ω as the vertices of a hypercube $\{0,1\}^n$, where the map Y sends a vertex $(x_1, \cdots, x_n)$ to $x_1 + \cdots + x_n$. (So random variables often "condense information" and are not one-to-one.) Then the law $\mathcal{L}_Y$ is a measure on $(\mathbb{R}, \mathcal{B})$ supported on $\{0, 1, \cdots, n\}$, so we just need to find the weight assigned to each integer to describe the distribution. In formal notation, we have

$$\mathcal{L}_Y(\{k\}) = \mathbb{P}\left(Y^{-1}(\{k\})\right) = \mathbb{P}\left(\{\omega \in \{0,1\}^n : Y(\omega) = k\}\right) = \frac{\binom{n}{k}}{2^n},$$

and this is the familiar **binomial distribution** with parameters $(n, \frac{1}{2})$.

Now that we have a more explicit description of our random variables, we'll move on to **Lebesgue integration**. For this part, assume that $(\Omega, \mathcal{F}, \mu)$ is a ***σ*-finite measure space**, meaning that there exists a sequence $\Omega_n \uparrow \Omega$ such that $\mu(\Omega_n) < \infty$ for all n. (For example, the real line with the (uniform) Lebesgue measure would work by taking $\Omega_n = [-n, n]$.) Suppose that we have a measurable function $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$; our goal is to define the quantity

$$\int_\Omega f\, d\mu = \int_\Omega f(\omega) d\mu(\omega),$$

such that in the case where μ is a probability measure $\mathbb{P}$, we get the mean of f. Since we build up measurable functions from simple functions, it makes sense to do the same for integrals:

1. For any **simple** function $f = \sum_{i=1}^n c_i 1_{A_i}$, define

$$\int f\, d\mu = \sum_{i=1}^n c_i \mu(A_i).$$

It's bookkeeping to check that this integral doesn't depend on the representation f as a simple function. Specifically, if we can write f as $\sum_{i=1}^n c_i 1_{A_i} = \sum_{j=1}^m d_j 1_{B_j}$), then each A_i must be contained within the region where f is nonzero,

which is contained within $B_1 \cup \cdots \cup B_n$. Thus,

$$\sum_{i=1}^{n} c_i \mu(A_i) = \sum_{i=1}^{n} c_i \mu\left(\bigcup_{j=1}^{m} A_i \cap B_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_i \mu\left(A_i \cap B_j\right).$$

Similarly, $\sum_{j=1}^{m} d_j \mu(B_j) = \sum_i \sum_j d_j \mu\left(A_i \cap B_j\right)$. But because the A_is are disjoint and so are the B_js we must have $c_i = d_j$ (equal to the value that f takes in the region) whenever $A_i \cap B_j$. Thus the two expressions are indeed equal.

2. Next, we define the integral for **bounded measurable functions**. For any measurable function $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ with $\sup_{\omega\in\Omega} |f(\omega)| < \infty$ and $\mu(\{f \neq 0\}) < \infty$ (the latter meaning that the function has **bounded support**), define

$$\int f\, d\mu = \sup\left\{\int \phi d\mu : \phi \leq f, \phi \text{ simple function}\right\} = S.$$

It may seem arbitrary to approach f from below, so we might ask whether it also makes sense to approximate from above. It turns out both methods yield the same answer:

Proposition 41

Let $I = \inf\left\{\int \phi d\mu : \phi \geq f, \phi \text{ simple function}\right\}$ and take the definition of S from above. Then $I = S$.

Proof. We clearly have $S \leq I$, because any $\phi \leq f$ in the definition of S is pointwise smaller than any $\phi \geq f$ in the definition of I (and thus the integral $\int \phi d\mu$ is also smaller). To show that $I \leq S$, we will sandwich our function between simple functions. We have

$$f_m^{\text{lower}} \leq f \leq f_m^{\text{upper}}, \text{ where } f_m^{\text{upper}} = \frac{\lceil mf \rceil}{m}, \quad f_m^{\text{lower}} = \left(\frac{\lceil mf \rceil - 1}{m}\right) \cdot 1\{f \neq 0\},$$

where the idea is that f_m^{upper} and f_m^{lower} sift the function f in increments of $\frac{1}{m}$, and the $1\{f \neq 0\}$ term ensures that both of these functions are zero whenever $f = 0$. Additionally, if $\sup|f| = M$ (finite by assumption here), then $f_m^{\text{upper}} \leq M + 1$ and $f_m^{\text{lower}} \geq -M - 1$. So both f_m^{lower} and f_m^{upper} are bounded and only take on values that are integer multiples of $\frac{1}{m}$, so they are simple. We can thus write

$$I - S \leq \int f_m^{\text{upper}} d\mu - \int f_m^{\text{lower}} d\mu = \frac{1}{m} \cdot \mu(\{f \neq 0\}),$$

because those two integrals only differ by $\frac{1}{m}$ and only on the region (of finite measure) where $f \neq 0$. Taking m to infinity, we find that $I - S \leq 0$, establishing the other inequality and proving that $I = S$. □

3. Next, we define the integral on **nonnegative** (but not necessarily bounded) measurable functions f by approximating with bounded functions from below:

$$\int f\, d\mu = \sup\left\{\int h d\mu : 0 \leq h \leq f,\ h \text{ bounded with } \mu(\{h \neq 0\}) < \infty\right\}.$$

Proposition 42

The above definition is equivalent to setting $\int f\, d\mu = \lim_{m\to\infty} \int_{\Omega_m} \min(f, m) d\mu$, where we integrate over sets $\Omega_m \uparrow \Omega$ of finite measure (meaning $\mu(\Omega_m) < \infty$ for each m).

This gives us a more explicit formula for the integral, since we don't need to actually compute the supremum over all possible h and can just take the "truncated" functions $\min(f, m)$.

Proof. Let $I_m = \int_{\Omega_m} \min\{f, m\} d\mu$ and $I = \int f\, d\mu$; we wish to show that $I_m \uparrow I$. The sequence $\{I_m\}$ is increasing as a function of m, because we're integrating over a larger set and integrating a larger nonnegative function as m grows. And we know that $I \geq \lim_{m\to\infty} I_m$, because each I_m is an example of a function h in the definition above, so we just need to show the other inequality.

By definition, for any $\varepsilon > 0$, we can find a bounded function h such that $\int_\Omega h d\mu \geq I - \varepsilon$. Because h is bounded, we can find some L such that $|h| \leq L$, and then for all $m \geq L$ (and thus $m \geq h$ everywhere),

$$\begin{aligned}\int_\Omega h d\mu &= \int_\Omega \min\{h, m\} d\mu \\ &= \int_{\Omega_m} \min\{h, m\} d\mu + \int_{\Omega\setminus\Omega_m} \min\{h, m\} d\mu \\ &\leq I_m + L \cdot \mu\left((\Omega \setminus \Omega_m) \cap \{h \neq 0\}\right),\end{aligned}$$

where we use that $h \leq f$ in the first term and $h \leq L$ in the second. Now L is a constant, and as $m \to \infty$, $\Omega \setminus \Omega_m \downarrow \varnothing$, so $(\Omega \setminus \Omega_m) \cap \{h \neq 0\} \downarrow \varnothing$. Also, the measure of any set $(\Omega \setminus \Omega_m) \cap \{h \neq 0\}$ is finite, because $\mu(\{h \neq 0\})$ is finite by definition. So by part (4) of Proposition 9, $\mu\left((\Omega \setminus \Omega_m) \cap \{h \neq 0\}\right)$ gets arbitrarily small, and in particular there will be some sufficiently large m such that

$$\int_\Omega h d\mu \leq I_m + \varepsilon.$$

Putting our inequalities together, we thus find that for any $\varepsilon > 0$,

$$I - \varepsilon \leq \int_\Omega h d\mu \leq I_m + \varepsilon,$$

so taking $\varepsilon \to 0$ yields $I \leq I_m$, giving us the other direction of the inequality and thus $I_m \uparrow I$. □

4. For the last step, we want to extend from nonnegative measurable functions to **all measurable** functions $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$. To do this, set

$$f = f_+ - f_-, \text{ where } f_+ = \max(f, 0), \quad f_- = \max(-f, 0)$$

(we can check that f_+ and f_- are both measurable if f is measurable) and define

$$\int f\, d\mu = \int f_+ d\mu - \int f_- d\mu,$$

where we compute the two integrals on the right-hand side using the previous step. There are a few cases here: if both integrals are finite, then $\int f\, d\mu$ is finite (in fact, $\int |f| d\mu < \infty$, and we say that f is **integrable**). Also, if only one of the integrals is infinite, then we get ∞ or $-\infty$. But if we have $\infty - \infty$ on the right-hand side, we say that $\int f\, d\mu$ is undefined.

Definition 43

The **expectation** of a random variable X is

$$\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\Omega) = \int_\Omega X d\mathbb{P}.$$

Defining this integral is useful because any function $f : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ that is Riemann-integrable is measurable, and the **Riemann integral is the same as the Lebesgue integral** with respect to the Lebesgue measure. On the other hand, the function $1_{\mathbb{Q}}$ is not Riemann integrable (because it is very "spiky"), but the Lebesgue integral is perfectly

well-defined – the Lebesgue measure of a countable set is 0, so we have $\int_{\mathbb{R}} f = 0$. Additionally, we can also interpret discrete sums as Lebesgue integrals by saying that

$$\sum_{i=1}^{n} f(i) = \int_{\mathbb{R}} f\, dn,$$

where n is the counting measure $n(A) = |A \cap \mathbb{N}|$. So Lebesgue integration can cover both very "smooth" and very discrete cases, and even some in between (which we'll see on the homework), which is useful because we have both discrete and continuous random variables in probability. And finally, we'll also soon see that Lebesgue integrals have nice convergence properties, which will be important for many of the results we show in this class!

5 September 18, 2019

Last time, we defined the Lebesgue integral. There are some problems about measurability on our homework (we should redownload the document because some typos have been fixed), so we'll start with a review of last lecture. (The homework is due next Wednesday during class, and Professor Sun says that it takes a while, so we should not leave it until the last minute.)

Recall that a function f is **simple** if it can be written as a linear combination $f = \sum_{i=1}^{n} c_i 1_{A_i}$ of indicator functions (where the A_is may have infinite measure). A **measurable** function is then the pointwise limit of simple functions (remember that on our homework, we show that this is equivalent to having $f^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}_{\mathbb{R}}$). So to define the Lebesgue integral, we suppose that $(\Omega, \mathcal{F}, \mu)$ is σ-finite, meaning that there exist sets $\Omega_m \uparrow \Omega$ with $\mu(\Omega_m) < \infty$ for each m. Then we can define our integral first over simple functions as

$$f = \sum_{i=1}^{n} c_i 1_{A_i} \implies \mu(A_i) < \infty \implies \int_{\Omega} f\, d\mu = \sum_{i=1}^{n} c_i \mu(A_i)$$

and using this to define the integral of any nonnegative function f:

$$\int f\, d\mu = \sup \left\{ \int h d\mu : 0 \le h \le f, h \text{ simple function}, \ \mu(\{h > 0\}) < \infty \right\}.$$

(We did this in two steps last time by defining the integral for bounded functions first, but this definition is equivalent.) Then we define the integral of any real-valued f to be the integral of f_+ (the positive part) minus the integral of f_- (the negative part), and $\int_{\Omega} f\, d\mu = \int_{\Omega} f(\omega) d\mu(\omega)$ is then called the **Lebesgue integral** of f. Importantly, this construction works for any σ-finite measure, not just the Lebesgue measure.

Last lecture, we also started talking about random variables: given a measurable function X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define its **law** to be the pushforward measure $\mathcal{L}_X = X_{\#}\mathbb{P}$. (This is a useful in connecting the mean of a random variable $\mathbb{E}[X]$ to the Lebesgue integral $\int X d\mathbb{P}$.)

Remark 44. *We can think of computing the Riemann integral as summing the area of a bunch of rectangles, obtained by partitioning the* **domain** *into small steps. In contrast, because the Lebesgue integral is built up by simple functions, we can think of the Lebesgue integral as summing rectangles coming from partitioning the* **target space** *into small steps.*

It can be proved that a function $f : [a, b] \to \mathbb{R}$ which is Riemann integrable is also Lebesgue integrable, and the two integrals both agree. (Basically, we can show this for simple functions and then approximate a Riemann integrable function from above and below with simple functions.) But as we mentioned last time, Lebesgue integration is more powerful because it can encode discrete sums (which Riemann integration cannot); additionally, we can also define

the Lebesgue integral in more abstract spaces $(\Omega, \mathcal{F}, \mu)$, while the Riemann integral requires some sort of Euclidean structure. And today, we'll talk about a third advantage of the Lebesgue integral, which is that it is well-suited for convergence theorems.

Remark 45. *In practice, the Lebesgue integral is rarely directly computed from the definition. Instead, we try to write down more explicit expressions – for example, if we can write $f = h + 1_{\mathbb{Q}}$ for a continuous function h, then we can Riemann integrate h and deal with $1_{\mathbb{Q}}$ with a discrete summation. And even if the Lebesgue integral is defined on a more abstract space, we can often push forward onto the real line to get a more explicit expression, but we often need convergence properties for that.*

Convergence theorems answer questions like "if $f_n \to f$ converges pointwise, does $\int f_n d\mu \to \int f d\mu$?" The answer turns out to be **no** in general:

Example 46

If we define the functions $f_n = n \cdot 1_{(0, \frac{1}{n})}$, the integral of each f_n (over the real line) is 1, but the sequence f_n converges pointwise to 0, so the limit has integral 0.

Our goal is thus to show conditions under which the integrals do converge. For the rest of this lecture, we'll assume that $(\Omega, \mathcal{F}, \mu)$ is a σ-finite measure space, and that f_n, f, g_n, g are all measurable functions on $(\Omega, \mathcal{F})$. We'll also let $\Omega_k \uparrow \Omega$ be a sequence of spaces with $\mu(\Omega_k) < \infty$ for all k.

Definition 47

A sequence of functions f_n converges to f **with respect to μ almost everywhere** if the numbers $f_n(\omega) \to f(\omega)$ converge except for ω in a set of μ-measure zero. Similarly, $f_n \to f$ converges **in μ-measure** if for all $\varepsilon > 0$, $\lim_{n \to \infty} \mu(\{\omega : |f_n(\omega) - f(\omega)| \geq \varepsilon\}) = 0$.

It will be a future homework problem to show that pointwise convergence implies μ-almost-everywhere convergence, which implies convergence in measure, but that the converses are not true. Our first convergence theorem will be the easiest to prove but unfortunately the least useful:

Theorem 48 (Bounded convergence theorem)

Suppose that a sequence of random variables f_n are uniformly bounded, meaning that there is some $M < \infty$ such that $|f_n| \leq M$ for all n. Also suppose that $\mu(\{\omega : f_n(\omega) \neq 0 \text{ for any } n\}) < \infty$. Then if $f_n \to f$ in μ-measure, then $\int f_n d\mu \to \int f d\mu$.

In other words, we have convergence of integrals if the functions are uniformly bounded in value and also in support.

Proof. Define $E = \{\omega : f_n(\omega) \neq 0 \text{ for some } n\}$. By assumption, notice that the limit f satisfies $|f| \leq M$ everywhere except a set of measure zero, and the set $\{f \neq 0\}$ is contained in E (because outside of E all of the f_ns are zero, so their limit is also zero). Thus, the only contributions to any of the integrals will come from E, and we can ignore $\Omega \setminus E$. Fix some $\varepsilon > 0$, and define $B_n = \{|f_n - f| \geq \varepsilon\}$ (this is a subset of E). We can now bound the difference

between our integrals as

$$\begin{aligned}\left|\int_\Omega f\,d\mu - \int_\Omega f_n d\mu\right| &= \left|\int_E (f_n - f)d\mu\right| \\ &= \left|\int_{E\setminus B_n} (f_n - f)d\mu\right| + \left|\int_{B_n} (f_n - f)d\mu\right| \\ &\le \varepsilon\mu(E\setminus B_n) + 2M\mu(B_n) \\ &\le \varepsilon\mu(E) + 2M\mu(B_n),\end{aligned}$$

where we use the fact that $|f_n - f| \le 2M$ everywhere and that $|f_n - f| \le \varepsilon$ in $E \setminus B_n$. But $\mu(B_n) \to 0$ by definition of convergence in μ-measure, so we have

$$\limsup_{n\to\infty}\left|\int (f_n - f)d\mu\right| \le \varepsilon\mu(E).$$

Because $\mu(E)$ is finite, sending $\varepsilon \to 0$ shows that $\int_\Omega f_n d\mu$ does indeed converge to $\int_\Omega f\,d\mu$, as desired. □

Theorem 49 (Fatou's lemma)
Let f_n be a sequence of nonnegative functions. Then

$$\int \left(\liminf_{n\to\infty} f_n\right) d\mu \le \liminf_{n\to\infty}\left(\int f_n d\mu\right).$$

For this result to make sense, we're assuming that if f_n are all measurable, then $\liminf f_n$ is also measurable – that's an exercise we can check ourselves. And as a note, $\liminf_{n\to\infty} f_n$ is only the pointwise liminf of the f_ns μ-**almost-everywhere**, but that's okay because we're integrating over it anyway and a measure-zero set does not change the integral.

Proof. Let $f = \liminf f_n$. By definition,

$$\liminf_{n\to\infty} f_n = \sup_{n\ge 1}\left(\inf_{\ell\ge n} f_\ell\right)$$

(recall that we can use sup instead of lim because the inner term $(\inf_{\ell\ge n} f_\ell)$ is nondecreasing with n). Let the inner term be $g_n = \inf_{\ell\ge n} f_\ell$; notice that $0 \le g_n \le f_n$ for any n and $g_n \uparrow f$. Because g_n converges pointwise to f, it also converges in μ-measure, so by the bounded convergence theorem (Theorem 48 above) we have

$$\boxed{\int_{\Omega_k} \min\{f, k\}} = \lim_{n\to\infty}\int_{\Omega_k}\min\{g_n, k\}d\mu.$$

(Here we are importantly using that Ω_k has finite measure by definition.) However, we have

$$\lim_{n\to\infty}\int_{\Omega_k}\min\{g_n,k\}d\mu \le \lim_{n\to\infty}\int g_n d\mu \le \boxed{\liminf_{n\to\infty}\int f_n d\mu},$$

where the first step comes from removing the upper cap on the function and also expanding the space Ω_k, and the second step comes from $g_n \le f_n$. By Proposition 42, if we take $k \to \infty$, $\int_{\Omega_k}\min\{f,k\}$ converges to $\int f\,d\mu$ (because all functions here are nonnegative), and thus the inequality between the two boxed terms yields the desired result. □

(If we look back at Example 46 and we plug in the functions $f_n = n\cdot 1_{(0,\frac{1}{n})}$ into Fatou's lemma, we see that $0 = \int \liminf_{n\to\infty} f_n d\mu$ is indeed at most $1 = \liminf_{n\to\infty}\int f_n d\mu$.) Fatou's lemma may look unmotivated, but it is useful because of its applications to some powerful results:

Theorem 50 (Monotone convergence theorem)
Let $f_n \geq 0$ be a sequence of nonnegative functions, and suppose that $f_n \uparrow f$. Then $\int f_n d\mu \uparrow \int f d\mu$, where integrals are allowed to be infinite.

Proof. Since the Lebesgue integral is monotone (if $f \leq g$, then $\int f d\mu \leq \int g d\mu$), $\int f_n d\mu$ is a nondecreasing sequence of extended real numbers, and it converges to some limit $I \leq \int f d\mu$ (here using monotonicity from $f_n \leq f$). For the other inequality, Fatou's lemma tells us that

$$\int f d\mu = \int \lim_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty} \left(\int f_n d\mu \right) = \lim_{n\to\infty} \left(\int f_n d\mu \right) = I.$$

Thus the $\int f_n d\mu$s indeed increase to $\int f d\mu$, as desired. □

Theorem 51 (Dominated convergence theorem)
Let f_n be a sequence of functions, and suppose there is some **integrable** function g (that is, $\int |g| d\mu < \infty$) such that $|f_n| \leq g$ for all n. If $f_n \to f$ μ-almost-everywhere, then $\int f_n d\mu \to \int f d\mu$.

Proof. By assumption, $-g \leq f_n \leq g$ for all n, so $f_n + g \geq 0$ and $g - f_n \geq 0$. Since f is the limit of the f_ns, it is also their lim inf, so by Fatou's lemma,

$$\int (g \pm f) d\mu \leq \liminf_{n\to\infty} \int (g \pm f_n) d\mu.$$

Cancelling out the gs and then rewriting this inequality with both $+f$ and $-f$, we find that

$$\limsup_{n\to\infty} \int f_n d\mu \leq \int f d\mu \leq \liminf_{n\to\infty} \int f_n d\mu,$$

but both the left and right terms are $\lim_{n\to\infty} \int f_n d\mu$, so we have our desired equality. □

(With these last two theorems, it's also useful to see how Example 46 fails each of the theorem assumptions.) We'll now turn to an application of these results in probability: suppose we're working on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we have a random variable Y with mean $\mathbb{E}[Y] = \int_\Omega Y d\mathbb{P}$. Additionally, suppose that Y can be written as $f(X)$, where X is another random variable. In other words, we can describe the random variables with the following diagram:

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (S, \mathcal{G}, \mu) \xrightarrow{f} (\mathbb{R}, \mathcal{B}, \nu), \quad Y = f \circ X$$

Since μ is the pushforward measure under X of $\mathbb{P}$, meaning that $\mu = X_\# \mathbb{P} = \mathbb{P} \circ X^{-1}$, and ν is the pushforward of μ under f and also of $\mathbb{P}$ under Y, meaning that $\nu = f_\# \mu = Y_\# \mathbb{P}$, it's natural to ask if we can compute the expected value of Y in both ways, giving us the **change of variables formula**

$$\int_\Omega Y d\mathbb{P} \stackrel{?}{=} \int_S f d\mu.$$

Such a result would be useful if X is a real-valued random variable, because S would be $\mathbb{R}$ and the integral $\int_S f d\mu$ could be computed more explicitly. We'll start with a simple case to see that this does make sense:

Example 52 (Change of variables formula on a finite state space)

Suppose Ω is finite, so we can describe $\mathbb{P}$ with a function p satisfying $\sum_{\omega\in\Omega} p(\omega) = 1$. Any function Y is then a simple function, since $Y = \sum_{\omega\in\Omega} Y(\omega)1_{\{\omega\}} = \sum_{\omega\in\Omega} f(X(\omega))1_{\{\omega\}}$, so the expectation of Y is (by definition)

$$\mathbb{E}[Y] = \int_\Omega Y\,d\mathbb{P} = \boxed{\sum_{\omega\in\Omega} f(X(\omega))p(\omega)}.$$

But f is also a simple function (because it takes on only finitely many values, which are the values of $Y(\omega)$), so (here we only care about the values of f on the finitely many points $X(\Omega)$)

$$f = \sum_{x\in S} f(x)1_{\{x\}} \implies \int_S f\,d\mu = \boxed{\sum_{x\in S} f(x)\mu(\{x\})}.$$

And now for any $x \in X(\Omega)$,$\mu(\{x\}) = (X_\#\mathbb{P})(\{x\}) = \mathbb{P}(X^{-1}(\{x\}))$, which is the measure of the set of ωs sent to x under X (and thus is the sum of the $p(\omega)$s). So the two sums both add up $f(X(\omega))p(\omega)$ over all $\omega \in \Omega$, proving that the two integrals are equal.

However, it turns out that the more general formula does require more work and is not just symbol pushing – we'll see our convergence theorems in action:

Theorem 53 (Change of variables formula)

Suppose we have maps $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (S, \mathcal{G}, \mu) \xrightarrow{f} (\mathbb{R}, \mathcal{B}, \nu)$, and let $Y = f(X)$. If either $f \geq 0$ (f is nonnegative) or $\int |f(X)|d\mathbb{P} < \infty$ (Y is integrable), then

$$\mathbb{E}[Y] = \int_\Omega f(X(\omega))d\mathbb{P}(\omega) = \int_S f\,d\mu = \int_S f(x)d(\mathbb{P}\circ X^{-1})(x).$$

(Here, the first and last equalities are definitions of $\mathbb{E}[Y]$ and $\int_S f\,d\mu$, but including them explains the name "change of variables:" the first and last integral change the variable of integration from ω to x.)

Proof. This is a pretty standard proof method, so we should make a note of it: whenever we want to show a result about Lebesgue integrals, **first prove it for indicators**, and then built it up for more and more general functions.

- First, assume $f = 1_B$ for some set $B \in \mathcal{G}$. Since f is simple on S,

  $$\int_S f\,d\mu = \mu(B) = \mathbb{P}(X^{-1}(B)).$$

  But then $Y = f(X) = 1_B(X)$ is a simple function on Ω, specifically $1_{X^{-1}(B)}$, and thus $\int_\Omega Y\,d\mathbb{P} = \mathbb{P}(X^{-1}(B))$ as well. Thus the formula holds when f is an indicator function.

- Since simple functions are finite linear combinations of indicator functions, it follows by linearity of the Lebesgue integral that $\int_S f\,d\mu = \int_\Omega f(X)d\mathbb{P}$ for all simple functions.

- Next, we'll prove the formula for general nonnegative functions $f \geq 0$ (this is one of the conditions in the theorem statement). The useful trick here is to define the functions $f_n = \min\left\{\frac{\lfloor f2^n\rfloor}{2^n}, 2^n\right\}$. (This wouldn't work with n in place of 2^n – we want to define these f_ns to be nondecreasing.) Each f_n is a simple function, and $f_n \uparrow f$ pointwise. From the previous step, we know that $\int f_n d\mu = \int f_n(X)d\mathbb{P}$ for all n, so by the monotone convergence theorem (because $f_n(X) \to f(X)$ pointwise) those integrals converge to $\int f(X)d\mathbb{P} = \int Y\,d\mathbb{P}$, as desired.

- Finally, for a general measurable function f, we prove the result by splitting f into its positive and negative parts, and this only works if Y is integrable.

□

As we mentioned, this formula is most commonly used when we're trying to compute a Lebesgue integral. If we want to find $\mathbb{E}X$ for some random variable X on a complicated space $(\Omega, \mathcal{F}, \mathbb{P})$, we can set up the change of variables formula as

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}, \mathcal{L}_X) \xrightarrow{\text{id}} (\mathbb{R}, \mathcal{B}, \mathcal{L}_X),$$

where this time we have $Y = X$. Theorem 53 then tells us that

$$\mathbb{E}[X] = \int_{\mathbb{R}} \text{id}(x) d\mathcal{L}_X(x),$$

which is what we expect – the expected value of X only depends on its properties on the real line, and we compute it as an integral of x with respect to the law of X.

6 September 23, 2019

This is a final reminder that our first homework assignment (which should be printed and submitted in class) is due on Wednesday. Two TAs will be reading and grading, so if we are handwriting our solutions, we should make sure to be neat. (Writing more is not generally helpful, and solutions should be aim to be "correct in some minimal way.") Also, because of a seminar this week, office hours have been changed to Monday 2–3 and Tuesday 11–12.

So far in this class, we've been considering measures on the one-dimensional space $\mathbb{R}$. Today, we'll be generalizing to higher dimensions – to do so, we'll actually need to study our construction on $\mathbb{R}$ in more detail. Recall that we started with the collection of half-open intervals equipped with a Lebesgue-Stieltjes function F, and we made the definitions

$$\mathcal{I} = \{(a, b]\}, \quad \mu((a, b]) = F(b) - F(a).$$

To get to a measure on $\mathcal{B}_{\mathbb{R}}$, we first extended μ to $\mathcal{A}$, which contained finite disjoint unions of elements of $\mathcal{I}$, and we had to check that μ was indeed countably additive on $\mathcal{A}$. Once we did this, we used the Caratheodory extension theorem to obtain a measure on $\sigma(\mathcal{A})$. This second step holds in an abstract setting, as long as $\mathcal{A}$ is an algebra (meaning it is closed under complementation and finite union) and μ is countably additive on $\mathcal{A}$. However, the first step (where we extend from $\mathcal{I}$ to $\mathcal{A}$) used some topological properties of $\mathbb{R}$. So if we want to do this step in general, we need to explain what properties of $\mathcal{I}$ are actually required:

Definition 54

A **semialgebra** $\mathcal{S}$ is a set of subsets closed under finite intersection and semiclosed under complementation (meaning that if $E \in \mathcal{S}$, then $\Omega \setminus E$ is a finite disjoint union of elements of $\mathcal{S}$).

For example, the set of half-open intervals $\mathcal{I}$ is a semialgebra because

$$\mathbb{R} \setminus (a, b] = (-\infty, a] \sqcup (b, \infty),$$

and the two terms on the right-hand side are both in $\mathcal{I}$. (If we're concerned about the open upper bracket on (b, ∞), recall that we actually defined $\mathcal{I}$ to be elements of the form $(a, b] \cap \mathbb{R}$.) And once we have our semialgebra $\mathcal{S}$, our goal

is to extend μ from $\mathcal{S}$ to the set $\mathcal{A}$ of finite disjoint unions of elements in $\mathcal{S}$ (in other words, the algebra generated by $\mathcal{S}$). We can verify the following fact:

Lemma 55

Let $\mathcal{S}$ be a semialgebra over Ω, and let $\mu : \mathcal{S} \to [0, \infty]$ be countably additive over $\mathcal{S}$ (so for any $E_i, E \in \mathcal{S}$ such that $E = \bigsqcup_{i=1}^{\infty} E_i$, we have $\mu(E) = \sum_{i=1}^{\infty} \mu(E_i)$. Then there is a unique $\mu : \mathcal{A} \to [0, \infty]$ which extends μ on $\mathcal{S}$ and is countably additive over $\mathcal{A}$.

(This was essentially proved during lectures 2 and 3, and checking that the proof carries over is a bit of bookkeeping. The details will be on our homework.) We'll now apply this result to **product measures**. The idea is that if we have (for instance) a Lebesgue measure defined on two different axes, then the measure of a rectangle should just be the product of the Lebesgue measures of the two side lengths. To formalize that, for the rest of this lecture, consider two measurable spaces $(S, \mathcal{G})$ and $(T, \mathcal{H})$. The **product space** will just be $\Omega = S \times T$, but it's a bit harder to write down what the measurable subsets of Ω need to look like:

Definition 56

The **product σ-field** $\mathcal{F}$ of $(S, \mathcal{G})$ and $(T, \mathcal{H})$ is the σ-field $\sigma(\mathcal{S})$, where $\mathcal{S}$ is the set of "rectangles"

$$\mathcal{S} = \{A \times B : A \in \mathcal{G}, B \in \mathcal{H}\}.$$

In particular, Lemma 55 will be useful now if we verify that $\mathcal{S}$. We should draw some pictures to convince ourselves of this: the intersection of two rectangles is a rectangle, and we can decompose $\Omega \setminus (A \times B)$ into rectangles in a picture too. But if we want to be more formal (for example, if we were writing this on a test), we would write that

$$(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$$

(and indeed both sets on the right side are in $\mathcal{S}$), and similarly

$$\Omega \setminus (A \times B) = (A \times (T \setminus B)) \sqcup ((S \setminus A) \times T).$$

So the (product) measurable space that we'll be trying to work with here is $(\Omega, \mathcal{F}) = (S \times T, \mathcal{G} \otimes \mathcal{H})$ – all that's left is for us to actually put a measure on it. Suppose that our original measurable spaces $(S, \mathcal{G}, \lambda)$ and $(T, \mathcal{H}, \rho)$ are equipped with measures, and assume here that we have finite measures $\lambda(S), \rho(T) < \infty$ (though σ-finite is sufficient). Our goal is to define a product measure $\mu = \lambda \otimes \rho$ on $(\Omega, \mathcal{F})$ using Lemma 55, which motivates the following verification:

Lemma 57

Given the measures λ and ρ above, define $\mu : \mathcal{S} \to [0, \infty)$ by setting $\mu(A \times B) = \lambda(A) \cdot \rho(B)$ for all $A \in \mathcal{G}$ and $B \in \mathcal{H}$. Then μ is countably additive over $\mathcal{S}$: that is, if we can write a rectangle $A \times B$ as a disjoint countable union of rectangles $A \times B = \bigsqcup_{i=1}^{\infty} (A_i \times B_i)$, then $\lambda(A)\rho(B) = \sum_{i=1}^{\infty} \lambda(A_i)\rho(B_i)$.

Proof. Assume without loss of generality that all sets are nonempty. (Otherwise, discard them.) Consider the **coordinate projection map** π defined by $\pi(x, y) = x$. If we apply π to both sides of $A \times B = \bigsqcup_{i=1}^{\infty} (A_i \times B_i)$, we get

$$A = \pi(A \times B) = \pi\left(\bigsqcup_{i=1}^{\infty} (A_i \times B_i)\right) = \bigcup_{i=1}^{\infty} A_i$$

because all sets are nonempty (though we no longer need to have a disjoint union on the right). Thus, A is the union of the A_is, and similarly B is the union of the B_is. Now by definition, (x, y) belongs to $A \times B$ if and only if there exists some unique index i such that $(x, y) \in A_i \times B_i$. Thus, for any $x \in A$, we can look at the cross-section $x \times B$ along B and break it up into disjoint pieces, meaning that

$$B = \bigsqcup_{i: x \in A_i} B_i.$$

Since we know that ρ is a measure on B, countable additivity then tells us that

$$\rho(B) = \sum_{i: x \in A_i} \rho(B_i)$$

for any $x \in A$. If we then define the function $f(x)$ to be $\rho(B)$ inside A and 0 otherwise (this is where we use our "area" intuition), we have

$$f(x) = 1\{x \in A\} \cdot \rho(B) = 1\{x \in A\} \sum_{i: x \in A_i} \rho(B_i) = \sum_{i=1}^{\infty} 1\{x \in A_i\} \rho(B_i).$$

Since f is a number $\rho(B)$ times an indicator function 1_A, f is a simplee function on S. Similarly, the right-hand side is a pointwise limit of simple functions (taking the partial sums). Thus we can integrate both sides over S (with respect to λ) to get

$$\boxed{\lambda(A)\rho(B)} = \int \left[\sum_{i=1}^{\infty} 1\{x \in A_i\} \rho(B_i) \right] \lambda(dx).$$

Since all functions are nonnegative, the partial sums are monotone, and we can apply the monotone convergence theorem, which tells us that the right-hand side is $\int \left[\sum_{i=1}^{\infty} 1\{x \in A_i\}\rho(B_i)\right] \lambda(dx) = \sum_{i=1}^{\infty} \int \left[1\{x \in A_i\}\rho(B_i)\right] \lambda(dx) = \boxed{\sum_{i=1}^{\infty} \lambda(A_i)\rho(B_i)}$, as desired. (So in other words, "we can interchange the sum and the integral" in this case.) □

With this, we can apply Lemma 55, we've found a unique $\mu = \lambda \otimes \rho$ that extends our definition on $\mathcal{S}$ and is countably additive over $\mathcal{A}$, so then applying the Caratheodory extension theorem gives us a measure μ on the product space $(\Omega, \mathcal{F})$. It is also natural to generalize to a product measure of the form $\mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_n$ – we can check that the order of operations doesn't matter, so such a product measure can also be defined. (And beyond that, it's also very common to work with infinite product spaces and define $\bigotimes_{i=1}^{\infty}(\Omega_i, \mathcal{F}_i, \mu_i)$, but these are more tricky and we'll talk about them next lecture.)

For now, we'll turn to discussing the generalization of the Lebesgue integral to product spaces. Recall that we're not just working with Riemann double integrals – because Lebesgue integral can also deal with summations, we're also in the setting of double sums of the form $\sum_{i,j} a_{ij}$. There is some trickiness here, because while we can always swap the order of summation for finite sums, like

$$\sum_{i=1}^{N} \sum_{j=1}^{M} a_{ij} = \sum_{j=1}^{M} \sum_{i=1}^{N} a_{ij},$$

we can run into issues with infinite sums:

Example 58

Define a doubly-indexed sequence

$$a_{ij} = \begin{cases} 1 & i = j, \\ -1 & i = j+1, \\ 0 & \text{otherwise.} \end{cases}$$

Then the order of summation changes the final answer $\sum_{i,j} a_{ij}$: we have

$$\sum_{i=1}^{\infty}\left(\sum_{j=1}^{\infty} a_{ij}\right) = 1+0+\cdots = 1, \quad \sum_{j=1}^{\infty}\left(\sum_{i=1}^{\infty} a_{ij}\right) = 0+0+\cdots = 0.$$

In cases like in this example, we'll say that the double sum $\sum_{i,j} a_{ij}$ is **undefined** (and we can't freely change the order). But there are conditions where swapping the order of summation (or more generally, integration) is allowed, and those are also the situations where we can define the Lebesgue double integral:

Theorem 59 (Fubini-Tonelli)

Let $(S, \mathcal{G}, \lambda)$ and $(T, \mathcal{H}, \rho)$ be finite measure spaces, and define $(\Omega, \mathcal{F}, \mu) = (S \times T, \mathcal{G} \otimes \mathcal{H}, \lambda \otimes \rho)$. Let f be a measurable function defined on $(\Omega, \mathcal{F})$. If $f \geq 0$ or $\int |f| d\mu < \infty$, then

$$\int_S \left[\int_T f(x,y) d\rho(y)\right] d\lambda(x) = \int_\Omega f\, d\mu = \int_T \left[\int_S f(x,y) d\lambda(x)\right] d\rho(y).$$

(More specifically, Tonelli's theorem covers the $f \geq 0$ case, and Fubini's theorem covers the $\int |f| d\mu < \infty$ case.). In particular, the middle integral is defined by the construction of the Lebesgue integral, but it's not clear a priori that the left and right double integrals actually exist.

Proof. We'll prove the first equality – the other one follows by symmetry – and we'll do so with the "start with indicators" method. First, suppose that $f = 1_E$ for a measurable $E \in \mathcal{F} = \mathcal{G} \otimes \mathcal{H}$. Fix some $x \in S$, and define $f_x(y) = f(x,y)$ to emphasize that f is just a function of y for now. Then

$$(1_E)_x(y) = 1\{(x,y) \in E\} = 1\{y \in E_x\},$$

where E_x is the cross-section of x across E in the space T.

Lemma 60

The cross-section E_x, as defined above, is a measurable set (in T) for any $x \in S$.

Proof of lemma. A slick trick here is to define

$$\mathcal{F}_x = \{E \in \mathcal{F} : E_x \text{ measurable}\} \subseteq \mathcal{F}.$$

Our goal is to show that $\mathcal{F}_x = \mathcal{F}$. We can check that $\mathcal{F}_x$ is a σ-field:

- If $E \in \mathcal{F}_x$, then E_x is measurable in T, and thus (because $\mathcal{H}$ is a σ-field) so is its complement $T \setminus E_x = (\Omega \setminus E)_x$ (the complement of the cross-section of E is the cross-section of the complement of E). Thus $\Omega \setminus E \in \mathcal{F}_x$ as well.

- If $E_i \in \mathcal{F}_x$ for all i, then each $(E_i)_x$ is measurable in T, so the countable union $\bigcup_{i=1}^{\infty}(E_i)_x = \left(\bigcup_{i=1}^{\infty} E_i\right)_x$ by the same logic (here we use that the countable union of the cross-section of E is the cross-section of the countable union of E), so $\bigcup_{i=1}^{\infty} E_i$ is also in $\mathcal{F}_x$.

Additionally, the cross-section of any rectangle $A \times B$ is always B or the empty set (depending on whether x is in A or not), which are both measurable, so $\mathcal{F}_x$ contains $\mathcal{S}$. But this means that $\sigma(\mathcal{F}_x) = \mathcal{F}_x$ contains $\sigma(\mathcal{S})$ (the whole σ-field), so $\mathcal{F}_x = \mathcal{F}$, as desired. □

Thus $f_x(y) = 1\{y \in E_x\}$ is a measurable function of y on $(T, \mathcal{H})$, meaning that the inner integral $\int_T f(x,y)d\rho(y)$ is indeed well-defined and satisfies

$$\int_T f(x,y)d\rho(y) = \int_T 1_E(x,y)d\rho(y) = \int_T 1\{y \in E_x\}d\rho(y) = \rho(E_x).$$

We now need to evaluate the outer integral, and to do so, we first need to show that the map $x \to \rho(E_x)$ is a measurable function on $(S, \mathcal{G})$. (The picture we should have in mind is that we slide x along the S-axis, outputting the measure of its cross-section in E, and we want to verify that this is a measurable function.) To do this, we use a similar trick as before, defining

$$\mathcal{F}_\rho = \left\{E \in \mathcal{F} : x \to \rho(E_x) \text{ is measurable on } (S, \mathcal{G}) \text{ and } \int_S \rho(E_x)d\lambda(x) = \mu(E)\right\} \subseteq \mathcal{F}.$$

(This final condition is included because it is the eventual result we want to show, which is that this double integral is indeed equal to $\int_\Omega 1_E d\mu = \int_\Omega f\,d\mu$.) We want to show that $\mathcal{F}_\rho = \mathcal{F}$, and we'll do so with a pi-lambda argument. Like before, $\mathcal{F}_\rho$ contains all rectangles in $\mathcal{S}$, because for any $E = A \times B$, $\rho(E_x)$ is equal to $\rho(B)$ if $x \in A$ and 0 otherwise, so $\rho(E_x) = \rho(B) \cdot 1\{x \in A\}$, which is indeed a measurable function integrating to $\int \rho(E_x)d\lambda(x) = \rho(B)\lambda(A) = \mu(E)$. Also, $\mathcal{S}$ is a π-system (semialgebras are closed under intersection by definition), and $\mathcal{F}_\rho$ is a λ-system:

- The whole space $\Omega = S \times T$ is an element of $\mathcal{F}_\rho$, because it is a rectangle.
- For any two sets $E_1, E_2 \in \mathcal{F}_\rho$ with $E_1 \supseteq E_2$, notice that

  $$\rho((E_1 \setminus E_2)_x) = \rho((E_1)_x \setminus (E_2)_x) = \rho((E_1)_x) - \rho((E_2)_x),$$

  because the cross-section of the difference of E_1 and E_2 is the difference of the cross-sections and because $(E_1)_x \supseteq (E_2)_x$. Since both terms are measurable functions of x and the difference of measurable functions is measurable, $\rho((E_1 \setminus E_2)_x)$ is indeed a measurable function. Additionally, the integral of $\rho((E_1 \setminus E_2)_x)$ is $\rho(E_1) - \rho(E_2) = \rho(E_1 \setminus E_2)$ by linearity of the integral. Thus $E_1 \setminus E_2 \in \mathcal{F}_\rho$ as well.
- Finally, we want to show that if $E_i \in \mathcal{F}_\rho$ for all i and $E_i \uparrow E$, then the limit E is also in $\mathcal{F}_\rho$. Since $(E_i)_x \uparrow E_x$, $\rho((E_i)_x) \uparrow \rho(E_x)$ for all x by continuity from below. So E_x is measurable (as the countable union of the E_is), and $\int \rho(E_x)d\lambda(x)$ is the limit of the $\int \rho((E_i)_x)d\lambda(x)$s by monotone convergence theorem, which is indeed $\lim_{i\to\infty} \rho(E_i) = \rho(E)$. Thus $E \in \mathcal{F}_\rho$.

So now by the π-λ theorem, because $\mathcal{F}_\rho$ is a λ-system containing $\mathcal{S}$, it also contains $\sigma(\mathcal{S})$. Thus $\mathcal{F}_\rho = \mathcal{F}$, meaning that for any $E \in \mathcal{F} = \mathcal{G} \otimes \mathcal{H}$, the indicator function 1_E satisfies Fubini-Tonelli. From here, we perform the usual machinery: the result is also true for simple functions by linearity, so it's also true for nonnegative functions by approximating from below with the monotone convergence theorem, and thus it's also true for all integrable functions by splitting into the positive and negative parts. □

7 September 25, 2019

Last time, we constructed the product measure on the product of two σ-finite measure spaces. Specifically, given $(S, \mathcal{G}, \lambda)$ and $T, \mathcal{H}, \rho)$, we explained how to define $\Omega = S \times T$, $\mathcal{F} = \mathcal{G} \otimes \mathcal{H}$, and $\mu = \lambda \otimes \rho$. Note that we can also define non-product measures, such as the uniform measure on a triangle in our space: some examples of these will be on our homework for next time. (And as mentioned, we can always extend this to arbitrary finite product spaces $\left(\prod_{i=1}^n \Omega_i, \bigotimes_{i=1}^n \mathcal{F}_i, \bigotimes_{i=1}^n \mu_i\right)$ by induction.)

We also mentioned during our discussion that infinite products are more subtle, and that's what we'll be going into today. For example, consider the **percolation model** on the infinite integer lattice grid, where at every point $(x, y) \in \mathbb{Z}^2$, we flip a coin to decide whether we have a 0 or a 1. The percolation problem then asks us about the large-scale structure of this model. And we can also build more complicated models off of this (for example in statistical physics), where perhaps we have some nearest-neighbor interaction between the sites. Our first instinct might be to avoid infinite products completely and only consider a finite large subset of the lattice, but often the boundary conditions of such a system are annoying to deal with. So here are the two main results we're going to prove today:

- a special case of the Ionescu-Tulcea theorem, which says that we can define an infinite **product** of probability spaces with no further conditions, and
- the Kolmogorov extension theorem, which says that we can define a **non-product** measure on an infinite product space, with a mild regularity condition.

We're using a product sigma-algebra in both cases, so we'll start by defining what that means for an infinite product:

Definition 61

Let $(\Omega_\alpha, \mathcal{F}_\alpha)$ be measurable spaces, indexed by $\alpha \in I$ (so there can be an uncountable set of αs). The **product σ-field** $\mathcal{F} = \bigotimes_{\alpha \in I} \mathcal{F}_\alpha$ is the minimal σ-field over $\Omega = \prod_{\alpha \in I} \Omega_\alpha$ such that all single-coordinate projections π_α are measurable. In other words, we require that $(\pi_\alpha)^{-1}(E_\alpha) \in \mathcal{F}$ for all $E_\alpha \in \mathcal{F}_\alpha$.

(This definition may look familiar if we've heard of a **product topology**.) The preimage of E_α can also be written

$$(\pi_\alpha)^{-1}(E_\alpha) = E_\alpha \times \prod_{\gamma \in I \setminus \{\alpha\}} \Omega_\gamma,$$

and because I can be uncountable, this is **not the same** as just taking arbitrary Cartesian products of measurable sets E_α in each $\mathcal{F}_\alpha$. Specifically, we are only allowed to take countable unions of these sets, so at most countably many of the sets in our product are allowed to be not the whole set Ω_α.

Definition 62 (Notation)

For any subset $J \subseteq I$ of the index set, we will define the partial products

$$\Omega_J = \prod_{\alpha \in J} \Omega_\alpha, \quad \mathcal{F}_J = \bigotimes_{\alpha \in J} \mathcal{F}_\alpha.$$

To understand why this is a useful definition, we'll bring in the probability intuition: we often care about the finite-dimensional distributions of a probability measure, also called its **marginals**. Specifically, suppose we have a product measure on an infinite product probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where the index set I is countable. Then any set

$E = \prod_{\alpha \in I} E_\alpha$ is measurable, so we would guess that

$$\mathbb{P}(E) \stackrel{?}{=} \prod_{\alpha \in I} \mathbb{P}_\alpha(E_\alpha).$$

(This turns out to not quite be true, but we're just talking about intuition for now.) In order for this probability to be positive (nonzero), most of the probabilities $\mathbb{P}_\alpha(E_\alpha)$ need to be pretty close to 1, meaning that most of the dimensions basically need to include the whole space. That motivates only looking at distributions along finitely many dimensions:

Definition 63

Let J be a finite subset of our index set I, and let $\mathbb{P}$ be a measure on the infinite product space Ω. The **finite-dimensional marginal** (or **finite-dimensional distribution**) of $\mathbb{P}$ on J is the pushforward measure on $(\Omega_J, \mathcal{F}_J)$, where for any $E_J \in \mathcal{F}_J$,

$$[(\pi_J)_\#(\mathbb{P})](E_J) = \mathbb{P}(\pi_J^{-1}(E_J)) = \mathbb{P}(E_J \times \Omega_{I \setminus J}).$$

If we want these marginals to make sense, then we should get consistent distributions no matter how we "project down." In other words, if we have two finite subsets $J' \subseteq J \subseteq I$, then the marginal $\mathbb{P}_{J'}$ of $\mathbb{P}$ on J' should be the same whether we're projecting directly from I to J', or from I to J and then J'. (And the point of the measurability condition in the definition of the product σ-field is necessary to make sure everything in this consistency statement is well-defined.)

So when we construct a measure $\mathbb{P}$ on an infinite product space, there are already some consistency conditions that $\mathbb{P}$ must satisfy – a natural question is to ask whether we can just use these conditions to construct $\mathbb{P}$. And in the results we'll show today, we'll start with the family of $\mathbb{P}_J$s (also called a **family of finite-dimensional distributions**), and we'll get a measure $\mathbb{P}$ consistent with those marginals.

Before we get to the main results, we'll go over a useful alternate version of the Carathéodory extension theorem to apply in the infinite-dimensional case (which doesn't require us to check countable additivity directly):

Lemma 64 (Variant of Carathéodory)

Let $\mathcal{A}$ be an algebra over Ω, and let $\mu : \mathcal{A} \to [0, \infty)$ be a finite measure which is finitely additive. Suppose that for any sequence $B_n \in \mathcal{A}$ with $B_n \downarrow \varnothing$, we have $\mu(B_n) \downarrow 0$. Then μ is also countably additive over $\mathcal{A}$ (so the Carathéodory extension theorem applies).

Proof. We need to prove countable additivity – let $A_n \in \mathcal{A}$ be disjoint sets such that $A = \bigsqcup_{n=1}^\infty A_n$ is also in $\mathcal{A}$. For each positive integer n, the set $B_n = A \setminus \bigsqcup_{i=1}^n A_i$ is also in $\mathcal{A}$ (because it can be constructed with just finite unions and complementations of A_is and A). By finite additivity, we thus have

$$\mu(A) = \sum_{i=1}^n \mu(A_i) + \mu(B_n).$$

But B_n decreases to the empty set, meaning that $\mu(B_n) \to 0$ by assumption. Thus taking $n \to \infty$ shows countable additivity, as desired. □

We now want to specialize this lemma to our infinite product space $(\Omega, \mathcal{F}) = (\prod \Omega_\alpha, \otimes \mathcal{F}_\alpha)$, and we'll set up the notation now. As an exercise, we can check that

$$\mathcal{A} = \left\{ (\pi_J)^{-1}(E_J) = E_J \times \Omega_{I \setminus J} : J \text{ finite set}, E_J \in \mathcal{F}_J \right\}.$$

is an algebra (we just need to check complementation and finite union, both of which are pretty easy). If we now consider an arbitrary sequence $B_n \in \mathcal{A}$ with $B_n \downarrow \varnothing$, each B_i only involves finitely many coordinates, so the total number of coordinates that can be involved in any of the B_is is countable. So we'll renumber the relevant coordinates to $Q = \{1, 2, 3, \cdots\}$ for convenience, and we may assume without loss of generality that B_1 depends on the first coordinate, B_2 depends on the first two, and so on (we can check that both of the following proofs work whether each coordinate in Q represents a single Ω_α or a product of them). This means that our sequence of B_ns can be written more explicitly as

$$\boxed{B_n = \overline{B_n} \times \Omega_{Q \setminus [n]} \times \Omega_{\text{rest}}},$$

where $[n] = \{1, 2, \cdots, n\}$, $\overline{B_n} \in \mathcal{F}_{[n]}$, and Ω_{rest} is the full probability space Ω on the index set $I \setminus Q$. (Here, the subset Q can depend on the B_n but is always countable.) Then to apply Lemma 64 in our theorems, we just need to show that $\mu(B_n) \downarrow 0$.

Theorem 65 (Ionescu–Tulcea theorem, special case)
Let $(\Omega_\alpha, \mathcal{F}_\alpha, \mathbb{P}_\alpha)$ be probability spaces indexed by $\alpha \in I$. There is a unique probability measure $\mathbb{P}$ on the product space $\left(\prod_{\alpha \in I} \Omega_\alpha, \bigotimes_{\alpha \in I} \mathcal{F}_\alpha\right)$, where the family of finite-dimensional distributions is given by $\mathbb{P}_J = \bigotimes_{\alpha \in J} \mathbb{P}_\alpha$ for any finite J.

This is essentially us creating a "product measure" on our infinite-dimensional space, since we are requiring product measures on all finite-dimensional Ω_Js.

Proof. As discussed above, we will be applying Lemma 64 using the algebra $\mathcal{A} = \left\{E_J \times \Omega_{I \setminus J}, J \text{ finite}, E_J \in \mathcal{F}_J\right\}$. To have the correct finite-dimensional marginals, we should define

$$\mathbb{P}(E_J \times \Omega_{I \setminus J}) = \left(\bigotimes_{\alpha \in J} \mathbb{P}_\alpha\right)(E_J)$$

for each $E_J \in \mathcal{F}_J$ (note that E_J may not be the product of E_αs, so we need to write the right-hand side in terms of the product measure $\mathbb{P}_J = \left(\bigotimes_{\alpha \in J} \mathbb{P}_\alpha\right)$). It's a bookkeeping exercise to show that $\mathbb{P}$ is well-defined and finitely additive over $\mathcal{A}$, so we'll be done by Lemma 64 if we can show that $\mathbb{P}(B_n) \downarrow 0$ for any B_n of the boxed form above. By definition, because B_n is of the form $E_J \times \Omega_{I \setminus J}$ for $J = [n]$ and $E_j = \overline{B_n}$,

$$\mathbb{P}(B_n) = \mathbb{P}_{[n]}\left(\overline{B_n}\right) = \mathbb{P}_{[n+1]}\left(\overline{B_n} \times \Omega_{n+1}\right) \geq \mathbb{P}_{[n+1]}\left(\overline{B_{n+1}}\right) = \mathbb{P}(B_{n+1}).$$

(where the second equality comes from consistency of the marginals, and the inequality comes from the B_ns being decreasing and thus decreasing within $E_{[n+1]}$). Thus, the $\mathbb{P}(B_n)$ form a nonincreasing sequence of numbers, and we just need to show that its limit is zero. By Fubini's theorem, we can write out our integral as

$$\mathbb{P}(B_n) = \mathbb{P}_{[n]}(\overline{B_n}) = \int \cdots \left[\int 1\{\overline{B_n}\} d\mathbb{P}_n\right] \cdots d\mathbb{P}_1.$$

To visualize the next step, consider the following doubly infinite array:

$$\begin{array}{cccc} \mathbb{P}(B_1) & 1\{\overline{B_1}\} & 1\{\overline{B_1}\} \times \Omega_2 & 1\{\overline{B_1}\} \times \Omega_2 \times \Omega_3 & \cdots \\ \mathbb{P}(B_2) & \int 1\{\overline{B_2}\} d\mathbb{P}_2 & 1\{\overline{B_2}\} & 1\{\overline{B_2}\} \times \Omega_3 & \cdots \\ \mathbb{P}(B_3) & \iint 1\{\overline{B_3}\} d\mathbb{P}_1 d\mathbb{P}_2 & \int 1\{\overline{B_3}\} d\mathbb{P}_2 & 1\{\overline{B_3}\} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

In other words, if a_{ij} is the entry in the ith row and jth column, (and we start the array from the first row and zeroth column), then $a_{ii} = 1\{\overline{B_i}\}$ for any i. Additionally, we multiply by Ω_{n+1} when we move from column n to column $n+1$, and we integrate with respect to $\mathbb{P}_{n+1}$ when we move from column $n+1$ to column n.

We then notice that the kth column consists of functions on the first k coordinates (the leftmost column is a sequence of numbers, the next column is a sequence of measurable functions on $(\Omega_1, \mathcal{F}_1)$, the following column is a sequence of measurable functions on $(\Omega_1 \otimes \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$, and so on). Additionally, each column is nonincreasing, so they all have (pointwise) limits. Let g_n be the pointwise limit of the nth column, and suppose for the sake of contradiction that $g_0 > 0$. By the bounded convergence theorem (which we can use because probability spaces are finite and our functions are bounded from above by 1), we can interchange the limit and the integration with respect to $\mathbb{P}_{n+1}$, meaning that

$$g_n = \int g_{n+1} d\mathbb{P}_{n+1} \text{ for all } n.$$

Now because g_0 is positive, there is some $\omega_1 \in \Omega_1$ such that $g_1(\omega_1)$ is positive. But then $g_1(\omega_1) = \int g_2(\omega_1, \omega_2) d\mathbb{P}_2(\omega_2)$, meaning there is some $\omega_2 \in \Omega_2$ such that $g_2(\omega_1, \omega_2)$ is positive. Repeating this, we have an infinite sequence $(\omega_1, \omega_2, \cdots)$ such that $g_n(\omega_1, \cdots, \omega_n) > 0$ for all n. Remembering that the columns are nonincreasing, we can only have $g_n(\omega_1, \cdots, \omega_n) > 0$ if $1\{\overline{B_n}\}(\omega_1, \cdots, \omega_n) > 0$, meaning that $(\omega_1, \cdots, \omega_n) \in \overline{B_n}$, so $(\omega_1, \omega_2, \cdots) \in \overline{B_n} \times \Omega_{Q\setminus[n]}$. Putting this together across all n, we find that

$$(\omega_1, \omega_2, \cdots) \in \bigcap_n \overline{B_n} \times \Omega_{Q\setminus[n]}.$$

But this is impossible: if the sets $B_n = \overline{B_n} \times \Omega_{Q\setminus[n]} \times \Omega_{\text{rest}}$ decrease to zero, then the intersection on the right-hand side must be empty. Thus we have a contradiction, meaning that we really have $g_0 = 0$, so we can apply Lemma 64 to finish the construction of the measure. □

Theorem 66 (Kolmogorov extension theorem)
Let Ω_α be metric spaces with Borel σ-fields $\mathcal{F}_\alpha$. Suppose we have a (not necessarily product) consistent family of finite-dimensional distributions $\{\mathbb{P}_J\}$, such that $\mathbb{P}_J$ is **inner-regular** for all J (meaning that the measure of any set can be approximated from within by compact subsets). Then there is a probability measure $\mathbb{P}$ on the whole space $(\Omega, \mathcal{F})$ with these finite-dimensional distributions $\{\mathbb{P}_J\}$.

Proof. Use the same $\mathcal{A}$ as in the previous proof, and again define $\mathbb{P} : \mathcal{A} \to [0,1]$ by setting $\mathbb{P}(E_J \times \Omega_{I\setminus J}) = \mathbb{P}_J(E_J)$. (Again, it is a bookkeeping exercise for us to show that $\mathbb{P}$ is well-defined and finitely additive over $\mathcal{A}$.) Then we again need to check the condition of Lemma 64: given sets of the form $B_n = \overline{B_n} \times \Omega_{Q\setminus[n]} \times \Omega_{\text{rest}}$ decreasing to the emptyset, we wish to prove that $\mathbb{P}(B_n) \downarrow 0$.

Suppose for the sake of contradiction that $\mathbb{P}(B_n) \downarrow \varepsilon > 0$. The key step is to replace each $\overline{B_n}$ with a compact set $\overline{K_n}$ sitting inside it such that the same hypotheses still hold. By assumption, $\mathbb{P}_{[n]}$ is in our family of finite dimensional distributions, so it is inner-regular, meaning we can find a compact $\overline{C_n} \subseteq \overline{B_n}$ such that

$$\mathbb{P}_{[n]}(\overline{C_n}) \geq \mathbb{P}_{[n]}(\overline{B_n}) - \frac{\varepsilon}{2^{n+1}}.$$

The sets $C_n = \overline{C_n} \times \Omega_{I\setminus[n]}$ are now no longer necessarily decreasing, but we can define

$$\overline{K_n} = \bigcap_{\ell=1}^{n} \overline{C_\ell} \times \Omega_{[n]\setminus[\ell]}$$

for each n – these are closed subsets of compact sets (because the nth term of the intersection is the compact set

$\overline{C_n}$) and are thus compact, and $\overline{K_{n+1}}$ is contained within $\overline{K_n} \times \Omega_{n+1}$. So the sets

$$K_n = \overline{K_n} \times \Omega_{I \setminus [n]}$$

are decreasing and contained in the B_ns, so they decrease to the empty set. (At this point, it's important to note that the K_ns are not necessarily compact – if they were, we'd already have a contradiction, because a nested sequence of compact sets has a nonempty intersection.) But now we can check (left as an exercise to us) that using the compact subsets does not lose us much measure, and we have

$$\mathbb{P}(K_n) \geq \mathbb{P}(B_n) - \frac{\varepsilon}{2}.$$

Thus, $\mathbb{P}(K_n) \downarrow \delta \geq \frac{\varepsilon}{2}$ decreases to some positive value. But K_n is contained in B_n, and the latter decrease to the empty set, so $K_n \downarrow \varnothing$ as well. Now K_n is nonempty for all n (because it has a positive measure), so we can find some $\omega^{(n)} \in K_n$. Let $\omega^{0,n} = \omega^{(n)}$, and for every $\ell \geq 1$, consider the sequence

$$\left(\pi_{[\ell]}\left(\omega^{\ell-1,n}\right)\right)_{n \geq 1},$$

meaning that we only consider the first ℓ components of each ω in the sequence. Since this is a sequence contained in the compact set $\overline{K_\ell}$, it has a convergent subsequence – let those elements form the next sequence $\omega^{\ell,1}, \omega^{\ell,2}, \cdots$. (In other words, the $\{\omega^{\ell,n}\}_{n \geq 1}$ sequence converges in the first ℓ coordinates.)

But we can now take the diagonal entries $\omega^{n,n}$ and form a new sequence out of them – because we defined our sequence to converge in the first ℓ coordinates for any ℓ, we know that $\{\pi_\ell(\omega^{n,n})\}_{n \geq 1}$ converges to some $x_\ell \in \Omega_\ell$ for all ℓ. But this means that there is some point $(x_1, x_2, \cdots)$ which is in all of the K_ns:

$$(x_1, x_2, \cdots) \in \bigcap_{n=1}^{\infty} \overline{K_n} \times \Omega_{Q \setminus [n]}.$$

And much like in our previous proof, this is a contradiction: the sets $K_n = \overline{K_n} \times \Omega_{Q \setminus [n]} \times \Omega_{\text{rest}}$ decrease to zero, so the intersection on the right-hand side must be empty. Thus $\mathbb{P}(B_n)$ must actually decrease to 0, so we can again apply Lemma 64 and get a valid measure $\mathbb{P}$, as desired. □

8 September 30, 2019

Last week, we did a lot of work to construct product measure spaces, especially in the case where we have an uncountable product of the form $(\Omega, \mathcal{F}, \mathbb{P}) = \left(\prod_{\alpha \in I} \Omega_\alpha, \bigotimes_{\alpha \in I} \mathcal{F}_\alpha, \bigotimes_{\alpha \in I} \mathbb{P}_\alpha\right)$. Today, we'll mostly talk about "easier things" motivated by undergraduated probability – we have likely seen concepts like variance, correlation, and independence before, and we'll basically review them now in a more formalized setting. Note that we're now considering arbitrary probability spaces (and are no longer assuming any kind of product structure).

Definition 67

Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ (in other words, let $X : \Omega \to \mathbb{R}$ be a measurable function). The **σ-field generated by X**, denoted $\sigma(X)$, is the minimal σ-field over Ω such that X is measurable. In other words,

$$\sigma(X) = \left\{X^{-1}(B) : B \in \mathcal{B}_{\mathbb{R}}\right\}.$$

We can check that this is a σ-field (for example, the preimage of the complement is the complement of the preimage). Intuitively, $\sigma(X)$ tells us about the events in Ω that we can describe just based on how X maps the space

into $\mathbb{R}$. For example, if X just sends everything to 0, we can't distinguish very much – the preimage of any Borel set B is Ω if it contains 0 and $\varnothing$ otherwise, so $\sigma(X) = \{\varnothing, \Omega\}$ and our random variable can't tell us very much about Ω. In other words, "the larger $\sigma(X)$ is, the more helpful X is for giving us information."

Definition 68

Two events $A, B \in \mathcal{F}$ are **independent** (denoted $A \perp\!\!\!\perp B$) if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

In particular, we also have that

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B^c),$$

so $A \perp\!\!\!\perp B \iff A \perp\!\!\!\perp B^c$, and similar with B^c instead of B. We can also generalize independence to more events:

Definition 69

The events $\{A_\alpha : \alpha \in I\}$ are **mutually independent** if for all finite subsets $J \subseteq I$,

$$\mathbb{P}\left(\bigcap_{\alpha \in J} A_\alpha\right) = \prod_{\alpha \in J} \mathbb{P}(A_\alpha).$$

Extending this definition, if $\{\mathcal{C}_\alpha : \alpha \in I\}$ are each a collection of events, then the collections are **independent** if the above equality holds for all finite $J \subseteq I$ and for each $A_\alpha \in \mathcal{C}_\alpha$ (for all corresponding α).

Definition 70

A set of random variables $\{X_\alpha : \alpha \in I\}$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are **mutually independent random variables** if and only if their σ-algebras $\{\sigma(X_\alpha) : \alpha \in I\}$ are independent.

In particular, for any finite set $J \subseteq I$ and any measurable subsets $B_\alpha \in \mathcal{B}_\mathbb{R}$, we have

$$\boxed{\mathbb{P}(X_\alpha \in B_\alpha \text{ for all } \alpha \in J)} = \mathbb{P}\left(\bigcap_{\alpha \in J} X_\alpha^{-1}(B_\alpha)\right) = \prod_{\alpha \in J} \mathbb{P}\left(X_\alpha^{-1}(B_\alpha)\right) = \boxed{\prod_{\alpha \in J} \mathbb{P}(X_\alpha \in B_\alpha)}.$$

by mutual independence (because the sets $X_\alpha^{-1}(B_\alpha)$s are each in the respective σ-algebras $\sigma(X_\alpha)$). And this probabilistic statement is what we may have seen in an undergraduate probability class as the definition of independence!

Example 71

By definition, a collection of events $\{E_\alpha\}$ are independent if and only if the corresponding indicator variables $1_{E_\alpha} : \alpha \in I$ are independent random variables.

Example 72

Consider the special case where we have a product probability space $(\Omega, \mathcal{F}, \mathbb{P}) = \left(\prod_{\alpha \in I} \Omega_\alpha, \bigotimes_{\alpha \in I} \mathcal{F}_\alpha, \bigotimes_{\alpha \in I} \mathbb{P}_\alpha\right)$ (as we've previously constructed). Then we can define random variables $X_\alpha(\omega) = f_\alpha(\omega_\alpha)$ for each α, which are measurable functions that only depend on the α-coordinate.

The preimage of any $B \in \mathcal{B}_\mathbb{R}$ is then

$$X_\alpha^{-1}(B) = f_\alpha^{-1}(B) \times \prod_{\gamma \in I \setminus \{\alpha\}} \Omega_\gamma$$

(because only the α-coordinate matters for X_α), so the σ-field for the random variable X_α is

$$\sigma(X_\alpha) = \sigma(f_\alpha) \otimes \bigotimes_{\gamma \in I \setminus \{\alpha\}} \mathcal{T}_\gamma,$$

where $\mathcal{T}_\gamma$ is the trivial σ-field $(\varnothing, \Omega_\gamma)$ (since we don't have any "information" about the other coordinates from X_α, and it's okay to include $\varnothing$ in any coordinate γ because that just gives us the emptyset overall, which is the preimage of the emptyset). We can then easily see (by applying the definition and writing out the preimages) that these σ-fields $\sigma(X_\alpha)$s are independent, meaning that all of the $(X_\alpha : \alpha \in I)$ will be independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.

The example above may seem like a special case for independence, but it turns out to be basically the only one! In general, suppose we have some probability space, and suppose that $\{X_\alpha : \alpha \in I\}$ are independent random variables on that space. Then define a new random variable X which is basically a tuple of the X_αs, setting

$$X(\omega) = (X_\alpha(\omega))_{\alpha \in I} \in \prod_{\alpha \in I} \mathbb{R}.$$

Then X is a map $(\Omega, \mathcal{F}, \mathbb{P}) \to \left(\prod_{\alpha \in I} \mathbb{R}, \bigotimes_{\alpha \in I} \mathcal{B}_\mathbb{R}, \mathcal{L}_X\right)$, where the law of X is $\mathcal{L}_X = X_\# \mathbb{P} = \mathbb{P} \circ X^{-1}$. But then we can check that the X_αs are independent random variables if and only if the law $\mathcal{L}_X$ is a product measure (by our discussion of product measure constructions from last lecture) – in other words, we should treat "independence of random variables" and "law is a product measure" as one and the same.

We've now reached (basically) **the end of measure theory** in this class – we now have enough theory developed to talk about more interesting aspects of probability.

Definition 73

Let X be a random variable in $(\Omega, \mathcal{F}, \mathbb{P})$. The ***p*th moment** of X is the value of $\mathbb{E}(|X|^p)$ (which may be infinite), and the ***L*^{p} norm** of X is $||X||_p = ||X||_{L^p(\Omega,\mathcal{F},\mathbb{P})} = \mathbb{E}(|X|^p)^{1/p}$. We say that X **belongs to *L*^{p}** if $||X||_p$ is finite.

The following is a useful fact about integrals that we won't prove here:

Theorem 74 (Jensen's inequality)

Let $G \subseteq \mathbb{R}$ be an open interval, and let $g : G \to \mathbb{R}$ be a convex function, meaning that $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$ for all $\lambda \in [0,1]$ and $x, y \in G$. Then if X is a random variable with $\mathbb{P}(X \in G) = 1$ and $\mathbb{E}|X|$ and $\mathbb{E}|g(X)|$ both finite, then $\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$.

In particular, this tells us a useful inequality for the L^p norms of a random variable:

Corollary 75 (L^p monotonicity)

Let $0 < r \leq p < \infty$, and let X be a random variable such that $||X||_r, ||X||_p$ are both finite. Then Jensen's inequality applied to the convex function $g(y) = |y|^{p/r}$ yields $\mathbb{E}(|X|^p) = \mathbb{E}((|X|^r)^{p/r}) \geq (\mathbb{E}|X|^r)^{p/r}$, so $||X||_r \leq ||X||_p$ when $r \leq p$.

Since an infinite L^p norm is "larger" than any finite L^p norm, we should expect that it is not possible for $||X||_r$ to be infinite but $||X||_p$ to be finite. This is indeed the case, but we have to deal with infinite expectations with a bit more care:

Proposition 76

Let $0 < r \leq p < \infty$. If X is a random variable with $||X||_p < \infty$, then $||X||_r < \infty$; in other words, $L^r(\Omega, \mathcal{F}, \mathbb{P}) \supseteq L^p(\Omega, \mathcal{F}, \mathbb{P})$ when $r \leq p$.

Proof. We split up the expectation into two terms

$$\mathbb{E}(|X|^r) = \mathbb{E}(|X|^r 1\{|X| \leq 1\}) + \mathbb{E}(|X|^r 1\{|X| > 1\}),$$

which we'll rewrite (from here on, this will be **standard notation** in this class) as

$$= \mathbb{E}(|X|^r; |X| \leq 1) + \mathbb{E}(|X|^r; |X| > 1).$$

The first term is bounded by $\mathbb{E}(1) = 1$ (because $|X|^r \leq 1$ whenever $|X| \leq 1$), so finiteness of the L^r norm only depends on the second term. But $\mathbb{E}(|X|^r; |X| > 1) \leq \mathbb{E}(|X|^p; |X| > 1)$ (because $|X|^p \geq |X|^r$ whenever $p \geq r$ and $|X| \geq 1$). Thus, if $||X||_p$ is finite, then $\mathbb{E}(|X|^p; |X| > 1)$ is finite, so $\mathbb{E}(|X|^r; |X| > 1)$ is finite, so $\mathbb{E}(|X|^r)$ is finite and thus $||X||_r$ is finite. □

However, it's important to note that this L^p monotonicity **only holds for probability spaces** – we'll verify the following claims on our homework:

- Consider the **sequence spaces** ℓ^p, which contain elements of the form $x = (x_1, x_2, \cdots,)$. The ℓ^p norm of such a sequence is defined as
$$||x||_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{1/p}.$$
In these cases, if $r < p$, then $||x||_r \geq ||x|_p$ (this inequality goes in the opposite direction as Corollary 75, and its proof does not come from Jensen's inequality), meaning that we actually have $\ell^r \subseteq \ell^p$ when $r \leq p$. One way to remember this is that the unit ball in ℓ^r is nested in the unit ball for ℓ^p – for example, the unit ball in ℓ^1 is the set of points such that $|x_1| + |x_2| + \cdots \leq 1$ (which is a "diamond"), while the unit ball in ℓ^2 is an actual "ball." And as p increases, this pattern continues – we can check that the ℓ^∞ unit ball is a cube of side length 2.
- On the other hand, consider the **classical function spaces** $L^p(\mathbb{R}, \mathcal{B}, \lambda)$ under the Lebesgue measure, where the norm of a function is
$$||f||_p = \left(\int |f(x)|^p dx \right)^{1/p}.$$
In this case, it turns out that the L^p and L^r function spaces are not nested in either direction.

If we now turn back to our random variables X on $(\Omega, \mathcal{F}, \mathbb{P})$, Proposition 76 tells us that

$$\mathbb{E}(|X|^2) < \infty \implies \mathbb{E}|X| < \infty,$$

and when these moments are finite, we have $\mathbb{E}|X| \leq \mathbb{E}(X^2)^{1/2}$ by Jensen's inequality. Thus, we can define the following nonnegative quantity (which we should be familiar with):

Definition 77

Let X be a random variable with finite second moment. The **variance** of X is

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}[X])^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Theorem 78 (Cauchy-Schwarz)
For any two random variables $X, Y \in L^2$, we have $\mathbb{E}(|XY|) \leq ||X||_2 ||Y||_2$.

Proof. If $||X||_2 = 0$ or $||Y||_2 = 0$, then either $X = 0$ or $Y = 0$ (almost everywhere), so $\mathbb{E}(|XY|) = 0$ and the equality holds. Otherwise, both $||X||_2$ and $||Y||_2$ are positive. First, consider the special case where $||X|_2 = ||Y|_2 = 1$ and we assume that $\mathbb{E}(|XY|) < \infty$. We know that

$$\mathbb{E}(|X|^2) + \mathbb{E}(|Y|^2) - 2\mathbb{E}(|XY|) = \mathbb{E}(|X| - |Y|)^2 \geq 0,$$

so $\mathbb{E}(|XY|) \leq 1$ and the equality holds. In the more general case whenever $\mathbb{E}(|XY|) < \infty$, we can scale X and Y to have norm 1 (by dividing by $||X||_2$ and

$$||Y|_2$$

respectively), so that the special case implies

$$\mathbb{E}\left|\frac{X}{||X||_2}\frac{Y}{||Y||_2}\right| \leq 1 \implies \mathbb{E}(|XY|) \leq ||X||_2 ||Y||_2.$$

Finally, if we do not know that $\mathbb{E}(|XY|)$ is finite to start with, we can use the previous case with capped random variables (so that we have finite expectation):

$$\boxed{\mathbb{E}\left(|\min\{|X|, n\}, \min\{|Y|, n\}|\right)} \leq ||\min\{|X|, n\}||_2 ||\min\{|Y|, n\}||_2 \leq \boxed{||X||_2 ||Y||_2}.$$

We can now send $n \to \infty$ and use the monotone convergence theorem on the left-hand side to get the desired result. □

This allows us to also define another familiar quantity that we use to study random variables:

Definition 79
Let $X, Y \in L^2$ be two random variables. The **covariance** of X and Y is given by

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}[X])(Y - \mathbb{E}[Y])) = \mathbb{E}(XY) - \mathbb{E}[X]\mathbb{E}[Y],$$

and X, Y are **uncorrelated** if $\text{Cov}(X, Y) = 0$.

In particular, any two random variables $X, Y \in L^2$ that are independent ($X \perp\!\!\!\perp Y$) are also uncorrelated, but the converse is false (random variables can be uncorrelated but not independent).

Now that we have all of our definitions set up, we're going to start to work towards **proving the laws of large numbers**. (As a joke, a "standard" example of such a law is that if someone goes to prison and they flip lots of coins in their free time, they should expect close to half of them to be heads.) We'll see that many of these laws will turn out to be applications of the Pythagorean theorem (like the proof of Cauchy-Schwarz secretly was).

Today, we'll start with the simplest case. Suppose $X_1, \cdots, X_n$ are pairwise uncorrelated random variables, and consider their **sample mean** or **empirical mean** $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Because covariance is linear in each argument (because the Lebesgue integral is), we have

$$\text{Var}\,\overline{X_n} = \frac{1}{n^2}\,\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right).$$

Because the X_is are uncorrelated, all cross-terms disappear and this simplifies to

$$\operatorname{Var} \overline{X_n} = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

Remark 80. *Another way to think of this (in the "Pythagorean theorem" sense) is that we can think of $v_i = X_i - \mathbb{E}(X_i)$ as vectors and note that all scalar products (v_i, v_j) for $i \neq j$ are zero by assumption. Thus, the Pythagorean theorem tells us that the variance is*

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left|\left|\frac{1}{n}\sum_{i=1}^{n} v_i\right|\right|_2^2 = \frac{1}{n^2}\sum_{i=1}^{n} ||v_i||^2$$

by the Pythagorean theorem.

If we now assume that **all X_is have the same law** (that is, they are all identically distributed) as some particular random variable $X \in L^2$, then

$$\boxed{\operatorname{Var}\left(\overline{X_n}\right) = \frac{1}{n^2} \cdot n \operatorname{Var}(X) = \frac{\operatorname{Var}(X)}{n}},$$

which gets smaller and smaller as n gets larger, meaning that we expect the average to converge to the mean $\mathbb{E}[X]$. To prove that, we need to make use of some inequalities we may have seen before (which basically bound the probability of having deviations far from the mean):

Theorem 81 (Markov's inequality)
Let X be a nonnegative random variable. Then for any $t > 0$, we have the inequality

$$\boxed{\mathbb{P}(X \geq t)} = \mathbb{E}\left[1\{X \geq t\}\right] \leq \mathbb{E}\left[\frac{X}{t}\right] = \boxed{\frac{\mathbb{E}[X]}{t}}.$$

Corollary 82 (Chebyshev's inequality)
Let $X \in L^2$ be a (not necessarily nonnegative) random variable. Then for any $t > 0$ (by Markov's inequality),

$$\boxed{\mathbb{P}(|X - \mathbb{E}[X]| \geq t)} = \mathbb{P}((X - \mathbb{E}[X])^2 \geq t^2) \leq \mathbb{E}\left[\frac{(X - \mathbb{E}[X])^2}{t^2}\right] = \boxed{\frac{\operatorname{Var}(X)}{t^2}}.$$

If we now apply Chebyshev's inequality to our expression for $\operatorname{Var}\left(\overline{X_n}\right)$, then we get the following result:

Theorem 83 (L^2 weak law of large numbers)
Let X, X_i are uncorrelated and identically distributed. Then the empirical mean $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to $\mathbb{E}[X]$, in L^2 and in probability, as $n \to \infty$.

(Note that on probability spaces, **convergence in μ-measure** becomes **convergence in probability**.) For completeness, we'll write out the proof in full:

Proof. From our above calculation, we have

$$\left|\left|\overline{X_n} - \mathbb{E}X\right|\right|_2^2 = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X])\right)^2\right] = \frac{\operatorname{Var}(X)}{n}.$$

Because this goes to 0 as $n \to \infty$, we do indeed have $\overline{X_n} \xrightarrow{L^2} \mathbb{E}[X]$. For convergence in probability, note that for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mathbb{E}[X]| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\operatorname{Var} \overline{X_n}}{\varepsilon^2} = 0,$$

which is the condition we wish to prove. □

This proves the simplest version of the law of large numbers, and the main idea is the geometric fact that if we are given a bunch of orthogonal vectors $v_1, v_2, \cdots$, their average will have small norm. And next time, we'll see how to extend these types of results beyond L^2 geometry.

9 October 2, 2019

Last time, we started studying moments of random variables, leading us to define the variance $\operatorname{Var} X = \mathbb{E}[(X - \mathbb{E}X)^2] \in [0, \infty)$ for any random variable $X \in L^2$ (meaning that the L^2 norm $||X||_2 = \mathbb{E}[X^2]^{1/2}$ is finite). We then proved the L^2 weak law of large numbers, which states that given pairwise uncorrelated, identical distributed random variables $X, X_i \in L^2$, we have $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$ converging to $\mathbb{E}[X]$ in L^2 and in probability. We'll strengthen this result in a few ways today, by relaxing the L^2 assumption, showing almost-surely convergence under some conditions, and generalizing to triangular arrays. For the rest of class, consider an array of random variables of the following form:

$$\begin{array}{cccc} X_{1,1} & & & \\ X_{2,1} & X_{2,2} & & \\ X_{3,1} & X_{3,2} & X_{3,3} & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

Defining $S_n = \sum_{k=1}^n X_{n,k}$ to be the sum of the nth row of the array, our central question of study is whether $\frac{S_n}{b_n}$ converges to a constant for some deterministic b_n (usually we'll have $b_n = n$, but we'll see some other examples as well). We'll first convert the proof from last time to this new notation:

Theorem 84 (L^2 weak law for triangular arrays)
Suppose we have a triangular array $X_{i,j}$ as above, and assume that $\operatorname{Var}(X_{n,k}) \leq C < \infty$ for all k, n. Then $\frac{S_n - \mathbb{E}[S_n]}{n}$ converges to 0, in L^2 and in probability, as $n \to \infty$.

Proof. By the Pythagorean theorem, because the S_is are uncorrelated, we can write

$$\left|\left|\frac{S_n - \mathbb{E}[S_n]}{n}\right|\right|_2^2 = \frac{1}{n^2}||S_n - \mathbb{E}[S_n]||_2^2 \leq \frac{1}{n^2}\sum_{k=1}^n \operatorname{Var}(X_{n,k}) \leq \frac{C}{n},$$

which goes to 0 as $n \to \infty$, showing L^2 convergence. Convergence in probability again follows from Chebyshev (much like Theorem 83) because $\operatorname{Var}\left(\frac{S_n - \mathbb{E}[S_n]}{n}\right) = \frac{1}{n^2}\operatorname{Var}(S_n) \leq \frac{C}{n}$. □

Under the L^2 assumption, it was okay for us to just assume that all of the vectors were pairwise orthogonal, but we're going to have to work with something stronger going forward. So from now on, we'll assume **independence of the random variables in each row**.

Theorem 85 (Weak law for triangular arrays)

Take the same triangular array as above, and assume we have a sequence $b_n \to \infty$ satisfying

$$\sum_{k=1}^{n} \mathbb{P}(|X_{n,k}| > b_n) \to 0, \quad \sum_{k=1}^{n} \frac{\mathbb{E}(X_{n,k}^2; |X_{n,k}| \le b_n)}{b_n^2} \to 0 \text{ as } n \to \infty.$$

Under these assumptions, define the random variables

$$Y_{n,k} = X_{n,k} \cdot 1\{|X_{n,k}| \le b_n\}, \quad T_n = \sum_{k=1}^{n} Y_{n,k}.$$

Then $\frac{S_n - \mathbb{E}[T_n]}{b_n}$ converges to 0 in probability.

This theorem is stated in an ugly way so that it is easier to prove – we are defining the random variables T_n because the expectations $\mathbb{E}[S_n]$ may not be defined in general.

Proof. First, we show that S_n is generally close to T_n, and we can in fact show that S_n and T_n are equal with high probability. Indeed,

$$\mathbb{P}(S_n \neq T_n) \le \sum_{k=1}^{n} \mathbb{P}(X_{n,k} \neq Y_{n,k}) = \sum_{k=1}^{n} \mathbb{P}(|X_{n,k}| > b_n)$$

because T_n is the sum of truncated $X_{n,k}$s, while S_n is just the sum of the original $X_{n,k}$s. This right-hand side goes to 0 by the first of our (convenient) assumptions, so $S_n - T_n \to 0$ in probability, and therefore $\frac{S_n - T_n}{b_n} \to 0$ in probability as well (since $b_n \to \infty$). Thus, it suffices to prove the statement for T_n in place of S_n (because $\frac{S_n - \mathbb{E}[T_n]}{b_n} = \frac{T_n - \mathbb{E}[T_n]}{b_n} + \frac{S_n - T_n}{b_n}$). But we've designed T_n so that it is nicely bounded: specifically, we have

$$\left|\left| \frac{T_n - \mathbb{E}T_n}{b_n} \right|\right|_2^2 = \frac{1}{b_n^2} \sum_{k=1}^{n} \text{Var}(Y_{n,k})$$

since the $Y_{n,k}$s are uncorrelated – they are functions of the $X_{n,k}$s, which are independent by assumption. (We do need the $X_{n,k}$s to be **independent**, because just knowing that the $X_{n,k}$s are uncorrelated wouldn't guarantee that the $Y_{n,k}$s are uncorrelated.) Then because $\text{Var}(Y_{n,k})^2 \le \mathbb{E}[Y_{n,k}]^2$, we can further bound

$$\left|\left| \frac{T_n - \mathbb{E}T_n}{b_n} \right|\right|_2^2 \le \frac{1}{b_n^2} \sum_{k=1}^{n} \mathbb{E}(Y_{n,k})^2,$$

which goes to zero by our second assumption. Thus we have convergence in L^2 for T_n and thus convergence in probability, proving the claim. □

We can now use this "ugly" theorem to prove nicer results: the rest of our theorems today will involve iid variables, specifically those in a triangular array as shown below:

$$\begin{array}{cccc} X_1 & & & \\ X_1 & X_2 & & \\ X_1 & X_2 & X_3 & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

In particular, our columns are now strongly dependent on each other, but the rows are independent if our X_is are independent.

Theorem 86 (Weak law of large numbers for iid sequences)
Let X, X_k be independent and identically distributed random variables, such that $g(x) = x\mathbb{P}(|X| \geq x)$ goes to 0 as $x \to \infty$. Then as $n \to \infty$, $\frac{S_n}{n} - \mu_n$ converges to 0 in probability, where $\mu_n = \mathbb{E}(X; |X| \leq n)$.

Notice that this condition does not require us to assume that $\mathbb{E}[X]$ is finite (which is equivalent to $\mathbb{E}|X|$ being finite, because we need both the positive and negative integrals to be finite). In particular, $\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| \geq x)dx$ (as we proved on our homework), and the condition that $g(x) \to 0$ only means that $\mathbb{P}(|X| \geq x)$ decays faster than $\frac{1}{x}$ as $x \to \infty$. But there are functions that decay faster than $\frac{1}{x}$ that aren't integrable – for instance,

$$\int_a^b \frac{dx}{x \log x} = \log\log x|_a^b$$

diverges as $b \to \infty$. So that's why we define μ_n in the way that we do – the assumptions on our random variables are slightly weaker than having a finite first moment.

Proof. We wish to apply Theorem 85 with $b_n = n$ (and having X_k take the role of $X_{n,k}$). To do so, we just need to check the two conditions. For the first one, we can rewrite

$$\sum_{k=1}^n \mathbb{P}(|X_{n,k}| \geq b_n) = n\mathbb{P}(|X| \geq n) = g(n),$$

which goes to 0 by assumption as $n \to \infty$. For the second condition, define $Y_n = X_n \cdot 1\{|X| \leq n\}$, so that we must prove that $n \cdot \frac{\mathbb{E}(Y_n^2)}{n^2} = \frac{\mathbb{E}(Y_n^2)}{n} \to 0$ as $n \to \infty$. But again using that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)dt$ for a nonnegative random variable X,

$$\frac{1}{n}\mathbb{E}(Y_n^2) = \frac{1}{n}\int_0^\infty \mathbb{P}(Y_n^2 \geq t)dt = \frac{1}{n}\int_0^\infty \mathbb{P}(|Y_n| \geq \sqrt{t})dt.$$

Making the change of variables $y = \sqrt{t}, dt = 2ydy$, our integral becomes

$$\frac{1}{n}\mathbb{E}(Y_n^2) = \frac{1}{n}\int_0^\infty \mathbb{P}(|Y_n| \geq y) \cdot 2ydy.$$

We can then cap the upper limit of the integral at n, because Y_n is always at most $b_n = n$ and thus $\mathbb{P}(Y_n \geq y) = 0$ above that point, leading us to

$$\frac{1}{n}\mathbb{E}(Y_n^2) = \frac{1}{n}\int_0^n \mathbb{P}(|Y_n| \geq y)2ydy = \frac{1}{n}\int_0^n \mathbb{P}(n \geq |X| \geq y)2ydy \leq \frac{2}{n}\int_0^n g(y)dy.$$

But since this last expression is twice the average value of g on $[0, n]$, and g decays to zero (and has a finite supremum), the average will also go to zero as $n \to \infty$. Thus both conditions of Theorem 85 hold, and μ_n is the value of $\frac{\mathbb{E}[T_n]}{n}$ in that theorem's language, implying the result. □

Corollary 87
Suppose X, X_k are iid random variables with finite mean $\mathbb{E}[X] = \mu$. Then $\frac{S_n}{n}$ converges to μ in probability.

Proof. Similarly to Markov's inequality, we can write

$$g(x) = x\mathbb{P}(|X| \geq x) = \mathbb{E}[x; |X| \geq x] \leq \mathbb{E}[|X|; |X| \geq x].$$

Now notice that the expression $(|X|; |X| \geq x) = |X| \cdot 1\{|X| \geq x\}$ goes to 0 as $x \to \infty$ almost surely, and all $|X|; |X| \geq x$s are bounded from above by $|X|$, which is integrable by assumption. So because the expectations $\mathbb{E}[|X|; |X| \geq x]$ are

Lebesgue integrals, we can apply the dominated convergence theorem to find that $g(x) \to 0$ as $x \to \infty$ (specifically, we can apply it to the functions $|X| \cdot 1\{|X| \geq x\}$ for integer x and notice that these functions are nonincreasing for all real x). Therefore, $\frac{S_n}{n} - \mu_n$ goes to 0 in probability by Theorem 86. It remains to check that $\mu_n \to \mu$, but we have

$$\mu - \mu_n = \mathbb{E}[X] - \mathbb{E}[X; |X| \leq n] \leq \mathbb{E}\left[|X|; |X| > n\right],$$

which, just like before, goes to zero. Thus $\frac{S_n}{n} - \mu$ converges to 0 in probability, as desired. □

We'll now show a stronger form of convergence under these same conditions:

Theorem 88 (Strong law of large numbers)
Suppose X, X_k are iid random variables with finite mean $\mathbb{E}[X] = \mu$. Then $\frac{S_n}{n}$ converges to μ **almost surely**.

Proof. We may assume without loss of generality that $X \geq 0$, because we can write $X = X_+ - X_-$ and prove separately that the average of the $(X_k)_+$s (resp. $(X_k)_-$s) converges to $\mathbb{E}[X_+]$ (resp. $\mathbb{E}[X_-]$) almost surely. For this proof, we need to use a different truncation than before, defining

$$Y_k = X_k \cdot 1\{X_k \leq k\}.$$

In the previous proof, we truncated each row of our triangular array at the same level n, so the sum of a given row would be $\sum_{k=1}^n X_k 1\{|X_k| \leq n\}$. But this time, we instead have

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

The first step of our proof is again to show that S_n and T_n are close. Notice that

$$\sum_{k=1}^\infty \mathbb{P}(X_k \neq Y_k) \leq \sum_{k=1}^\infty \mathbb{P}(X_k \geq k) \leq \sum_{k=1}^\infty \mathbb{P}(X \geq k)$$

because the X_ks are all identically distributed as X, and now we can convert the right-hand side from a Riemann sum to an integral (with overall error at most $\mathbb{P}(X \geq 0) = 1$) to get

$$\sum_{k=1}^\infty \mathbb{P}(X_k \neq Y_k) \leq 1 + \int_0^\infty \mathbb{P}(X \geq t) dt = 1 + \mathbb{E}[X] < \infty.$$

By the Borel-Cantelli lemma (homework), this means that $\mathbb{P}(X_k \neq Y_k \text{ i.o.})$ (infinitely often) must be zero, where the event is defined as

$$\{X_k \neq Y_k \text{ i.o.}\} = \{\omega \in \Omega : X_k(\omega) \neq Y_k(\omega) \text{ for infinitely many } k\}.$$

This implies that $\boxed{\frac{S_n - T_n}{n} \to 0}$ almost surely, because with probability one, there are only finitely many indices where $X_k - Y_k$ will be nonzero, so $S_n - T_n$ converges to some finite value C. (Thus $\frac{S_n - T_n}{n}$ is then bounded by $\frac{C}{n}$, which goes to zero.)

Next, we can also show that $\boxed{\frac{\mathbb{E}[T_n]}{n} \to \mu}$ almost surely, because

$$\mu - \frac{\mathbb{E}[T_n]}{n} = \frac{\mathbb{E}[S_n - T_n]}{n} = \frac{1}{n} \sum_{k=1}^n \mathbb{E}\left[X; X \geq k\right].$$

As we've shown before, $\mathbb{E}[X; X \geq k] \to 0$ as $k \to \infty$ by the dominated convergence theorem, so the right-hand side

(which is an average of these types of terms) also converges to 0, as desired. In other words, the two boxed statements above allow us to replace S_n with T_n, and μ with $\frac{\mathbb{E}[T_n]}{n}$, so it now suffices to show that $\frac{T_n - \mathbb{E}[T_n]}{n} \to 0$ almost surely. To do so, we'll start by calculating the L^2 norm as usual. We have

$$\left|\left|\frac{T_n - \mathbb{E}T_n}{n}\right|\right|_2^2 = \frac{1}{n^2}\sum_{k=1}^{n} \text{Var}(Y_k) \leq \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}[Y_k^2].$$

However, this expression is more difficult to work with than the ones in previous proofs, because our X_ks aren't even assumed to have finite variance. **Suppose** we knew that $\sum_{k=1}^{\infty} \frac{\mathbb{E}[Y_k^2]}{k^2}$ is finite and equal to some value A. Then we could bound the L^2 norm by breaking up the sum into parts:

$$\begin{aligned}
\frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}[Y_k^2] &= \sum_{k=1}^{n} \left(\frac{k}{n}\right)^2 \frac{\mathbb{E}[Y_k^2]}{k^2} \\
&\leq \sum_{k \leq \log n} \left(\frac{\log n}{n}\right)^2 \cdot \frac{\mathbb{E}[Y_k^2]}{k^2} + \sum_{k > \log n} \frac{\mathbb{E}[Y_k^2]}{k^2} \\
&\leq \left(\frac{\log n}{n}\right)^2 A + \sum_{k > \log n} \frac{\mathbb{E}[Y_k^2]}{k^2}.
\end{aligned}$$

The first term would then go to 0 as $n \to \infty$, and so would the second because we'd be taking the tail of a finite sum. So we will calculate the expression in blue above, converting to an integral just like in Theorem 86, to get

$$\sum_{k=1}^{\infty} \frac{1}{k^2}\mathbb{E}[Y_k^2] = \sum_{k=1}^{\infty} \frac{1}{k^2}\int_0^{\infty} 2y\mathbb{P}(X \geq y)dy.$$

Because the integrand is positive, we can swap the sum and the integral to get

$$\sum_{k=1}^{\infty} \frac{1}{k^2}\mathbb{E}[Y_k^2] = \int_0^{\infty} \mathbb{P}(X \geq y)\left[2y\left(\sum_{k \geq y} \frac{1}{k^2}\right)\right]dy.$$

Now the inner bracketed term is uniformly bounded over all $y \geq 0$ by some finite constant c, because it is bounded by $2y$ times a constant for small y and the inner sum is proportional to $\frac{1}{y}$ for large y. Thus, we can write

$$\sum_{k=1}^{\infty} \frac{1}{k^2}\mathbb{E}[Y_k^2] \leq c\int_0^{\infty} \mathbb{P}(X \geq y)dy = c\mathbb{E}[X] < \infty,$$

as desired. Thus the L^2 norm of $\frac{T_n - \mathbb{E}[T_n]}{n}$ goes to zero and we have convergence in probability as usual, but that's not what we wanted to prove. Instead, notice that this calculation gives us a useful rate of decay. For any subsequence of the integers $n(\ell)$ (going to infinity), we have (by the same logic but now summing over the subsequence)

$$\sum_{\ell \geq 1}\left|\left|\frac{T_{n(\ell)} - \mathbb{E}T_{n(\ell)}}{n(\ell)}\right|\right|_2^2 \leq \sum_{\ell \geq 1} \frac{1}{n(\ell)^2}\sum_{k=1}^{n(\ell)} \mathbb{E}[Y_k^2].$$

Swapping the order of summation (again allowed because everything is positive here), this simplifies to

$$= \sum_{k=1}^{\infty} \mathbb{E}[Y_k]^2 \sum_{\ell \geq 1} \frac{1\{n(\ell) \geq k\}}{n(\ell)^2}.$$

So if we can pick a sequence $n(\ell)$ such that $\sum_{\ell \geq 1} \frac{1\{n(\ell) \geq k\}}{n(\ell)^2} \leq \frac{C}{k^2}$, we will be in the case above (where the blue expression is finite). Taking $n(\ell) = \alpha^\ell$ for some $\alpha > 1$, we do indeed find by a geometric series bound that $\sum_{\ell \geq 1} \frac{1\{n(\ell) \geq k\}}{n(\ell)^2} \leq \frac{C(\alpha)}{k^2}$

for some constant $C(\alpha)$ and for all $k \geq 1$. So the argument above with the L^2 norm tells us that

$$\sum_{\ell \geq 1} \left\| \frac{T_{\alpha^\ell} - \mathbb{E}T_{\alpha^\ell}}{\alpha^\ell} \right\|_2^2 < \infty.$$

Therefore, for any $\varepsilon > 0$, $\left| \frac{T_{\alpha^\ell} - \mathbb{E}[T_{\alpha^\ell}]}{\alpha^\ell} \right| > \varepsilon$ only occurs finitely many times almost surely (because the sum of the probabilities of these events is finite by Chebyshev's inequality, and then we can apply Borel-Cantelli). So $\frac{T_{\alpha^\ell} - \mathbb{E}[T_{\alpha^\ell}]}{\alpha^\ell} \to 0$ as $\ell \to \infty$ almost surely, and we've shown almost-sure convergence along a subsequence. Finally, for the full sequence, we can use that the Xs are nonnegative. In particular, for any $\alpha^\ell \leq n \leq \alpha^{\ell+1}$, because $T_{\alpha^\ell} \leq T_n \leq T_{\alpha^{\ell+1}}$ and $\alpha^{\ell+1} \geq n \geq \alpha^\ell$, we also have

$$\frac{T_{\alpha^\ell}}{\alpha^{\ell+1}} \leq \frac{T_n}{n} \leq \frac{T_{\alpha^{\ell+1}}}{\alpha^\ell}.$$

Taking $\ell \to \infty$ and using almost-sure convergence along a subsequence, we get

$$\frac{\mu}{\alpha} \leq \liminf \frac{T_n}{n} \leq \limsup \frac{T_n}{n} \leq \alpha\mu.$$

Finally, taking $\alpha \downarrow 1$ shows that the lim inf and lim sup of $\frac{T_n}{n}$ are both μ, so $\frac{T_n}{n}$ converges to μ almost surely. Combining this with the fact that $\frac{S_n - T_n}{n} \to 0$ almost surely (from above) finishes the proof. □

10 October 7, 2019

Class today started with an attendance quiz "not intended to be difficult:"

Problem 89

Let X be a random variable satisfying

$$\mathbb{P}(X \leq x) = \begin{cases} 0 & x \leq 1, \\ 1 - \frac{1}{x} & x \geq 1. \end{cases}$$

Calculate $\mathbb{E}[X^p; X \leq n]$ without explicitly finding the probability density of X.

Solution. We're given the tail bound $\mathbb{P}(X \geq x) = \frac{1}{x}$ for all $x \geq 1$, and we wish to compute

$$\mathbb{E}[X^p; X \leq n] = \mathbb{E}[Y^p], \text{ where } Y = X \cdot 1\{X \leq n\}.$$

To do this, we can use the useful formula (like we used in last lecture)

$$\mathbb{E}[Y^p] = \int_0^\infty \mathbb{P}(Y^p \geq t)dt.$$

If we now change variables by setting $t = y^p$, $dt = py^{p-1}dy$, this simplifies to

$$\mathbb{E}[Y^p] = \int_0^\infty py^{p-1}\mathbb{P}(Y \geq y)dy = \int_0^n py^{p-1}\mathbb{P}(Y \geq y)dy = \int_0^n py^{p-1}\mathbb{P}(y \leq X \leq n)dy,$$

at which point we can plug in $\mathbb{P}(y \leq X \leq n) = \frac{1}{y} - \frac{1}{n}$ and directly integrate to get the answer. □

(The main takeaway is that this strategy gives us a way to bound $\mathbb{E}[X^p; X \leq n]$ without needing an explicit density function, as long as we know something like $\mathbb{P}(X \geq x) \leq \frac{c}{x}$.)

Last lecture, we finished by proving the strong law of large numbers, which states that if X and X_i are iid and $\mathbb{E}[X]$ is finite, then $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ almost surely. Today, we'll work "around that value" and find bounds on the probability that $\frac{S_n}{n}$ is approximately $\mu + \varepsilon$ for some constant ε. We'll start with a special case:

Example 90

Suppose $X \sim N(0,1)$ is standard Gaussian with density $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$.

In this case, $S_n = \sum_{i=1}^{n} X_i \sim N(0,n)$ is also Gaussian with variance n, meaning that $\frac{S_n}{\sqrt{n}} \sim N(0,1)$, S_n has probability density $\phi\left(\frac{x}{\sqrt{n}}\right)\frac{dx}{\sqrt{n}}$, and the strong law of large numbers tells us that $\frac{S_n}{n} \to 0$ almost surely as $n \to \infty$. But we might be interested in behavior away from the mean as well – for example, we may want to calculate

$$\mathbb{P}\left(\frac{S_n}{n} \geq \varepsilon\right) = \mathbb{P}\left(N(0,1) \geq \varepsilon\sqrt{n}\right) = \int_{z \geq \varepsilon\sqrt{n}} \phi(z) dz.$$

Since ϕ decays very fast, this is approximately on the order of $\phi(\varepsilon\sqrt{n}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{n\varepsilon^2}{2}\right)$. And we can show more precisely that the integral is in fact $\exp\left(-\frac{n\varepsilon^2}{2} + o(n)\right)$, so the corrections are not leading order, and differing from the mean by ε is exponentially unlikely in n.

Example 91

Suppose $X \sim \text{Ber}(p)$ is a Bernoulli random variable, meaning that X is 1 with probability p and 0 with probability $(1-p)$. Then S_n is Binomial $\sim \text{Bin}(n,p)$ with probability mass function $\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Again by the strong law of large numbers, $\frac{S_n}{n}$ approaches p almost surely, and the **central limit theorem** further tells us that

$$\frac{S_n - \mu}{\sigma} = \frac{S_n - np}{\sqrt{np(1-p)}} \approx N(0,1).$$

We haven't proved the central limit theorem yet, but we can do a heuristic calculation to explain why we do get an approximately normal distribution. Applying Stirling's approximation $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, we find that for any $0 < x < 1$ where nx is an integer,

$$\mathbb{P}(S_n = nx) = \binom{n}{nx} p^{nx}(1-p)^{n(1-x)} \sim \frac{\sqrt{2\pi n}}{\sqrt{2\pi nx}\sqrt{2\pi n(1-x)}} \cdot \frac{(n/e)^n}{(nx/e)^{nx}(n(1-x)/e)^{n(1-x)}} \cdot p^{nx}(1-p)^{n(1-x)}.$$

(This approximation is good when x is not too close to 0 or 1, because then the factorials in the binomial coefficient are all large.) The $(n/e)^n$ factors cancel out in the top and bottom, leaving us with

$$\mathbb{P}(S_n = nx) \sim \frac{1}{\sqrt{2\pi nx(1-x)}} \left(\frac{p}{x}\right)^{nx} \left(\frac{1-p}{1-x}\right)^{n(1-x)},$$

which can be rewritten in terms of an exponential as

$$\mathbb{P}(S_n = nx) \sim \frac{1}{\sqrt{2\pi nx(1-x)}} \exp\left(-n\left[x \log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}\right]\right).$$

This inner bracketed term is often denoted $I_p(x)$ or $\mathcal{H}(x|p)$, and it is also called the **binary relative entropy**. Remembering that our goal is to make this look like a Gaussian, and specifically that $z = \frac{S_n - np}{\sqrt{np(1-p)}} = \frac{n(x-p)}{\sqrt{np(1-p)}}$ should be approximately standard Gaussian, we're motivated to make a change of variables from x to z. Implicitly, we need z to be bounded ($O(1)$) for the next calculation to be valid. Specifically, assuming that $x = p \pm \frac{O(1)}{\sqrt{n}}$ (so S_n is pretty

close to the mean), we can plug in $x = p + \frac{1}{n}\sqrt{np(1-p)}z$ to get

$$\mathbb{P}\left(S_n = np + \sqrt{np(1-p)}z\right) \sim \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-nI_p\left(p + \frac{\sqrt{np(1-p)}z}{n}\right)\right),$$

where we have replaced the xs in the denominator of the prefactor with p (which is allowed because x is close to p by assumption). Additionally, because I_p is being evaluated close to its minimizer p, we can Taylor expand around that minimum. The first two terms of the Taylor series are zero because $I_p(p) = I_p'(p) = 0$, and thus (after some calculus) we have $I_p(x) = \frac{z^2}{2n} + \frac{O(1)}{n^{3/2}}$. Plugging everything back in, we find that

$$\mathbb{P}\left(S_n = np + \sqrt{np(1-p)}z\right) \sim \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(\frac{-z^2}{2} + \frac{O(1)}{\sqrt{n}}\right),$$

at which point the error term here can be absorbed as a constant of proportionality. So we've recoverved the Gaussian density – specifically, our random variable $\frac{S_n - np}{\sqrt{np(1-p)}}$ converges to the Gaussian in the sense that for any fixed $a < b$

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) \sim \sum_{z \in [a,b]} \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{z^2}{2}\right),$$

where z sums over all points in the lattice $\frac{\mathbb{Z} - np}{\sqrt{np(1-p)}}$ within $[a, b]$ to guarantee that we only sum over integer values of S_n (since we still actually have a discrete binomial distribution). So this hopefully gives a good sense of how the central limit theorem statement will look in general.

Remark 92. *It's important to note that the calculation above does not actually tell us how to estimate $\mathbb{P}(S_n - np \geq n\varepsilon)$ for some constant ε. We made the assumption in our argument that $S_n - np$ is on the order of $\sqrt{n}$ – because of the extra $\sqrt{n}$ term in our Taylor expansion, we can go a little farther out, but we can check that we get all the way to $S_n - np$ being linear in n and still expecting Gaussian behavior to hold.*

To actually understand the tail behavior of a binomial random variable, we have to go back to

$$\mathbb{P}(S_n - np \geq n\varepsilon) = \sum_{k \geq n(p+\varepsilon)} \binom{n}{k} p^k (1-p)^{n(1-k)}.$$

We can approximate this by Stirling's formula – because we're now summing over k larger than the mean, the sum will be dominated by the smallest values of k. Skipping the calculations, we find that

$$\mathbb{P}(S_n - np \geq n\varepsilon) \approx \exp\left(-n(I_p(p+\varepsilon)) + o(n)\right).$$

(Here $o(n)$ is not the best possible approximation, but it's good enough for us because we only care about the leading term.) And this is not the same as the naive guess we might have (of just plugging into the Gaussian density) because this time we can't Taylor expand I_p around p. In other words, we find that even if we have the central limit theorem result above, we still have

$$\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \geq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right) \not\sim \exp\left(-\frac{n\varepsilon^2}{2p(1-p)}\right).$$

Our main goal for next lecture will thus be to find conditions for when iid random variables X, X_i with finite mean satisfy a relation of the form

$$\mathbb{P}(S_n - n\mu \geq n\varepsilon) = \exp\left(-nf(\varepsilon, \mu) + o(n)\right)$$

for some function f. This is a subject called **large deviations theory**, which is pretty deep, but we'll only study it

briefly in this class. We'll finish today with a final example which has some important large deviations concepts:

Example 93

Consider a $k \times n$ table of bins, where we fix the number of rows k and study the limit as $n \to \infty$. Place nkp balls into the bins of the table uniformly at random (where each bin can hold at most one ball). We wish to estimate the probability of the event $F = \{\text{every column has at least one ball}\}$.

First of all, we must have $nkp \geq n$ (so that there are enough balls to have one per column). For simplicity, let's assume this inequality is strict so that we have $kp > 1$.

We'll start with the simple case $k = 2$, in which the probability can be calculated exactly. In this case, we have a $2 \times n$ table with $2np$ filled bins. If all columns have at least one ball (that is, if F occurs), then nx of them have a single ball and the remaining $n(1-x)$ have two balls, where $nx + 2n(1-x) = 2np \implies x = 2(1-p)$. The probability of this occurring if $2np$ of our $2n$ bins are randomly chosen to be filled is then

$$\mathbb{P}(F) = \frac{1}{\binom{2n}{2np}} \binom{n}{nx} 2^{nx} = \frac{1}{\binom{2n}{2np}} \binom{n}{2n(1-p)} 2^{2n(1-p)},$$

because we pick which nx columns only get one ball and whether the top or bottom bin is filled. (And if we want a nicer expression, we can always use Stirling's approximation to get an asymptotic bound.) But even moving up to $k = 3$ already complicates the problem – we could separate our columns into those with one, two, and three occupied entries, but now the number of each is no longer uniquely determined just by the total number of balls (for example, we could fill two adjacent columns with one and three balls, or with two and two). So combinatorial calculations won't be able to get us an immediate answer.

Instead, we'll turn to probability to help us solve this problem. Let $\theta \in (0,1)$ be a parameter (which is arbitrary for now), and let $X_1, \cdots, X_n$ be iid $\text{Bin}(k, \theta)$ random variables. We claim that regardless of the value of θ, we have (here $\mathbb{P}_\theta$ denotes the probability measure coming from the X_is)

$$\mathbb{P}(F) = \mathbb{P}_\theta \left(X_i \geq 1 \text{ for all } 1 \leq i \leq n \,\middle|\, \sum_{i=1}^n X_i = nkp \right).$$

To understand this, the idea is that X_i represents the number of balls in the kth column if we fill every bin independently with probability θ. Then if we condition on the total number of balls $\sum_{i=1}^n X_i$ being nkp, we're indeed placing nkp balls into the bins uniformly at random, and we want each column to have at least one ball. For notational convenience, define $G = \{X_i \geq 1 \text{ for all } 1 \leq i \leq n\}$ and $S = \sum_{i=1}^n X_i = nkp$. By Bayes' rule, we can rewrite the probability that we want as

$$\mathbb{P}(F) = \mathbb{P}_\theta(G|S) = \frac{\mathbb{P}_\theta(G \cap S)}{\mathbb{P}_\theta(S)} = \frac{\mathbb{P}_\theta(G)\mathbb{P}_\theta(S|G)}{\mathbb{P}(\text{Bin}(nk, \theta) = nkp)}.$$

All of the terms on the right-hand side can now be dealt with relatively easily. The denominator can be calculated using the work in Example 91, and $\mathbb{P}_\theta(G)$ is much easier to calculate than $\mathbb{P}_\theta(G|S)$ because the events in the various columns are now independent (meaning that it is just $\mathbb{P}(\text{Bin}(k, \theta) \geq 1)^n$). Finally, the conditional probability $\mathbb{P}_\theta(S|G)$ can just be bounded from above by 1. This leaves us with

$$\mathbb{P}_\theta(G|S) \leq \frac{\mathbb{P}(\text{Bin}(k, \theta) \geq 1)^n}{\mathbb{P}(\text{Bin}(nk, \theta) = nkp)},$$

and now we can get the best possible bound by optimizing the right-hand side over θ. And now we may want to know whether the optimal θ actually gives us a bound on $\mathbb{P}(F)$ – to see that, we can take a closer look at the term we

bounded by 1 and write it out as

$$\mathbb{P}_\theta(S|G) = \mathbb{P}_\theta\left(\sum_{i=1}^n X_i = nkp \,\middle|\, X_i \geq 1 \text{ for all } 1 \leq i \leq n\right) = \mathbb{P}\left(\sum_{i=1}^n Y_i = nkp\right),$$

where Y_i is distributed as X_i conditioned on $\{X_i \geq 1\}$. For our bound to be good, we should try to make this probability large, so it makes sense to choose θ so that $\mathbb{E}_\theta Y = kp$. Explicitly, this means that we want to pick θ such that

$$\mathbb{E}_\theta(X|X \geq 1) = \frac{k\theta}{1-(1-\theta)^k} = kp.$$

Picking this value of θ, we know that $\frac{\sum Y_i}{n}$ converges almost surely to kp as $n \to \infty$, and with a bit more work (specifically the local central limit theorem), it can be shown that $\mathbb{P}(\sum Y_i = knp) = \exp(-o(n))$, meaning that our estimate of $\mathbb{P}(F)$ is fairly good.

In summary, this example shows that a specific counting configuration problem can be made simpler with probability. Next lecture, we'll look at the more general question we've proposed about large deviations!

11 October 9, 2019

(Another homework assignment has been posted, and it will be due in a week.) Last lecture, we studied sums of independent random variables. In particular, we've shown that if X, X_i are iid with finite mean, then their average approaches μ, but we're now curious about the probability that this average is more than $\mu + \varepsilon$ for some constant $\varepsilon > 0$. Last time, we did a calculation to verify that when X is standard normal, we have

$$\mathbb{P}(S_n \geq n\varepsilon) = \exp\left(-\frac{n\varepsilon^2}{2} + o(n)\right),$$

and similarly when X is Bernoulli with probability p, we have

$$\mathbb{P}(S_n - np \geq n\varepsilon) = \exp(-nI_p(p+\varepsilon) + o(n)).$$

Today, we'll see that these formulas are part of a more general structure.

Definition 94

The **moment generating function** (also **mgf**) of a random variable X is the function $m(\theta) = \mathbb{E}[e^{\theta X}]$.

Because $e^{\theta X} > 0$ for all θ, we have $m(\theta) \in (0, \infty]$ for any θ. Additionally, we always have $m(0) = 1$, but it's possible that this is the only finite value for the moment generating function.

Lemma 95

For any random variable X, let $m(\theta) = \mathbb{E}[e^{\theta X}]$, and define

$$\theta_+ = \sup\{\theta : m(\theta) < \infty\}, \quad \theta_- = \inf\{\theta : m(\theta) < \infty\}.$$

(Because $m(0) < \infty$, $\theta_+ \in [0, \infty]$ and $\theta_- \in [-\infty, 0]$.) Then on the open interval (θ_-, θ_+), $m(\theta)$ is a smooth function with kth derivative $m^{(k)}(\theta) = \mathbb{E}[X^k e^{\theta X}] < \infty$. Additionally, if $\theta_+ > 0$, then $m^{(k)}(\theta) \to \mathbb{E}[X^k]$ as $\theta \downarrow 0$.

While we can't generally exchange derivatives and expectation values, this lemma tells us that we can indeed do so for the moment generating function. We won't do the proof in full here, but to show that $m(\theta)$ is finite on the interval

(θ_-, θ_+), notice that for any $a < b < c$, $e^{bx} \le e^{ax} + e^{cx}$, so $\mathbb{E}[e^{bX}] \le \mathbb{E}[e^{aX}] + \mathbb{E}[e^{cX}]$ (meaning that if the moment generating function is finite at a and c, it is also finite in between). Smoothness follows by checking the definition of the derivative (which also exists on (θ_-, θ_+) because $\mathbb{E}[e^{\theta X}]$ is a power series in θ), and the last statement follows from the dominated convergence theorem.

Definition 96

The **cumulant generating function** (also **cgf**) of a random variable is the function $\kappa(\theta) = \log m(\theta)$.

We're now almost ready to answer or question (about calculating $\mathbb{P}(S_n \ge na)$ for some $a > \mu$), and we just need a bit more notation. Let $x_{\max}$ be the **essential supremum** of the random variable X, meaning that

$$x_{\max} = \sup(\operatorname{supp} \mathcal{L}_X) = \sup\{x \in \mathbb{R} : \mathbb{P}(X > x) > 0\}.$$

The idea is that nothing interesting happens above $x_{\max}$, because for any $a > x_{\max}$, $\mathbb{P}(X \ge a) = 0$, so $\mathbb{P}(S_n \ge na) = 0$ (each term in the sum S_n is less than a, so their sum must be less than na). And if $a = x_{\max}$, then $\mathbb{P}(S_n \ge na) = \mathbb{P}(X = a)^n$ (because we must have all terms exactly a to have sum na), which can be positive if we have a point of positive measure at a. So we already know the answer in those cases, and from here on **we will assume** $\mathbb{E}X < a < \text{ess sup } X$. (In particular, this also rules out the case where $\mathbb{E}X = \text{ess sup } X$, in which case the distribution is concentrated entirely at one value and nothing interesting happens.)

Theorem 97 (Cramér)

Let X, X_i be iid random variables. If $m(\theta) = \mathbb{E}[e^{\theta X}]$ is finite for some $\theta > 0$, then $\mathbb{E}[X] = \mu \in [-\infty, \infty)$ is well-defined (by Lemma 95). In addition, for any $\mu < a < x_{\max} = \text{ess sup } X$, the sum $S_n = \sum_{i=1}^n X_i$ satisfies the "large deviations principle"

$$-\lim_{n\to\infty} \frac{1}{n} \log(\mathbb{P}(S_n \ge na)) = I(a) = \sup_{\theta \ge 0} [\theta a - \kappa(\theta)] = \sup_{\theta \in \mathbb{R}} [\theta a - \kappa(\theta)] .$$

To explain the bounds $\mu \in [-\infty, \infty)$, $\mathbb{E}[e^{\theta X}]$ being finite for some $\theta > 0$ rules out the case where X is really big (since $e^{\theta x} > x$ for sufficiently large x), but it doesn't prove that X cannot be really small (so we could have $\mathbb{E}X_+ - \mathbb{E}X_- = -\infty$). The function $\kappa^*(a) = \sup_{\theta \in \mathbb{R}} [\theta a - \kappa(\theta)]$ is often called the **Legendre dual** of the function κ.

The actual statement of the large deviations principle is a bit more advanced, but we can read that on our own: the central idea of our first equality is still that $\mathbb{P}(S_n \ge na) = \exp(-nI(a) + o(n))$. Intuitively, this result says that there may be many ways to have $X_1, \cdots, X_n$ add up to an atypically large value na, but that it's (intuitively) unlikely to have a case like $X_1 = \cdots = X_{n-1} = 0$ and $X_n = na$.

Proof. We'll first prove an upper bound on $\mathbb{P}(S_n \ge na)$. For any $\theta \ge 0$, we have

$$\mathbb{P}(S_n \ge na) \le \mathbb{P}(e^{\theta S_n} \ge e^{n\theta a}) \le \frac{\mathbb{E}(e^{\theta S_n})}{e^{n\theta a}} = \frac{m(\theta)^n}{e^{n\theta a}} = \exp\left[-n(\theta a - \kappa(\theta))\right],$$

where we have inequality instead of equality in the first relation only because of $\theta = 0$, and where we apply Markov's inequality in the second. Since this inequality holds for any $\theta \ge 0$, we thus have

$$\boxed{\mathbb{P}(S_n \ge na) \le \exp\left[-n \sup_{\theta \ge 0}(\theta a - \kappa(\theta))\right]}.$$

Next, we'll prove that the last two expressions are equal (so the supremum is the same whether we take $\theta \ge 0$ or

$\theta \in \mathbb{R}$). For any $\theta \in (\theta_-, \theta_+)$, define a probability measure $\mathbb{P}_\theta$ which assigns to any event A the probability

$$\mathbb{P}_\theta(A) = \mathbb{E}\left[1_A \frac{e^{\theta X}}{m(\theta)}\right],$$

where the expectation $\mathbb{E}$ is taken relative to the original probability measure. We'll cover this kind of operation more in a later lecture, but it is basically a reweighting of the original measure $\mathbb{P}$ (known as an **exponential tilting**), and it is a probability measure because $\mathbb{P}_\theta(\Omega) = \mathbb{E}\left[\frac{e^{\theta X}}{m(\theta)}\right] = \frac{1}{m(\theta)}\mathbb{E}[e^{\theta X}] = 1$. Thus for any measurable function f,

$$\mathbb{E}_\theta[f(X)] = \mathbb{E}\left[f(X)\frac{e^{\theta X}}{m(\theta)}\right]$$

(because this is true for simple functions by the definition and then we can approximate measurable functions by simple functions). This also means (replacing f with $f\frac{m(\theta)}{e^{\theta X}}$) that $\mathbb{E}_\theta\left[f(X)\frac{m(\theta)}{e^{\theta X}}\right] = \mathbb{E}[f(X)]$, and in particular

$$\kappa(\theta) = \log \mathbb{E}[e^{\theta X}] \implies \kappa'(\theta) = \frac{m'(\theta)}{m(\theta)} = \frac{\mathbb{E}[Xe^{\theta X}]}{\mathbb{E}[e^{\theta X}]} = \mathbb{E}\left[\frac{Xe^{\theta X}}{m(\theta)}\right] = \mathbb{E}_\theta[X],$$

where we've applied Lemma 95 to $m'(\theta)$. We can also calculate the second derivative of κ to be

$$\kappa''(\theta) = \frac{m''(\theta)}{m(\theta)} - \left(\frac{m'(\theta)}{m(\theta)}\right)^2 = \frac{\mathbb{E}[X^2e^{\theta X}]}{\mathbb{E}[e^{\theta X}]} - \left(\frac{\mathbb{E}[Xe^{\theta X}]}{\mathbb{E}[e^{\theta X}]}\right)^2 = \mathbb{E}_\theta[X^2] - \mathbb{E}_\theta[X]^2 = \mathrm{Var}_\theta(X),$$

which is strictly positive for any θ because we assumed that $\mu < \text{ess sup } X$ and thus X is not concentrated at a single point. In other words, $\kappa(\theta)$ is a **strictly convex function** on (θ_-, θ_+). And because $\kappa'(\theta) \to \mu$ as $\theta \downarrow 0$ (again by Lemma 95), we may pick some sufficiently small $\theta = \theta_0$ such that $\mu < \kappa'(\theta_0) < a$. We then find that

$$\frac{d}{d\theta}(\theta a - \kappa(\theta))_{\theta=\theta_0} = a - \kappa'(\theta_0) > 0.$$

Additionally, $\theta a - \kappa(\theta)$ is strictly concave on (θ_-, θ_+) (because θa has zero second derivative, and then we subtract $\kappa(\theta)$), so κ is increasing on (θ_-, θ_0). In particular, this implies that we indeed have

$$\boxed{\sup_{\theta \geq 0}(\theta a - \kappa(\theta)) = \sup_{\theta \in \mathbb{R}}(\theta a - \kappa(\theta))}.$$

Finally, we can prove the lower bound – we'll only do it in a nice case, assuming that the supremum $\sup_{\theta\in\mathbb{R}}(\theta a - \kappa(\theta))$ is actually attained at some value $\theta_a \in (\theta_-, \theta_+)$. This is an assumption which doesn't always hold, and the theorem is true without making it, but this will help us avoid some infinite behavior. We have just shown that $\theta_a > 0$ – pick some $\theta \in (\theta_a, \theta_+)$. Since θ_a is a critical point of $\theta a - \kappa(\theta)$, convexity of κ tells us that

$$\kappa'(\theta) > \kappa'(\theta_a) = a.$$

We'll now look at the change of measure applied to the X_is, defining

$$\mathbb{P}_\theta\left(\prod_{i=1}^n 1\{X_i \in A_i\}\right) = \mathbb{E}\left[\prod_{i=1}^n \left(1\{X_i \in A_i\}\frac{e^{\theta X_i}}{m(\theta)}\right)\right]$$

for any events A_i. (In particular, the X_is are also iid under $\mathbb{P}_\theta$ because of the product form of this equation.) We may then extend this to measurable functions in the same way as above, and in particular we can write (notice that we are now changing measure on the right-hand side instead of the left)

$$\mathbb{P}(S_n \geq na) = \mathbb{E}_\theta\left[\frac{m(\theta)^n}{e^{\theta S_n}} \cdot 1\{S_n \geq na\}\right].$$

We are trying to prove a lower bound, so it is okay to decrease the right-hand side: picking some $b > \kappa'(\theta)$, it is also true that

$$\mathbb{P}(S_n \geq na) \geq \mathbb{E}_\theta\left[\frac{m(\theta)^n}{e^{\theta S_n}} \cdot 1\{S_n \in [na, nb]\}\right] \geq \mathbb{E}_\theta\left[\frac{m(\theta)^n}{e^{\theta nb}} \cdot 1\{S_n \in [na, nb]\}\right],$$

because $S_n \geq nb$ within this event $\{S_n \in [na, nb]\}$. But here's the magic: under our changed measure, $\mathbb{E}_\theta[X] = \kappa'(\theta) \in (a, b)$, so $\mathbb{P}_\theta(S_n \in [na, nb]) \to 1$ as $n \to \infty$ by the law of large numbers. In other words, doing our change of measure has turned a rare event into a typical event! Thus, we indeed have

$$\boxed{\mathbb{P}(S_n \geq na)} \geq \frac{m(\theta)^n}{e^{\theta nb}}(1 - o(1)) = \boxed{(1 - o(1)) \exp\left[-n(\theta b - \kappa(\theta))\right]}$$

for any $\theta > \theta_a$ and $b > \kappa'(\theta)$. Taking $\theta \downarrow \theta_a$ and $b \downarrow \kappa'(\theta_a) = a$ and then sending $n \to \infty$ gives us the desired lower bound (because θ_a is where the supremum is achieved). □

Something magical is happening here – the convenient exponentiation in the upper bound when we used Markov's inequality was a bit arbitrary, and we also performed a pretty arbitrary-looking change of measure to get the lower bound, but they turn out to yield the same Legendre dual. So in some sense, exponentiating is optimal on both sides of the inequality, and we now have another way to understand the theorem statement: "the most efficient (probable) way to achieve a large deviation $S_n \geq na$ is to have $X_1, \cdots, X_n$ all behave like a collection of samples from our tilted measure $\mathbb{P}_\theta$, choosing θ such that $\mathbb{E}_\theta[X] = a$."

Example 98

In some special cases, we can calculate these large deviations more explicitly, and we'll finish this lecture by doing so in the case where $\mathcal{L}_X$ has finite support of size k. Suppose we have a random variable X which can only take on the values $x_1, \cdots, x_k$, with probabilities $\pi_1, \cdots, \pi_k$ respectively.

Definition 99

The **empirical measure** of the random variables $X_1, \cdots, X_n$ (in general, not just in the case of finite support) is

$$L_n^X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}.$$

(In other words, we sample the X_is and put a weight of $\frac{1}{n}$ at each achieved value.) L_n^X is a probability measure on $\mathbb{R}$, and it is a random variable that is $\sigma(X_1, \cdots, X_n)$-measurable. In our particular case where the support of X is finite, we can express the empirical measure as a vector $L_n^X \in [0, 1]^k$, with jth component equal to the empirical fraction

$$(L_n^X)_j = \frac{1}{n}\{\#1 \leq i \leq n : X_i = x_j\}.$$

Since there are only finitely many possibilities for L_n^X, we can explicitly calculate any particular probability

$$\mathbb{P}(L_n^X = \nu = (\nu_1, \cdots, \nu_k)) = \frac{n!}{(n\nu_1)!(n\nu_2)! \cdots (n(\nu_k)!} \cdot \pi_1^{\nu_1}\pi_2^{\nu_2} \cdots \pi_k^{\nu_k}.$$

(this is the probability that a ν_i fraction of the random variables land on π_i). Expanding the factorials with Stirling's formula and simplifying, we find that

$$\mathbb{P}(L_n^X = \nu) = \exp\left[-n\mathcal{H}(\nu|\pi) + o(n)\right],$$

where

$$\mathcal{H}(\nu|\pi) = \sum_{j=1}^{k} = \nu_j \log \frac{\nu_j}{\pi_j} = D_{\mathrm{KL}}(\nu|\pi)$$

is the **relative entropy** or **Kullback-Leibler divergence** between the two measures (though note that this expression is not symmetric between ν and π.) This is just a combinatorical calculation, but we can now do something with it. Suppose the support of $\mathbb{P}$ contains all k values $x_1, \cdots, x_k$, so that $\mathcal{H}(\nu|\pi)$ is strictly convex in ν. Notice that S_n doesn't depend on the ordering of X_is, and in fact it is a function of L_n^X because

$$S_n = n \sum_{j=1}^{k} x_j L_n^X(x_j) = n\langle x, L_n^X \rangle.$$

For any a such that $\mathbb{E}[X] < a <$ ess sup X, we thus have (summing over all values ν that L_n^X can take)

$$\mathbb{P}(S_n \geq na) = \sum_{\nu} \mathbb{P}(L_n^X = \nu) \cdot 1\{n\langle x, \nu \rangle \geq na\} = \sum_{\nu} 1\{\langle x, \nu \rangle \geq a\} \exp(-n\mathcal{H}(\nu|\pi) + o(n)).$$

But the number of values ν is polynomial in n (by stars and bars), which can be absorbed into the $o(n)$ term in the exponential. In other words, we find that (taking the largest exponential across all νs)

$$\boxed{\mathbb{P}(S_n \geq na) = \exp\left[-n \inf_{\nu}\{\mathcal{H}(\nu|\pi) : \langle x, \nu \rangle \geq a\} + o(n)\right]}.$$

To find the optimal ν, we now have a constraint optimization problem with Lagrangian (remembering that we have the constraint that our probability needs to sum to 1)

$$L = \mathcal{H}(\nu|\pi) + \rho(1 - \sum_j \nu_j) + \theta\left(a - \sum_j x_j \nu_j\right),$$

and taking its derivative gives us the solution

$$\frac{\partial L}{\partial \nu_j} = 1 + \log \nu_j - \log \pi_j - \rho - \theta x_j = 0 \implies \nu_j = \frac{\pi_j e^{\theta x_j}}{C},$$

where we should choose the normalization constant to be $C = m(\theta)$ so probabilities do add to one. But this is the same change of measure $\mathbb{P}_\theta$ as we talked about in the proof of Theorem 97! (This time, it comes from the final optimization problem rather than pulling them out of nowhere.) In other words, when we are trying to achieve this large deviation event $S_n \geq na$, the primary contribution (the slowest decaying exponential) comes from $L_n^X \sim \nu = \mathbb{P}_\theta$. And as an exercise, we can also show that

$$\mathbb{P}(S_n \geq na) = \exp\left[-n \sup_{\theta : \mathbb{E}_\theta[X] \geq a} \mathcal{H}(\mathbb{P}_\theta|\pi) + o(n)\right],$$

with optimal θ being achieved at equality $\mathbb{E}_\theta[X] = a$.

12 October 16, 2019

(Our first midterm is on Monday, so we should make sure to remember to show up to class.) Today, we'll cover some loose ends that have been mentioned but not in detail, and we'll also introduce the Fourier transform (which is a topic for after the exam). Part of this is fair game for the exam, but this will be communicated clearly. We'll start with a sample problem for the midterm:

Problem 100

Prove that the law of a real-valued random variable X is uniquely determined by its cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$.

This is the level of formality we should expect: we should understand, for example, that X is a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and its law is defined as $\mathcal{L}_X = X_{\#}\mathbb{P} = \mu$, which is a probability measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Solution. The cdf $F(x) = \mathbb{P}(X \leq x)$ is also the measure $\mu((-\infty, x])$, so knowing μ also determines F. For the converse, suppose there are two measures μ, ν on $\mathbb{R}$ that both correspond to the same cumulative distribution function, so $F(x) = \mu((-\infty, x]) = \nu((-\infty, x])$ for all x. This means that

$$\mu((a, b]) = F(b) - F(a) = \nu((a, b])$$

for all half-open intervals. Let $\mathcal{F}$ be the set of Borel sets such that $\mu(B) = \nu(B)$ – we need to show that $\mathcal{F} = \mathcal{B}_{\mathbb{R}}$, which we can do with a π-λ argument. Specifically, $\mathcal{F}$ contains the set of half-open intervals, which is a π-system, and we can check that $\mathcal{F}$ is a λ-system as well (by properties of a measure). So by the π-λ theorem, $\mathcal{F}$ contains the σ-algebra generated by the set of half-open intervals, which is $\mathcal{B}_{\mathbb{R}}$. Therefore $\mu = \nu$ and the measure μ is uniquely determined by F as well. □

Remark 101. *We won't be allowed to bring cheat sheets, but some useful things will be written on the cover page. (However, it is reasonable for us to remember, for instance, the full weak law of large numbers.) If a result follows directly from a theorem in class, we can cite the theorem and just check that all the conditions hold. But if we're instead being told to reproduce a very similar proof, that will be written out.*

For today's class, we'll start by discussing a probabilistic concept which we haven't talked about too formally:

Definition 102

Let X and Y be independent real-valued random variables with laws $\mathcal{L}_X = \mu$ and $\mathcal{L}_Y = \nu$, so that $\mathcal{L}_{(X,Y)} = \mu \otimes \nu$. Let their cdfs be $F(x) = \mathbb{P}(X \geq x)$ and $G(y) = \mathbb{P}(Y \geq y)$. Because μ and ν are in one-to-one correspondence with F and G respectively (by Problem 100), we may define the **convolution** $\mu * \nu = \nu * \mu$ to be the law of $Z = X + Y$. We denote the corresponding cdf by $F * G = G * F$.

Note that μ, ν can be discrete, continuous, or anything in between, and this definition is still valid. To get a more explicit expression for $F * G$, notice that

$$(F * G)(z) = \mathbb{P}(X + Y \leq z) = \int \int 1\{x + y \leq z\} d\mu(x) d\nu(y),$$

where we've implicitly used the change of variables formula to write the integral in terms of μ and ν and Tonelli's theorem to break up the double integral. Then the inner integral is just the probability that $x \leq z - y$, so we have

$$\boxed{(F * G)(z) = \int F(z - y) d\nu(y) = \int F(z - y) dG(y)},$$

where the last equality is just saying that ν and G carry the same information. (And by symmetry, this is also equal to $\int G(z - x) dF(x)$.) We can say something even nicer in the case where μ and ν have **densities** – in particular, $\mu * \nu$ also has a density. To see this, suppose μ and ν correspond to the density functions f and g. By definition, that means that for all Borel sets B, $\mu(B) = \int 1_B(x) f(x) dx$ (where dx is the standard Lebesgue measure), where f is a

nonnegative measurable function on $\mathbb{R}, \mathcal{B}_{\mathbb{R}}$. In particular, $F(x) = \int_{-\infty}^{x} f(t)dt$ (by plugging in the set $(-\infty, x]$), and for any integrable random variable $h(X)$, we have $\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(t)f(t)dt$. So plugging in $dG(y) = g(y)dy$, we find that

$$(F*G)(z) = \int_{-\infty}^{\infty} F(z-y)dG(y) = \int_{-\infty}^{\infty} F(z-y)g(y)dy = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f(t-y)dt\ g(y)dy = \int_{-\infty}^{z}\int_{-\infty}^{\infty} f(t-y)g(y)dydt$$

by Tonelli's theorem. So we've now written the cdf $F * G$ as an integral $\int_{-\infty}^{\infty}(f*g)(t)dt$, and therefore $\mu * \nu$ has density

$$\boxed{(f*g)(z) = \int_{-\infty}^{\infty} f(z-y)g(y)dy = \int_{-\infty}^{\infty} g(z-x)f(x)dx}.$$

We'll now move on to our next topic, the **Fourier transform**. (This topic won't be on the first exam, and not everything mentioned today is necessary for us to know, but it may be useful intuition.)

Definition 103

The **Fourier transform** or **characteristic function** of a probability measure μ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a function $\phi_\mu = \hat{\mu} : \mathbb{R} \to \mathbb{C}$ defined by

$$\phi_\mu(t) = \int e^{itx} d\mu(x).$$

In particular, if X is a random variable with law μ, its characteristic function is $\phi_\mu(t) = \mathbb{E}\left[e^{itX}\right]$.

Because X is real-valued, e^{itX} is always on the unit circle in $\mathbb{C}$, so $\phi_\mu(t)$ is inside the unit ball in $\mathbb{C}$. Thus, $\phi_\mu(t)$ is well-defined for all $t \in \mathbb{R}$, and it's continuous in t (by the bounded convergence theorem). The reason this definition is useful (and in particular good for proving something like the Central Limit Theorem) is that **the Fourier transform turns convolution into multiplication**. Specifically, if X is a random variable with law μ and Y is a random variable with law ν, and the two random variables are independent, then we have the useful identity

$$\boxed{\phi_{\mu*\nu}(t)} = \mathbb{E}\left[e^{it(X+Y)}\right] = \mathbb{E}\left[e^{itX}\right]\mathbb{E}\left[e^{itY}\right] = \boxed{\phi_\mu(t)\phi_\nu(t)}.$$

Example 104

We will indeed eventually study the Fourier transform on $\mathbb{R}$, and we'll see that it captures a lot of properties of the measure μ. But today, we'll look at the Fourier transform on a finite space.

We'll look at the space $\Omega = \mathbb{Z}/n\mathbb{Z}$ (the integers mod n), which will keep our story simple. Let V be the set of functions $f : \Omega \to \mathbb{C}$, which is a finite-dimensional complex vector space isomorphic to $\mathbb{C}^{|\Omega|}$. If we view these functions as vectors, then V has a Hermitian inner product given by

$$\langle f, g\rangle = \sum_{x=0}^{n-1} f(x)\overline{g(x)} = g^* f,$$

where g^* is the conjugate transpose of g. Then for any (nonzero) $z \in \Omega$, we can consider the translation operator T_z acting on functions via

$$(T_z f)(x) = f(x-z).$$

T_z is then essentially sending vectors to vectors, so we can think of it as a matrix – specifically, it is a **circulant** matrix with shift depending on z. We can then calculate its eigenvectors and eigenvalues: f is an eigenvector of T_z with

eigenvalue λ if and only if for all $x \in \Omega$,

$$\lambda f(x) = T_z f(x) = f(x-z) \implies \lambda^k f(x) = f(x-kz) = f(x) \text{ if } kz \equiv 0 \bmod n.$$

In particular (plugging in $k = n$), this means that $\lambda^n = 1$, so any eigenvalue λ must be an nth root of unity. Now if n is prime, then T_z has n distinct eigenvalues, which are exactly those nth roots of unity, and we can check that the eigenvectors are of the form

$$\chi_\ell(x) = \frac{1}{\sqrt{n}} \exp\left(\frac{2\pi i \ell x}{n}\right)$$

with corresponding eigenvalues $\lambda_{z,\ell} = \exp\left(-\frac{2\pi i \ell z}{n}\right)$. On the other hand, if n is not prime, we no longer have a unique eigenbasis. Luckily, it turns out that $\chi_0, \cdots, \chi_{n-1}$ still do form an eigenbasis of T_z for all $z \in \Omega$, and χ_ℓ is still an eigenvector of T_z with that same eigenvalue $\lambda_{z,\ell} = \exp\left(-\frac{2\pi i \ell z}{n}\right)$. In other words, because our finite space has a periodic structure, the translation operators have a particularly nice form. We can also find that the scalar products between two vectors in our eigenbasis is

$$\langle \chi_k, \chi_\ell \rangle = (\chi_\ell)^* \chi_k = \sum_x \chi_k(x)\overline{\chi_\ell(x)} = \sum_{x=0}^{n-1} \frac{1}{n} \exp\left(\frac{2\pi i x(k-\ell)}{n}\right).$$

This is 1 when $k = \ell$ and 0 otherwise, which means that with respect to the Hermitian inner product, the χ_is form an orthonormal basis. Thus, the matrix

$$U = \begin{bmatrix} | & \cdots & | \\ \chi_0 & \cdots & \chi_{n-1} \\ | & \cdots & | \end{bmatrix}$$

is **unitary** (meaning that $U^*U = UU^* = I$). We can call $(\chi_0, \cdots, \chi_{n-1})$ the **Fourier basis** for V, and the **Fourier transform** is just the change-of-basis operation sending a function f to $\hat{f} = U^* f$ (which basically tells us the coordinates of f in the Fourier basis). In other words,

$$f = UU^* f = U\hat{f} = \sum_\ell \hat{f}(\ell)\chi_\ell$$

is how we write f as a linear combination of the eigenvectors χ_ℓ, and the coefficients $\hat{f}(\ell)$ are given explicitly by

$$\hat{f}(\ell) = (U^* f)_\ell = \langle f, \chi_\ell \rangle = \sum_{x=0}^{n-1} f(x)\overline{\chi_\ell(x)} = \sum_{x=0}^{n-1} \frac{1}{\sqrt{n}} f(x) \exp\left(-\frac{2\pi i \ell x}{n}\right).$$

Ignoring the constants, this expression should look very similar to how we multiply f by some complex exponential in the regular Fourier transform in Definition 103. So the whole point is that **the Fourier transform is a change-of-basis operation which diagonalizes the translation operators**, and the exponentials in the definitionare motivated by the discrete case in which they pop out of the eigenvalue calculation.

But we can study **convolution** in this finite space as well, because we can think of the expression for $f * g$ in terms of translations of one of the two functions:

$$(f * g)(x) = \sum_z g(z) f(x-z) = \sum_z g(z)(T_z f)(x).$$

Thus, we can think of $f * g$ as applying an operator C_g, which is a linear combination of translation operators, to f:

$$C_g = \sum_z g(z) T_z.$$

But since the Fourier basis diagonalizes all of the T_zs simultaneously, we should expect that the Fourier basis is also good for C_g. Specifically, we know that $T_z = U\Lambda_z U^*$ (where Λ_z is the diagonal matrix with entries $\lambda_{z,0}, \cdots, \lambda_{z,n-1}$) for each z, or equivalently (expanding out the matrix multiplication)

$$T_z = \sum_{\ell=0}^{n-1} \lambda_{z,\ell} \chi_\ell \chi_\ell^*.$$

Plugging this into our expression for C_g, we thus find that

$$C_g = \sum_z g(z) \sum_\ell \lambda_{z,\ell} \chi_\ell \chi_\ell^* = \sum_\ell \left(\sum_z \lambda_{z,\ell} g(z) \right) \chi_\ell \chi_\ell^*.$$

Now because $\lambda_{z,\ell} = \exp\left(-\frac{2\pi i \ell z}{n}\right)$ is the zth entry in the ℓth row of the matrix U^*, the inner sum can also be thought of as the ℓth coordinate of the vector $U^* g$. So we can write C_g in matrix form as the product of three terms,

$$C_g = \begin{bmatrix} | & \cdots & | \\ \chi_0 & \cdots & \chi_{n-1} \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} (U^*g)_0 & & & \\ & (U^*g)_1 & & \\ & & \ddots & \\ & & & (U^*g)_{n-1} \end{bmatrix} \begin{bmatrix} — & \chi_0^* & — \\ \vdots & \vdots & \vdots \\ — & \chi_{n-1}^* & — \end{bmatrix} = U \operatorname{diag}(U^* g)\, U^*.$$

But this now allows us to understand how convolution and the Fourier transform interact: we have

$$\widehat{f * g} = \widehat{C_g f} = U^* C_g f = U^* U \operatorname{diag}(U^* g)\, U^* f = \operatorname{diag}(U^* g)\, U^* f.$$

But now $U^* g$ is the Fourier transform of $\hat{g}$, and $U^* f$ is the Fourier transform of f. Thus this last expression is the entry-wise product of $\hat{f}$ and $\hat{g}$, and we do indeed see that $\widehat{f * g}(x) = \hat{f}(x)\hat{g}(x)$, analogous to the identity $\phi_{\mu * \nu}(t) = \phi_\mu(t)\phi_\nu(t)$ that we derived earlier! So the Fourier transform can also be thought of as the unique object behaving in this particularly nice way under convolution, and as we mentioned before, it will play a role in some of the results we will prove later in this class.

We'll finish this lecture with one more sample problem:

Problem 105

Suppose $X_1, X_2, \cdots$ are random variables that are **Cauchy in probability**, meaning that that for all $\varepsilon > 0$, $\mathbb{P}(|X_m - X_n| \geq \varepsilon)$ goes to 0 as $m, n \to \infty$. Prove there exists a random variable X such that $X_n \to X$ in probability.

Solution. We proved in our homework that if a sequence of random variables converges in probability, it also converges almost surely along some subsequence. Motivated by this, we'll start by constructing that subsequence. Because the X_is are Cauchy in probability, there is some sequence n_k (which we can choose to go to infinity) such that for all $m, n \geq n_k$,

$$\mathbb{P}\left(|X_m - X_n| \geq \frac{1}{2^k}\right) \leq \frac{1}{2^k}.$$

Now define the event $A_k = \{\omega : |X_{n_{k+1}}(\omega) - X_{n_k}(\omega)| \geq \frac{1}{2^k}\}$. By Borel-Cantelli, the probability that infinitely many of $A_1, A_2, \cdots$ occur is 0, so on $\Omega \setminus \{A_k \text{ i.o.}\}$ (an event of probability 1), these events eventually stop occurring and thus the X_{n_k}s converge (because their values are Cauchy) to some X. In other words, the subsequence X_{n_k} converges to some random variable X almost surely.

It remains to show that the full sequence converges to X. By the triangle inequality and a union bound,

$$\begin{aligned}\mathbb{P}(|X_n - X| \geq \varepsilon) &\leq \mathbb{P}(|X_n - X_{n_k}| + |X_{n_k} - X| \geq \varepsilon) \\ &\leq \mathbb{P}\left(|X_n - X_{n_k}| \geq \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|X_{n_k} - X| \geq \frac{\varepsilon}{2}\right).\end{aligned}$$

Taking $k \to \infty$, the second term goes to 0 because X_{n_k} converges to X almost surely. And if we also send $n \to \infty$ (so that n and n_k both go to infinity), the first term also goes to zero because the X_is are Cauchy in probability. Thus $\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0$ as $n \to \infty$, which is the desired result. □

In preparation for the midterm, Professor Sun's office hours will be Friday 11-12, and the TAs will have office hours as well.

s

13 October 23, 2019

Last time, we gave some motivation for the Fourier transform of a random variable by considering the discrete case $\Omega = \mathbb{Z}/n\mathbb{Z}$. Letting V be the set of functions from $\Omega \to \mathbb{C}$, we have a finite-dimensional vector space isomorphic to $\mathbb{C}^\Omega$ with Hermitian inner product $\langle f, g\rangle = g^* f = \sum_{x\in\Omega} f(x)\overline{g(x)}$. Under this inner product, we found an orthonormal basis $(\chi_0, \cdots, \chi_{n-1})$ for V, with $\chi_\ell(x) = \frac{1}{\sqrt{n}} \exp\left(\frac{2\pi i \ell x}{n}\right)$. In particular, we found that the matrix $U = \begin{bmatrix}\chi_0 & \chi_1 & \cdots & \chi_{n-1}\end{bmatrix} \in \mathbb{C}^{n\times n}$ is **unitary** and gives rise to Fourier transform as a change of basis-operation

$$\hat{f} = U^{-1} f = U^* f.$$

In other words, the coefficient $\hat{f}(\ell)$ of f in the Fourier basis is $\langle f, \chi_\ell\rangle$. It was particularly nice that **Fourier inversion** could be easily performed (since $f = U\hat{f}$) and that $\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle$ because U is unitary. Today's class will now cover the Fourier transform for functions on $\mathbb{R}$, and we'll see some of these properties come up again.

We'll start by fixing the normalization. Recall that if μ is a probability measure on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$, we defined the **Fourier transform** or **characteristic function** of μ as

$$\phi_\mu(t) = \int e^{itx} d\mu(x) = \mathbb{E}\left[e^{itX}\right],$$

for a random variable $X \sim \mu$. We discussed some basic properties of this function last time, and we'll formalize them now:

Proposition 106

For any probability measure μ on $\mathbb{R}$, ϕ_μ is a uniformly continuous mapping from $\mathbb{R}$ to the closed unit disk $\overline{D} = \{z \in \mathbb{C} : |z| \leq 1\}$.

Proof. As mentioned previously, ϕ_μ indeed takes values on $\overline{D}$ because e^{itx} is always on the unit circle in $\mathbb{C}$. To prove **uniform** continuity, we use that

$$|\phi_\mu(t+h) - \phi_m u(t)| = \left|\mathbb{E}[e^{i(t+h)}X - e^{itX}]\right|.$$

Applying Jensen's inequality in the form $|\mathbb{E}[Y]| \leq \mathbb{E}[|Y|]$ allows us to bring the absolute value into the expectation and find that this is

$$\leq \mathbb{E}\left[\left|e^{itX}(e^{ihX} - 1)\right|\right] = \mathbb{E}\left[\left|e^{ihX} - 1\right|\right].$$

But the right-hand side does not depend on the point $t \in \mathbb{R}$ we choose, and it tends to zero as $h \to 0$ by applying the bounded convergence theorem to any sequence $h_n \to 0$ (and remembering that the integral over a probability space is indeed over a set of finite support). □

In the special case where μ has a density function $f(x)$, we can write the Fourier transform as an integral

$$\phi_\mu(t) = \int e^{itx} f(x)dx = \hat{f}(t),$$

and we call this the **L^1 Fourier transform** (because f integrates to 1, so it's in L^1). There is an important result from Fourier analysis that we should keep in mind:

Theorem 107 (Planchard/Parseval)
Suppose we have two functions $f, h \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Then $\hat{f}, \hat{h} \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$, and we have

$$\langle \hat{f}, \hat{h} \rangle = 2\pi \langle f, h \rangle.$$

Thus, the mapping $U : L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \to L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$ defined by $U(f) = \frac{\hat{f}}{\sqrt{2\pi}}$ has a unique continuous extension to a unitary map $L^2(\mathbb{R}) \to L^2(\mathbb{R})$.

The idea is that it is nice to be able to have U map a space to itself, but the integral $\int e^{itx} f(x)dx$ may not be defined if f is not in L^1. But instead, we can approximate f by functions in the space $L^1 \cap L^2$, and it turns out that the limit will make sense and live in L^2. And having a unitary mapping means that we also have a simple inverse – for any function $h \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, we can define

$$U^*(h(x)) = \frac{1}{\sqrt{2\pi}} \int e^{-itx} h(t)dt = \frac{\hat{h}(-x)}{\sqrt{2\pi}}$$

(which is like $U(f) = \frac{\hat{f}}{\sqrt{2\pi}}$ but with an additional negative sign). The logic from Theorem 107 applies here as well, so this map also extends to a map $U^* : L^2(\mathbb{R}) \to L^2(\mathbb{R})$. This gives us an important result: the **Fourier inversion theorem** for functions tells us that

$$f(x) = U^* U f = U^* \left(\frac{\hat{f}}{\sqrt{2\pi}} \right) = \frac{1}{2\pi} \hat{\hat{f}}(-x),$$

meaning that up to a scaling factor and a change of sign, the Fourier transform is an involution. Unfortunately, the Fourier transform for general probability measures (instead of functions) is not quite as nice:

Theorem 108 (Fourier inversion formula for probability measures on $\mathbb{R}$)
Suppose μ is a probability measure on $\mathbb{R}$ and $\phi_\mu(t)$ is its characteristic function. Then we have

$$\int_{-T}^{T} \phi_\mu(t) \left(\frac{e^{-ita} - e^{-itb}}{2\pi it} \right) dt \xrightarrow{T \to \infty} \tilde{\mu}(a, b) = \mu((a, b)) + \frac{\mu(\{a, b\})}{2}.$$

The point of this result is that we use μ to define ϕ_μ, but we can also use ϕ_μ to determine μ (it's left as an exercise that even though $\tilde{\mu}$ and μ are not exactly the same, we can use $\tilde{\mu}$ to uniquely determine μ). By the way, we do need to integrate from $-T$ to T to ensure the integral actually exists, and it's part of the content of the theorem that those integrals converge as $T \to \infty$:

Example 109

Consider the random variable X which takes on the values ± 1 with probability $\frac{1}{2}$. Then

$$\phi_\mu(t) = \frac{e^{it} - e^{-it}}{2} = \cos t.$$

This function does not decay as $t \to \infty$, so plugging it into the integral on the left-hand side of Theorem 108 would not work.

Before we go into the proof, we'll discuss the intuitive explanation for this result. We're basically being given ϕ_μ, and we want to recover the distribution by finding, for example, $\mu((a,b))$ for some $a < b$. In other words, we wish to compute $\mu((a,b)) = \int h(x) d\mu(x)$, where h is the indicator function $1_{(a,b)}$. Because h has bounded support, we can compete the Fourier transform of this indicator function directly to be

$$\hat{h}(t) = \int e^{itx} h(x) dx = \int_a^b e^{itx} dx = \frac{e^{itb} - e^{ita}}{it}.$$

So if we think of $\mu(a,b) = \int h \, d\mu = \langle \mu, h \rangle$ as some kind of inner product (whatever that means, because μ is a measure rather than a function), then Theorem 107 should yield

$$\mu((a,b)) = \frac{1}{2\pi} \langle \hat{\mu}, \hat{h} \rangle = \frac{1}{2\pi} \langle \phi_\mu, \hat{h} \rangle = \int_{-\infty}^{\infty} \phi(t) \frac{e^{-ita} - e^{-itb}}{2\pi i t} dt.$$

In reality, this integral may not even be defined, but we at least see why the integrand on the left-hand side takes the form that it does. To understand this further, we can approximate ϕ by **truncating in the Fourier domain** by integrating the last expression from $-T$ to T only:

$$\int_{-T}^{T} \phi(t) \frac{e^{-ita} - e^{-itb}}{2\pi i t} dt = \int_{-\infty}^{\infty} \phi(t) \overline{\hat{h}(t)} \frac{k_T(t)}{2\pi} dt,$$

where k_T is the truncation function $1\{|t| \le T\}$. We can rewrite this as an inner product $\frac{1}{2\pi} \langle \phi, \hat{h} k_T \rangle = \langle \mu, h_T \rangle$, where h_T is the function satisfying $\widehat{h_T} = \hat{h} k_T$. We can solve for h_T explicitly, but we don't need to – the idea is that h is a not-so-smooth function (the indicator of (a,b)) mapped into the Fourier domain, so removing the high-frequency domain by multiplying by k_T should mean that h_T is essentially a "smooth version" of h. So if that smoothing goes away as $T \to \infty$, it makes sense that we have

$$\lim_{T \to \infty} h_T(x) = \tilde{h}(x) = 1\{(a,b)\} + \frac{1}{2} \cdot 1\{\{a,b\}\},$$

because h is 0 on one side of a (and also b) and 1 on the other, so the h_Ts will have half of each contribution. This completes the intuition for why we get only half of the measure on the endpoints, and we're now almost ready to dive into the proof. We'll first figure out what h_T should be so that it satisfies $\widehat{h_T} = \hat{h} k_T$:

Proposition 110

The function $\text{sinc}(t) = \frac{\sin t}{t}$ is not in $L^1(\mathbb{R})$, because it decays like $\frac{1}{t}$. However, the truncated version of that function

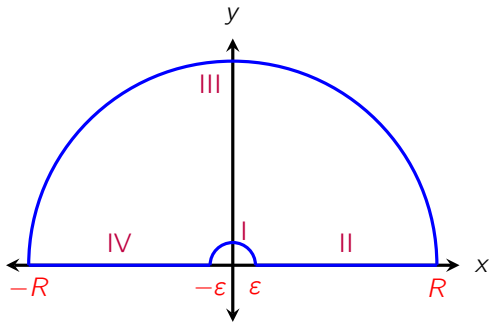
$$S(T) = \int_{-T}^{T} \text{sinc}(t)$$

converges to π as $T \to \infty$.

Proof. There are various ways to do this, but we'll calculate using complex analysis. We can first rewrite

$$S(T) = \int_{-T}^{T} \frac{e^{it} - e^{-it}}{2it} dt = \int_{-T}^{T} \frac{e^{it}}{it} dt$$

by symmetry. Now the function $h(z) = \frac{e^{iz}}{iz}$ is holomorphic on $\mathbb{C} \setminus \{0\}$, so if we integrate around any closed contour γ not containing or passing through the origin, then $\oint_\gamma h(z)dz = 0$. We will integrate over the following indented semicircle once counterclockwise and take $R \to \infty, \varepsilon \to 0$:



The integral of $\text{sinc}(t)$ from $-\varepsilon$ to ε vanishes as $\varepsilon \to 0$ (because $\frac{\sin t}{t} \to 1$ as $t \to 0$), so the integral we want to compute is the contribution along parts II and IV of the contour. It suffices to show that (1) as $\varepsilon \to 0$, the integral along I goes to $-\pi$, and (2) as $R \to \infty$, the integral along III goes to 0. (2) can be done with some careful bounding which we'll skip (by parameterizing the semicircle and then bounding the resulting integral $\int_0^\pi e^{-R\sin\theta}$ by breaking it up into a region with small length and a region with small integrand). Meanwhile, (1) can be done by parametrizing z as $\varepsilon e^{i\theta}$ for θ from π to 0, so that the integral along region I is

$$\int_{\gamma_\varepsilon} h(z)dz = \int_\pi^0 \frac{e^{i\varepsilon e^{i\theta}}}{i\varepsilon e^{i\theta}} i\varepsilon e^{i\theta} d\theta = \int_\pi^0 e^{i\varepsilon e^{i\theta}} d\theta.$$

But the numerator converges to 1, so this integral converges to $\int_\pi^0 1 d\theta = -\pi$ as $\varepsilon \to 0$. So because the total contour integral is 0, the integral we're trying to find (along II and IV) is π, as desired. □

We will use this function $S(T)$ to find h_T by applying the Fourier inversion formula for functions. Inverting $\hat{h}k_T$, we have

$$h_T(x) = \frac{1}{2\pi} \int_{-T}^{T} e^{-itx} \hat{h}(t) dt = \frac{1}{2\pi} \int_{-T}^{T} e^{-itx} \cdot \left(\frac{e^{itb} - e^{ita}}{it} \right) dt.$$

If we average between the values at x and $-x$ for each point in the integrand, the integral can be rewritten as

$$h_T(x) = \int_{-T}^{T} \left[\frac{\sin(t(b-x))}{2\pi t} - \frac{\sin(t(a-x))}{2\pi t} \right] dt,$$

where we use that $\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}$. The two terms can now be integrated separately – the integral of $\frac{\sin(t(b-x))}{2\pi t}$ is zero if $b - x = 0$, and in all other cases we can make a u-substitution (flipping the limits of integration if $b - x, a - x$ are negative) to write this in terms of the function S from above. We thus have

$$h_T(x) = \frac{\text{sgn}(b-x)S(|b-x| \cdot T) - \text{sgn}(a-x)S(|a-x| \cdot T)}{2\pi},$$

where we define $\text{sgn}(0) = 0$. Now if $b - x, a - x$ are both positive or both negative, both terms converge to the same value (because S converges to π as its argument goes to infinity), so this is only nonzero when $x \in [a, b]$. If x is strictly between a and b, we have $\frac{\pi - (-\pi)}{2\pi} = 1$, and otherwise (when $x = a$ or $x = b$) it goes to $\frac{\pi}{2\pi} = \frac{1}{2}$. Putting this all together, we indeed find that h_T converges pointwise to $\tilde{h}(x) = 1_{(a,b)} + \frac{1}{2} \cdot 1_{\{\{a,b\}\}}$, as desired.

Remark 111. *Another way we can see that h_T is smooth is that Fourier transforms take convolution to multiplication, so $\hat{h_T} = \hat{h}k_T \implies h_T = h * S_T$, where $\hat{S_T} = k_T$. This function S_T is "wavy" and becomes more oscillatory as T gets larger, but it is smooth, and thus the convolution h_T is also smooth.*

With this, we're finally ready to explain the Fourier inversion theorem:

Proof of Theorem 108. We start by manipulating the left-hand side in terms of the functions we've been describing: we have

$$I_T = \int_{-T}^{T} \phi_\mu(t) \left(\frac{e^{-ita} - e^{-itb}}{2\pi it} \right) dt = \frac{1}{2\pi} \int_{-T}^{T} \phi_\mu(t) \overline{\hat{h}(t)} dt = \frac{1}{2\pi} \int_{-T}^{T} \left(\int_{-\infty}^{\infty} e^{itx} d\mu(x) \right) \overline{\hat{h}(t)} dt.$$

Using that $e^{itx} = \overline{e^{-itx}}$ and swapping the order of integration (which is allowed by Fubini's theorem because the inner integral is bounded and the outer integral is a continuous function integrated over a finite domain),

$$I_T = \int_{-\infty}^{\infty} \overline{\int_{-T}^{T} \frac{e^{-itx} \hat{h}(t)}{2\pi} dt} \, \mu(dx).$$

Applying the Fourier inversion formula for functions to the inner integral, we can simplify to

$$I_T = \int \overline{h_T(x)} d\mu(x) = \int h_T(x) d\mu(x)$$

(where the last step follows from h_T being real by our previous computation). All of this holds for a fixed T, and to finish, we need to show that $I_T = \int h_T(x) d\mu(x)$ converges to $\int \tilde{h}(x) d\mu(x)$ as $T \to \infty$ (where $\tilde{h} = 1\{(a,b)\} + \frac{1}{2} \cdot 1\{\{a,b\}\}$ as defined above). We already know that $h_T \to \tilde{h}$ converges pointwise from earlier work, and we know that h_T is of the form $\frac{1}{2\pi}(\pm S(\text{something}) \pm S(\text{something else}))$. But since S is bounded for both small and large values of T, it must be uniformly bounded by some constant overall. Thus, h_T is bounded uniformly as well ($\sup_T ||h_T||_\infty$ is finite), so the bounded convergence theorem applies. This means that the left-hand side I_T does indeed converge to $\int \tilde{h}(x) d\mu(x) = \mu((a,b)) + \frac{1}{2}\mu(\{a,b\})$, as desired. □

Example 112

For a concrete example of a Fourier transform, we'll compute the characteristic function of a Gaussian with density

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

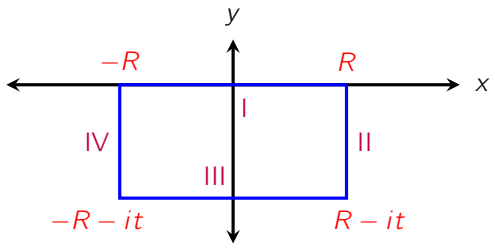
Since we have an explicit density function, we can write

$$\phi(t) = \hat{g}(t) = \int_{-\infty}^{\infty} e^{itx} g(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-it)^2}{2} \right) e^{-t^2/2} dx,$$

which we can think of as an integral along a line in the complex plane as

$$\phi(t) = e^{-t^2/2} \int_{x \in \mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-it)^2 \right) dx = e^{-t^2/2} \int_{\mathbb{R}-it} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

We'll now again appeal to complex analysis by integrating this normal density $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ once counterclockwise around a rectangle (as shown below) as $R \to \infty$:



Since $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is holomorphic everywhere, the integral around this closed loop is 0, and we can check that the integrals along II and IV go to zero as $R \to \infty$. Thus, the integrals along the real line $\mathbb{R}$ and the shifted real line $\mathbb{R} - it$ are the same, and they're both equal to 1 because we're integrating a probability density. Plugging this in, we find that $\phi(t) = e^{-t^2/2}$. This is an important example because (ignoring factors of 2π) **the Fourier transform of a Gaussian is a Gaussian** – in other words, the Gaussian is a sort of eigenvector for the Fourier transform.

To finish today's lecture, we'll do a weak form of the L^2 isometry (basically proving a simple case of Theorem 107):

Proposition 113

Suppose $f, h \in L^1 \cap L^2 \cap L^\infty$ are continuous functions. Then $\langle \hat{f}, \hat{h} \rangle = 2\pi \langle f, h \rangle$.

Proof. We will assume the part of Theorem 107 that if $f \in L^1 \cap L^2$, then $\hat{f} \in L^2 \cap L^\infty$ (this comes from Fourier analysis). That means that $\hat{f}$ and $\hat{h}$ are in L^2, so by Cauchy-Schwarz $\hat{f}\overline{\hat{h}}$ is integrable. By the dominated convergence theorem, we can therefore write

$$\langle \hat{f}, \hat{h} \rangle = \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \hat{f}(t) \overline{\hat{h}(t)} \exp\left(-\frac{\varepsilon t^2}{2}\right) dt.$$

Expanding out the definitions of the Fourier transforms, we have

$$\langle \hat{f}, \hat{h} \rangle = \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{itx} f(x) dx \right) \left(\int_{-\infty}^{\infty} e^{-ity} \overline{h(y)} dy \right) \exp\left(-\frac{\varepsilon t^2}{2}\right) dt.$$

By Fubini's theorem, we can now swap the order of integration to get

$$\begin{aligned} \langle \hat{f}, \hat{h} \rangle &= \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) \overline{h(y)} \int_{-\infty}^{\infty} e^{it(x-y)} \exp\left(-\frac{\varepsilon t^2}{2}\right) dt dx dy \\ &= \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) \overline{h(y)} \sqrt{\frac{2\pi}{\varepsilon}} \int_{-\infty}^{\infty} e^{it(x-y)} \frac{1}{\sqrt{2\pi/\varepsilon}} \exp\left(-\frac{\varepsilon t^2}{2}\right) dt dx dy. \end{aligned}$$

Since $\frac{1}{\sqrt{2\pi/\varepsilon}} \exp\left(-\frac{\varepsilon t^2}{2}\right)$ is the density of $\frac{Z}{\sqrt{\varepsilon}}$, where Z is standard Gaussian, the blue inner integral is $\mathbb{E}[e^{i(x-y)\cdot Z/\sqrt{\varepsilon}}] = e^{-(x-y)^2/(2\varepsilon)}$. This leaves us with

$$\langle \hat{f}, \hat{h} \rangle = \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) \overline{h(y)} \sqrt{\frac{2\pi}{\varepsilon}} e^{(x-y)^2/(2\varepsilon)} dx dy.$$

(So we start off with a "spread out Gaussian" in the Fourier space, which has now become a very "peaked" Gaussian in regular space.) Taking the limit $\varepsilon \to 0$, the $e^{(x-y)^2/(2\varepsilon)}$ will shift all of the weight of integral to the region around $x = y$. Thus by continuity of f and h, we have

$$\langle \hat{f}, \hat{h} \rangle = 2\pi \int f(x) \overline{h(x)} dx = 2\pi \langle f, h \rangle,$$

as desired. □

14 October 28, 2019

(Our midterm exams will be brought to class on Wednesday.) Last time, we discussed the Fourier inversion theorem for probability measures on $\mathbb{R}$, which lets us recover the measure μ for a random variable X given its characteristic function. (One key example was that the characteristic function for the standard Gaussian measure is $\phi(t) = \exp\left(-\frac{t^2}{2}\right)$, which will come up again.) The main goal of today will be to prove the **central limit theorem** for iid sequences. Basically, we'll show that if X, X_i are iid with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$, then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0,1)$ for some sense of convergence ("in distribution") which we'll now formally define.

Definition 114

Let S be a **metric space** with Borel sigma-algebra $\mathcal{B}$ generated by the open sets in the **metric topology**, and let μ_n, μ be probability measures on $(S, \mathcal{B})$. We say that μ_n **converges weakly** to μ (written as $\mu_n \implies \mu$), if $\int f d\mu_n \to \int f d\mu$ for all bounded continuous functions $f : S \to \mathbb{R}$. If X_n, X are random variables, then we say that $X_n \xrightarrow{d} X$ **converges in distribution** if $\mathcal{L}_{X_n} \implies \mathcal{L}_X$.

To understand why we require f to be continuous, consider the (deterministic) random variables $X_n = \frac{1}{n}$. Then X_n converge to $X = 0$ almost surely (so it also makes sense to have them converge in distribution), but the function $f(x) = 1\{x > 0\}$ does not satisfy $f(X_n) \to f(X)$.

Fact 115

Every probability measure on a metric space $(S, \mathcal{B})$ is **regular**, which means that for any $A \in \mathcal{B}$, we can approximate $\mu(A)$ from above and below:

$$\mu(A) = \sup\{\mu(F) : F \subseteq A, F \text{ closed}\} = \inf\{\mu(G) : G \supseteq A, G \text{ open}\}.$$

The proof in general is similar to the one we did on our homework for $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. In particular, this implies that the measure μ is determined by $\{\mu(F) : F \text{ closed}\}$.

We can thus make the following claim:

Lemma 116

Suppose we have two measures μ and ν such that $\int f d\mu = \int f d\nu$ for all bounded continuous functions $f : S \to \mathbb{R}$. Then $\mu = \nu$.

Proof. For any closed set $F \subset S$, we can approximate its indicator function with a continuous function, namely

$$f(x) = \max\left(0, 1 - \frac{\text{dist}(x, F)}{\varepsilon}\right).$$

In other words, $f(x) = 1$ if $x \in F$, $f(x) = 0$ if x is more than ε away from F, and we linearly interpolate in between. If we let F^ε denote the ε-neighborhood of F, we then have that for all $\varepsilon > 0$,

$$1_F \le f \le 1_{F^\varepsilon} \implies \boxed{\mu(F)} \le \int f d\mu = \int f d\nu \le \boxed{\nu(F^\varepsilon)},$$

where the middle equality holds by assumption because f is a bounded continuous function. Since F is a closed set, $F^\varepsilon \downarrow F$ as $\varepsilon \downarrow 0$, so $\nu(F^\varepsilon) \downarrow \nu(F) \ge \mu(F)$ by continuity from above. Repeating the argument in the other direction, we see that $\mu(F) = \nu(F)$. Since this holds for all closed sets, we find that $\mu = \nu$ by Fact 115, as desired. □

Corollary 117

A sequence of measures μ_n cannot converge weakly to two different limits.

With that fact, we can now think about how to study the space of probability measures. Let μ be any probability measure on S, and for any $\varepsilon > 0$ and (finitely many) bounded continuous functions $f_1, \cdots, f_k$, define the set

$$U_{\varepsilon, f_1, \cdots, f_k}(\mu) = \left\{ \nu \text{ probability measures on } S : \left| \int f_i d\nu - \int f_i d\mu \right| < \varepsilon \text{ for all } 1 \leq i \leq k \right\}$$

to be a kind of neighborhood around the measure μ. We can check the following fact from the definition:

Proposition 118

Let $\mathcal{P}$ be the space of probability measures on $(S, \mathcal{B})$, and let $\mathcal{T}$ be the topology on $\mathcal{P}$ generated by the neighborhoods $U_{\varepsilon, f_1, \cdots, f_k}(\mu)$. Then weak convergence (as in Definition 114) is equivalent to convergence in the topology $\mathcal{T}$.

Because we have neighborhoods around measures, it's natural to also define a distance between measures:

Definition 119

The **Prohorov measure** is defined via

$$\pi(\mu, \nu) = \inf \{ \varepsilon > 0 : \mu(A) \leq \nu(A^\varepsilon) + \varepsilon \text{ and } \nu(A) \leq \mu(A^\varepsilon) + \varepsilon \quad \forall A \in \mathcal{B} \}$$

for any $\mu, \nu \in \mathcal{P}$, where A^ε denotes the ε-neighborhood of the set A in the metric space S.

Fact 120

If S is a **Polish space**, meaning that it is **complete** (Cauchy sequences converge) and **separable** (it has a countable dense subset), then weak convergence of measures is equivalent to convergence in π-measure.

In particular, if S is complete and separable, then $\mathcal{P}$ is also complete and separable and itself satisfies the conditions above. We often want to pick a random probability measure from the set of probability measures, so having the Polish space assumption is nice – this is why almost all of probability happens on Polish spaces, and we'll always be in this setting for this class.

With this, we can turn to the central limit theorem. In general, the main strategy for showing convergence in distribution is to **break the proof into two parts**:

- First, show that the family of measures μ_n is confined in a **compact subset** of $\mathcal{P}$, which implies that there are subsequences of μ_n that converge. (This can usually be done with rough estimates.)
- Once we know that subsequences converge, leverage that fact to show that all subsequential limits coincide.

This might be a bit abstract, so it's important for us to understand what compact spaces in $\mathcal{P}$ actually look like:

Definition 121

Let $\{\mu_\alpha : \alpha \in I\} \subseteq \mathcal{P}$ be a set of probability measures. We say that $\{\mu_\alpha\}$ is **tight** if for all $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subseteq S$ such that $\mu_\alpha(K_\varepsilon) \geq 1 - \varepsilon$ for all $\alpha \in I$.

Tightness of a set of measures basically means that the mass is mostly concentrated within a compact set – two example families of probability measures that are not tight are $\{\nu_n = N(0, n)\}$ and $\{\gamma_n = N(n, 1)\}$ (because significant mass can be arbitrarily far away from the origin). It turns out that tightness and compactness are essentially equivalent in the following way:

Theorem 122 (Prohorov)
Let μ_α be probability measures on a metric space $(S, \mathcal{B})$. If $\{\mu_\alpha\}$ is tight, then the family of measures is **relatively compact** in $\mathcal{P}$ (that is, it has compact closure). Furthermore, if S is a Polish space, then the converse also holds (though this is less useful).

It turns out the case $S = \mathbb{R}$ is easier to prove, and this leads to the **Helly selection theorem**. We should make sure we understand the statement of Prohorov's theorem, and we're responsible for reading and understanding the proof of the Helly selection theorem. (This can be found in the course textbook, but the main idea is to take the family of cumulative distribution functions F_n, repeatedly take subsequences of this family that converge at each rational to get limits at each $q \in \mathbb{Q}$, and then define a function F on $\mathbb{R}$ by making $F(x)$ the infimum of those limits across all $q > x$. This function will be increasing and right-continuous but not necessarily a distribution function because the left and right limits may not be 0 and 1, but they will be if and only if the measures are tight.)

Turning now to the central limit theorem, the key to the proof will be to relate weak convergence to characteristic functions:

Lemma 123
Suppose μ is a probability measure on $\mathbb{R}$ with characteristic function ϕ. Then for any $u \in \mathbb{R}$, we have

$$\mu\left(\left\{x \in \mathbb{R} : |x| \geq \frac{2}{u}\right\}\right) \leq \frac{1}{u}\int_{-u}^{u}(1 - \phi(t))dt.$$

In other words, the tail behavior of the measure μ is related to the characteristic function's values near 0.

Proof. Writing out the definition of the characteristic function and using Tonelli's theorem to change the order of integration, we have

$$\frac{1}{u}\int_{-u}^{u}(1 - \phi(t))dt = \frac{1}{u}\int_{-u}^{u}\int_{-\infty}^{\infty}(1 - e^{itx})d\mu(x)dt = \int_{-\infty}^{\infty}\int_{-u}^{u}\frac{1}{u}(1 - e^{itx})dtd\mu(x).$$

The inner integral can be directly computed, and we end up with

$$= \int_{-\infty}^{\infty} 2\left(1 - \frac{e^{ixu} - e^{-ixu}}{2ixu}\right)d\mu(x) = \int_{-\infty}^{\infty} 2\left(1 - \frac{\sin(ux)}{ux}\right)d\mu(x).$$

The integrand is always nonnegative, so restricting the domain to the region where $|ux| \geq 2$ only makes it smaller. Additionally, because $\sin(ux) \leq 1$, the integrand here is at most 1 whenever $|ux| \geq 2$. Thus we indeed have

$$\frac{1}{u}\int_{-u}^{u}(1 - \phi(t))dt \geq \int_{|ux|\geq 2} 1 d\mu(x) = \mu\left(\{x : |ux| \geq 2\}\right),$$

as desired. □

From this lemma, we can get the following important result:

Theorem 124 (Continuity theorem)

Let μ_n be probability measures on $\mathbb{R}$ with characteristic functions ϕ_n. Then

- If μ_n converge weakly to some μ, then ϕ_n converge pointwise to the characteristic function ϕ of μ.
- Conversely, if ϕ_n converge pointwise to ϕ, and ϕ is continuous at $t = 0$, then ϕ is a valid characteristic function of some probability measure μ with $\mu_n \implies \mu$.

Proof. For the forward direction, if $\mu_n \implies \mu$, then $\phi_n(t) = \int_{-\infty}^{\infty} e^{itx} d\mu_n(x)$ is the integral of the bounded continuous function $f(x) = e^{itx}$, so $\phi_n(t) \to \phi(t)$ converges pointwise by the definition of weak convergence.

For the reverse direction, start with the equation $\mu_n\left(\{|x| \geq \frac{2}{u}\}\right) \leq \frac{1}{u}\int_{-u}^{u}(1 - \phi_n(t))dt$ from Lemma 123. Fix u and take $n \to \infty$; by the bounded convergence theorem (and the assumption that $\phi_n \to \phi$ pointwise), the right-hand side converges to $\frac{1}{u}\int_{-u}^{u}(1 - \phi(t))$.

Additionally, since ϕ is continuous at $t = 0$ and $\phi(0) = 1$, this integral converges to 0 as $u \to 0$. This shows that the μ_n are tight – indeed, $f(u) = \limsup_{n\to\infty} \mu_n\left(|x| \geq \frac{2}{u}\right)$ goes to 0 as $u \downarrow 0$, so for any $\varepsilon > 0$ we can take an appropriately small u such that $f(u) < \varepsilon$. That implies that $\mu_n\left(\{|x| \geq \frac{2}{u}\}\right) < \varepsilon$ for all but finitely many n, and then we can take the union of this compact set with the finitely many compact sets that include at least $1 - \varepsilon$ of the measures at the beginning of our sequence.

Thus, μ_n weakly converges along subsequences by Prohorov's theorem, so for any subsequence where we have weak convergence $\mu_{n_k} \implies \nu$, we must also have $\phi_{\mu_{n_k}} \to \phi_\nu$ (by the forward direction argument). But we have $\phi_\nu = \phi$ for any subsequence (because the whole sequence ϕ_n converges too ϕ by assumption), so because ϕ is the limit of such a weak convergence, it is a valid characteristic function. This argument applies for any subsequence, so all subsequential limits exist and are identical. Therefore every subsequence of μ_n has a further subsequence converging to μ, and (as a general property of topological spaces) this implies that μ_n converges to μ as desired. □

Example 125 (A non-example)

Consider the wide Gaussians $\mu_n = N(0, n)$ from before (which are not tight). Then $\phi_n(t) = \exp\left(\frac{-nt^2}{2}\right)$ converges to the indicator function $1\{t = 0\}$, which is not continuous. Indeed, the measures μ_n do not converge weakly to any probability measure μ.

Remark 126 (Joke)**.** *In response to "Does this imply CLT?", the answer was "I don't know, does it?"*

Thus, the continuity theorem tells us that all we need is to show that the characteristic functions of the left and right side are the same in our central limit theorem statement. We can check the following fact with a calculus bash (which can also be found in the textbook):

Lemma 127

For all $x \in \mathbb{R}$, $\left|e^{ix} - \left(1 + ix - \frac{x^2}{2}\right)\right| \leq \min\left(x^2, \frac{|x|^3}{6}\right)$.

We're now ready to prove two versions of the central limit theorem:

Theorem 128 (Central limit theorem for iid sequences)
Let X, X_i be iid random variables with $\mathbb{E}[X] = 0$, and $\mathbb{E}[X^2] = 1$. (These values can always be rescaled, but we do require a finite second moment, which is stronger than the assumption for the strong law of large numbers.) Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0,1).$$

Proof. By linearity of expectation, we know that

$$\mathbb{E}\left[e^{itX} - \left(1 + itX - \frac{t^2X^2}{2}\right)\right] = \phi_X(t) - 1 - \frac{t^2}{2}.$$

But plugging in tx into both sides of Lemma 127, taking expectations (which is allowed because both sides are nonnegative random variables), and then dividing by t^2 yields

$$\frac{1}{t^2}\left|\phi_X(t) - \left(1 - \frac{t^2}{2}\right)\right| \leq \mathbb{E}\left[\min\left(X^2, \tfrac{t}{6}|X|^3\right)\right].$$

The expression $\min\left(X^2, \frac{t}{6}|X|^3\right)$ converges to zero pointwise almost surely as $t \to 0$ (because $\frac{t}{6}|X|^3$ goes to zero), and it is dominated by the integrable X^2. Thus, the right-hand side goes to zero as $t \to 0$ by the dominated convergence theorem, which implies that

$$\phi_x(t) = 1 - \frac{t^2}{2} + o(t^2).$$

From here, the random variable $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ has characteristic function

$$\phi_{Z_n}(t) = \phi_X\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

For any fixed t, this converges to $\exp\left(-\frac{t^2}{2}\right)$ as $n \to \infty$ (we can check this by taking a log first and using L'Hopital's rule), which is the characteristic function of the standard Gaussian. Thus $\phi_{Z_n} \to \phi_{N(0,1)}$ pointwise, so Theorem 124 yields the desired result. □

Remark 129. *If we had tried to produce a similar proof with the moment generating function instead of the characteristic function, we would run into more problems. In particular, it's possible to have finite second moment and not have the moment generating function be defined anywhere but $t = 0$, and we also don't have such a nice inversion theorem in that case.*

Finally, we'll prove a more general central limit theorem:

Theorem 130 (Lindeberg-Feller central limit theorem)
Suppose that we have a triangular array of random variables such that for each n, $\{X_{n,j} : 1 \leq j \leq n\}$ are **mutually independent** with mean 0 but not necessarily iid. Also suppose that the functions are normalized such that $\mathrm{Var}(X_{n,1} + \cdots + X_{n,n}) = \sum_{j=1}^{n} \mathbb{E}\left[X_{n,j}^2\right]$ converges to some $\sigma^2 > 0$ as $n \to \infty$, and (here is the key assumption) for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{E}\left[X_{n,j}^2; |X_{n,j}| > \varepsilon\right] = 0.$$

Then $S_n = \sum_{i=1}^{n} X_n$ converges in distribution to $N(0, \sigma^2)$.

Theorem 128 is a special case of this – if X_i are our iid random variables, then we can plug in $X_{n,i} = \frac{X_i}{\sqrt{n}}$ and check that the conditions hold. The reason we don't have any $\sqrt{n}$ factor here is that we are assuming we have already done the normalization for the variance to work out, and the key condition says intuitively that each individual variable cannot contribute significantly to the variance.

Proof. We'll show again that the characteristic function of S_n converges to that of a Gaussian. Let $\phi_{n,j}$ be the characteristic function for $X_{n,j}$ and let $\sigma_{n,j}^2 = \mathbb{E}(X_{n,j}^2)$. By the same inequality as in the previous proof (but this time with variance $\sigma_{n,j}^2$), we have

$$\left|\phi_{n,j}(t) - \left(1 - \frac{\sigma_{n,j}^2 t^2}{2}\right)\right| \le \mathbb{E}\left[\min\left(t^2 X_{n,j}^2, \frac{t^3 |X_{n,j}|^3}{6}\right)\right].$$

The second term is smaller when $X_{n,j}$ is small, and otherwise the first term can be bounded with the theorem assumption. So using one or the other inequality depending on whether $|X_{n,j}|$ is larger than ε, we get the bound

$$\left|\phi_{n,j}(t) - \left(1 - \frac{\sigma_{n,j}^2 t^2}{2}\right)\right| \le t^2\left[\frac{t\varepsilon \mathbb{E}(X_{n,j}^2)}{6} + \mathbb{E}(X_{n,j}^2; |X_{n,j}| > \varepsilon)\right]$$

for any ε (where the first term comes from the $|X|^3$ term and the second from the X^2 term). We can now use the fact that for any $z_i, w_j \in \mathbb{C}$ with modulus at most 1, $\boxed{\left|\prod_{i=1}^n z_i - \prod_{j=1}^n w_j\right| \le \sum_{j=1}^n |z_j - w_j|}$ (by writing a telescoping sum where we switch from z_1 to w_1, then z_2 to w_2, and so on). The characteristic function $\phi_{n,j}(t)$ always has modulus at most 1, and

$$\sigma_{n,j}^2 = \mathbb{E}(X_{n,j}^2) = \mathbb{E}(X_{n,j}^2; |X_{n,j}| < \varepsilon) + \mathbb{E}(X_{n,j}^2; |X_{n,j}| \ge \varepsilon) \le \varepsilon^2 + \mathbb{E}(X_{n,j}^2; |X_{n,j}| \ge \varepsilon)$$

can be made arbitrarily small by taking $\varepsilon \to 0$ and $n \to \infty$ (by our key assumption). Thus, for sufficiently large n and using sufficiently small ε, the term $\left|\frac{\sigma_{n,j}^2 t^2}{2}\right|$ is always at most 1. So applying the boxed identity above, we find that

$$\left|\prod_{j=1}^n \phi_{n,j}(t) - \prod_{j=1}^n \left(1 - \frac{\sigma_{n,j}^2 t^2}{2}\right)\right| \le \sum_{j=1}^n \left[t^2\left[\frac{t\varepsilon \mathbb{E}(X_{n,j}^2)}{6} + \mathbb{E}(X_{n,j}^2; |X| \ge \varepsilon)\right]\right].$$

For any fixed t, both terms tend to 0 as $\varepsilon \to 0$ and $n \to \infty$ (the first term does so because the $\sum_{n,j} \mathbb{E}[X_{n,j}^2]$ converges to a finite value and then we multiply by ε, and the second term does so by our key assumption), so the left-hand side converges to zero as well. In other words, for any t,

$$\lim_{n\to\infty} \phi_{S_n} = \lim_{n\to\infty} \prod_{j=1}^n \phi_{n,j}(t) \to \prod_{j=1}^n \left(1 - \frac{\sigma_{n,j}^2 t^2}{2}\right).$$

We've shown that $\sigma_{n,j}^2$ can be made arbitrarily small for large enough n, which is enough to make the product on the right-hand side converge to $e^{-t\sigma^2/2}$ (again this can be checked by taking logs on both sides). This is the characteristic function of the Gaussian $N(0, \sigma^2)$ as desired, so Theorem 124 again yields the desired result. □

15 October 30, 2019

Office hours will run from 4–6 pm today, and we can come to see our midterms (for logistical reasons, we won't get them back during class). Our grades are on Stellar already – the average is fairly low and the distribution is very interesting. The score is out of 40, and when we see our score, we should take it and divide it by 40, and then multiply

it by 200 to think about how well we did. If we scored a 10 out of 40, this is problematic (and we should talk to Professor Sun about whether we should stay in the class), but otherwise we should just interpret the score as how well we're doing. Moving forward, the class won't change much – the pace and difficulty will be about the same as it has been so far, but maybe the problem with the first exam is that we ran out of time. So there might be a time arranged outside of class for us to take the exam (at night, for longer than an hour and a half).

Last time, we talked about the topology of weak convergence (for probability measures on a metric space). Notably, we discussed Prohorov's theorem, which was used to prove the Lindeberg-Feller central limit theorem. This last result essentially states that if we have a triangular array of independent mean-zero random variables $X_{n,j}$, and we know that $\sum_{j=1}^n \mathbb{E}[X_{n,j}^2]$ converges to a finite value σ^2 and that the contribution from large values $\sum_{j=1}^n \mathbb{E}[X_{n,j}^2; |X_{n,j}| \geq \varepsilon]$ goes to zero as $n \to \infty$, then the row sums $\sum_{j=1}^n X_{n,j}$ converge in distribution to $N(0, \sigma^2)$. Let's see an example of this in action:

Example 131

Let π be a uniformly random permutation on $[n] = \{1, \cdots, n\}$ (there are $n!$ such possibilities). We'll study the behavior of the **number of cycles** in π as n grows large.

We'll write the permutations in **(sorted) cycle notation**. For example, having $\pi = (136)(2975)(48)$ indicates that 1 maps to 3 maps to 6 maps to 1, and so on. To ensure uniqueness of representation, we make sure that the smallest number in each cycle appears at the beginning, and we sort the cycles from left to right by minimum element (as we've done above). This is useful because we can now **sample π sequentially** from left to right: start by writing down "(1", and then pick a random integer (uniformly from 1 to n) for 1 to go to, say 3. (If it's 1, we instead immediately close the parentheses.) Next, pick another random integer that is not 3 (for example 6) – if it's 1, close the parentheses, and otherwise, pick another random integer which is not 3 or 6. Once we return to 1, we close that cycle and start again with the next smallest integer not yet picked, ignoring all of the numbers already in a determined cycle.

This sampling method is convenient, because we can now define the random variables

$$I_{n,k} = 1\{\text{a right parenthesis appears after the } k\text{th number in sorted cycle notation}\}.$$

(For example, $I_{n,k}$ takes on the values $(0, 0, 1, 0, 0, 0, 1, 0, 1)$ in our example permutation π.) Since the number of right parentheses is the same as the number of cycles, our goal is to determine the behavior of $S_n = \sum_{k=1}^n I_{n,k}$. And because of the way we sample, the $I_{n,k}$ are actually **independent Bernoulli variables** of parameter $\frac{1}{n-(k-1)}$, because there are $(k-1)$ choices that the kth number cannot go to, and out of the remaining $n-(k-1)$ choices, the probability of having a right parenthesis is the probability that we choose the current first entry of the cycle. Additionally, we can see that any permutation will be sampled with probability $\frac{1}{n} \cdot \frac{1}{n-1} \cdot \cdots \cdot 1 = \frac{1}{n!}$, so this does indeed yield a uniformly random permutation of $[n]$.

Thus, we're in a setting to apply the Lindeberg-Feller central limit theorem to our random variables $I_{n,k}$. The expected number of cycles in π is

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

and the variance is (using that a $\text{Ber}(p)$ random variable has variance $p(1-p)$)

$$\text{Var}(S_n) = \sum_{k=1}^n \text{Var}(I_{n,k}) = \sum_{j=1}^n \frac{1}{j}\left(1 - \frac{1}{j}\right) = \mathbb{E}[S_n] - \sum_{j=1}^n \frac{1}{j^2} = \log n + O(1).$$

To apply Lindeberg-Feller, we recenter and rescale our random variables by defining the random variables

$$X_{n,k} = \frac{I_{n,k} - \mathbb{E}[I_{n,k}]}{\sqrt{\log n}}.$$

This rescaling makes the first condition of Lindeberg-Feller hold (the sum of the variances of the $X_{n,k}$s is $\frac{1}{\sqrt{\log n^2}}\text{Var}(S_n)$, which goes to 1 as $n \to \infty$). The second condition is pretty easy to check too, because each term of the expression

$$\sum_{k=1}^{n} \mathbb{E}\left[\left(\frac{I_{n,k} - \mathbb{E}[I_{n,k}]}{\sqrt{\log n}}\right)^2 ; \left|\frac{I_{n,k} - \mathbb{E}[I_{n,k}]}{\sqrt{\log n}}\right| \geq \varepsilon\right]$$

is just equal to 0 for all n large enough (the numerator $I_{n,k} - \mathbb{E}I_{n,k}$ is at most 1, so multiplying it by $\frac{1}{\sqrt{\log n}}$ makes $\left|\frac{I_{n,k} - \mathbb{E}[I_{n,k}]}{\sqrt{\log n}}\right|$ smaller than ε almost surely for sufficiently large n). Thus Lindeberg-Feller applies, and we get the result

$$\sum_{k=1}^{n} \frac{I_{n,k} - \mathbb{E}[I_{n,k}]}{\sqrt{\log n}} \xrightarrow{d} N(0,1).$$

Substituting in $S_n = \sum_{k=1}^{n} I_{n,k}$ and $\sum_{k=1}^{n} \mathbb{E}[I_{n,k}] = \log n + O(1)$, we thus find that

$$\boxed{\frac{S_n - \log n}{\sqrt{\log n}} \xrightarrow{d} N(0,1)},$$

and we've seen an example in action where the central limit theorem can be applied even though the random variables are not identically distributed.

However, there are some settings in which sums of random variables do not converge to a Gaussian:

Example 132

Suppose the random variables $I_{n,k}$ are all distributed according to $\text{Ber}(\frac{\lambda}{n})$ for some constant λ, and we want to study the behavior of $S_n = \sum_{k=1}^{n} I_{n,k}$ as n grows large.

This time, $S_n =\sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ has expectation λ by linearity of expectation, so the probability that S_n is extremely large (like 100λ) is pretty small. So because most of the mass is supported on a constant range (λ is a constant not depending on n), but S_n is only supported on the nonnegative integers, it doesn't really make sense to expect that the distribution will approach that of a Gaussian. We can carry out the calculations to make this more clear: notice that

$$\mathbb{P}(S_n = k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}\frac{(n)_k}{n^k}\left(1 - \frac{\lambda}{n}\right)^{n-k},$$

where $(n)_k$ denotes the falling factorial $\frac{n!}{(n-k)!}$. If we fix k and take $n \to \infty$, the middle term goes to 1, and the $-k$ in the right term's exponent becomes irrelevant, and thus

$$\lim_{n\to\infty} \mathbb{P}(S_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

So in this case, S_n actually converges in law to the **Poisson distribution** $\text{Pois}(\lambda)$. Again, this is supported on the integers, so it's definitely not Gaussian. However, if we take λ large enough, we do recover the Gaussian limit – in other words, $\frac{\text{Pois}(\lambda)-\lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1)$ as $\lambda \to \infty$. Another way of saying this is that if we consider the random variable $\frac{\text{Bin}(n,p)-np}{\sqrt{np(1-p)}}$, which is a random variable with mean 0 and variance 1, then (by direct calculation or by the central limit theorem), it will converge to a standard normal if we take p fixed and n large, but it will converge to a Poisson distribution if we take $p = \frac{\lambda}{n}$ and n large. Let's verify that we do get the correct limit behavior as we take $\lambda \to \infty$:

Proposition 133

The Poisson distribution Pois(λ) converges in distribution to a Gaussian as $\lambda \to \infty$.

Proof. We'll make use of characteristic functions. If $Y_\lambda \sim \text{Pois}(\lambda)$, then the characteristic function of Y_λ is

$$\phi(t) = \mathbb{E}\left[e^{itY}\right] = \sum_{k\geq 0} \frac{e^{itk}e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k\geq 0}\frac{\left(\lambda e^{it}\right)^k}{k!} = \exp(-\lambda(1-e^{it})).$$

Since Y has mean and variance λ, we can define a normalized version $Z_\lambda = \frac{Y_\lambda - \lambda}{\sqrt{\lambda}}$. The characteristic function for Z_λ is then

$$\psi(t) = \mathbb{E}\left[\exp\left(it\left(\frac{Y_\lambda - y}{\sqrt{\lambda}}\right)\right)\right] = \exp(-it\sqrt{\lambda})\mathbb{E}\left[e^{itY_\lambda/\sqrt{\lambda}}\right].$$

Plugging this into the form of ϕ above, we thus find that

$$\boxed{\psi(t) = \exp\left(-\lambda\left(1 - e^{it/\sqrt{\lambda}}\right) - it\sqrt{\lambda}\right)}.$$

But as λ gets large, we can do a series expansion, and we'll find that

$$\psi(t) = \exp\left(-\lambda\left(-\frac{it}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} - O\left(\frac{1}{\lambda^{3/2}}\right)\right) - it\sqrt{\lambda}\right),$$

which converges pointwise to $e^{-t^2/2}$ as $\lambda \to \infty$. Thus, the usual continuity theorem tells us that the rescaled Poisson random variables do indeed converge to a standard normal. □

Professor Sun once taught a class similar to 18.600, so she feels a little silly going over this next topic, but this next part may be useful because apparently many of us haven't taken an intro probability class before.

Example 134

We're going to review all of the standard probability distributions. Consider **Bernoulli trials**, which means that we have random variables I_k distributed iid according to Ber(p) for all $k \geq 1$.

We'll say that $I_k = 1$ (which occurs with probability p) is a "success" and $I_k = 0$ (which occurs with probability $1-p$) is a "failure." Then the number of successes $B_m = \sum_{k=1}^m I_k$ after time m (that is, m trials) is distributed according to the **binomial distribution** Bin(m, p), with

$$\mathbb{P}(B_m = k) = \binom{m}{k} p^k (1-p)^{m-k}$$

for any $0 \leq k \leq m$. Meanwhile, the time G of the first success, or equivalently the first index k such that $I_k = 1$, is distributed according to the **geometric distribution** Geo(p), with

$$\mathbb{P}(G = k) = (1-p)^{k-1}p.$$

As an extension of this, the time of the rth success X_r is distributed as $X_r \stackrel{d}{=} G_1 + \cdots + G_r$ where G, G_i are iid Geo(p) random variables. (In other words, the negative binomial distribution is the geometric distribution convolved with itself r times.) X_r is then distributed according to the **negative binomial distribution** NegBin(r, p), with

$$\mathbb{P}(X_r = t) = \binom{t-1}{r-1}(1-p)^{t-r}p^r$$

for any $t \geq r$ (because we need exactly $(r-1)$ successes in the first $(t-1)$ trials, followed by a success).

Next, suppose that we scale $p = \frac{1}{n}$, so that the vast majority of our experiments are failures. If we also scale time by n (so that we can see successes at a reasonable rate instead of very rarely), then we'll have done nt trials by time t. Then as we calculated in Example 132, $B_{nt} \sim \text{Bin}\left(nt, \frac{1}{n}\right) \xrightarrow{d} \text{Pois}(t)$ converges to the **Poisson distribution**. Meanwhile, the time E of the first success converges to a **continuous** random variable $\frac{1}{n}\text{Geo}\left(\frac{1}{n}\right) \xrightarrow{d} \text{Exp}$, the **exponential distribution**, which has density

$$f_E(t) = 1\{t \geq 0\}e^{-t}$$

(we can prove this with characteristic functions or direct calculations). The time of the rth success then also converges to $\frac{G_1+\cdots+G_r}{n} \sim \frac{1}{n}\text{NegBin}\left(r, \frac{1}{n}\right) \xrightarrow{d} E_1 + \cdots + E_r$, where the E_i are iid exponential. This is the exponential law convolved with itself r times, which gives us the **Gamma distribution** $\text{Gamma}(r)$. The gamma density is good to remember, and it can be derived by taking the limit of the negative binomial distribution:

$$\mathbb{P}\left(\frac{\text{NegBin}(r, 1/n)}{n}\right) \in [z, z+dz] = \binom{nz-1}{r-1}\left(1-\frac{1}{n}\right)^{nz-r}\left(\frac{1}{n}\right)^r (n\ dz) = \frac{(nz-1)_{r-1}}{(r-1)!n^r}\left(1-\frac{1}{n}\right)^{nz-r}(ndz).$$

If we now fix r and z and take $n \to \infty$, we find that

$$\mathbb{P}\left(\frac{\text{NegBin}(r, 1/n)}{n}\right) \in [z, z+dz] = \frac{z^{r-1}e^{-z}}{(r-1)!}dz \implies f_{E_1+\cdots+E_r}(t) = 1\{t \geq 0\}\frac{t^{r-1}e^{-t}}{(t-1)!}.$$

For general real numbers $\alpha > 0$ (not necessarily an integer), the $\text{Gamma}(\alpha)$ distribution similarly has density $1\{t \geq 0\}\frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}$, where $\Gamma(\alpha)$ is the normalizing constant. (In particular, Γ is a generalization of the factorial, because $\Gamma(r) = (r-1)!$ for positive integer r.) And the Gamma distribution will approach a Gaussian if the amount of time passed is large relative to our sampling rate – in other words, we converge to a Gaussian as $\alpha \to \infty$.

Fact 135

We showed in Example 131 that the number of cycles S_n of a random permutation of $[n]$ satisfies $\frac{S_n - \log n}{\sqrt{\log n}} \xrightarrow{d} N(0,1)$. If we let $C_{n,k}$ be the number of cycles in π of length k, then it turns out that the tuple $(C_{n,1}, C_{n,2}, \cdots, C_{n,n})$ converges in distribution to $(Y_k)_{k \geq 1}$, where $Y_k \sim \text{Pois}(\frac{1}{k})$. The proof is a tiny bit outside the scope of what we've done so far, and we can see Arratia and Tavaré's paper [3] for more details. (The main idea is that a large number of trials for a rare event is approximately Poisson, and the events that two different vertices are both in cycles of length k are close to independent.) For now, we can at least check the expectations $\mathbb{E}[C_{n,k}] = \frac{1}{k}$, because there are (combinatorially) $\frac{(n)_k}{k}$ possible cycles of length k, and each one occurs with probability $\frac{1}{(n)_k}$ (this can be checked by exploring the cycle one element at a time).

With the remaining time in this lecture, we'll discuss some weak convergence concepts, with a reminder of what results and proofs we should know for this class in general. For what follows, let S be a complete separable metric space (that is, a Polish space) with Borel sigma-algebra $\mathcal{B}$. The space of probability measures $\mathcal{P}$ on S is then also a Polish space. This next result gives us useful alternative characterizations of weak convergence:

Theorem 136 (Portmanteau theorem)
Let μ_n be a sequence of probability measures. Then the following are equivalent:

- $\mu_n \implies \mu$.
- $\int f d\mu_n \to \int f d\mu$ for all bounded **uniformly** continuous f.
- $\limsup \mu_n(F) \leq \mu(F)$ for all closed $F \subseteq S$.
- $\liminf \mu_n(G) \geq \mu(G)$ for all open $G \subseteq S$.
- $\lim \mu_n(A) = \mu(A)$ for all A with $\mu(\partial A) = 0$.

There is also a way to convert weak convergence of measures into a statement about convergence of random variables:

Theorem 137 (Skorohod)
If $\mu_n \implies \mu$, then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and measurable mappings $X_n, X : \Omega \to S$ such that $\mathcal{L}_{X_n} = \mu_n, \mathcal{L}_X = \mu$, and X_n converges to X almost surely under $\mathbb{P}$.

We should know the statements of these proofs and of Prohorov's theorem (Theorem 122), and we should know a bit more when we're looking at the special case under the real line ($S = \mathbb{R}$). Let's do a bit of review for that: if we have a set of measures on the real line, then each μ_n can be represented by its F_n, and μ can be represented with its cdf F. We then say that $F_n \implies F$ if $F_n(x) \to F(x)$ at all points of continuity of F – the portmanteau theorem then implies that $\mu_n \implies \mu$ if and only if $F_n \implies F$. From there, tightness in $\mathbb{R}$ is easy to characterize, because the definition is equivalent to requiring that for any $\varepsilon > 0$, $\inf_n\{F_n(x) - F_n(-x)\} \geq 1 - \varepsilon$ for some sufficiently large x. So showing Prohorov's theorem for $\mathbb{R}$ means that we just need to extract subsequences for our F_ns which don't have mass escaping from the side, which follows by a diagonalization argument (the Helly selection theorem).

Skorohod's theorem is somewhat abstract in general, but (just like Prohorov's theorem) it's simpler over $\mathbb{R}$, and it can also be very useful:

Proof of Theorem 137 for $S = \mathbb{R}$. Represent μ with its cdf F – first suppose that F is one-to-one. In this case, let U be uniform on $[0, 1]$; we claim that $F^{-1}(U) \sim \mu$. Indeed, for any $x \in \mathbb{R}$, we have

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$$

(where the first equality comes from F being monotone and one-to-one, and the second comes from $F(x)$ being between 0 and 1). More generally, if F is not one-to-one, define the function $X(u) = \inf\{y : F(y) \geq u\}$ in place of $F^{-1}(u)$. Then notice that we have $X(u) \leq x$ if and only if $F(x) \geq u$ – the reverse direction is clear, and for the forward direction, $X(u) \leq x$ implies that there are $y < x + \varepsilon$ such that $F(y) \geq u$ for all $\varepsilon > 0$, so $F(x) \geq u$ by right-continuity of F. Thus we again have

$$\mathbb{P}(X(U) \leq x) = \mathbb{P}(F(x) \geq U) = F(x).$$

Basically, given any cdf F, we can define a "rough inverse" of it using this mapping X, so we can construct the desired random variables as follows. Let the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ be $([0, 1], \mathcal{B}, \text{Leb})$, and define the random variables $X, X_n : \Omega \to \mathbb{R}$ via

$$X(\omega) = \inf\{x : F(x) \geq \omega\}, \quad X_n(\omega) = \inf\{x : F_n(x) \geq \omega\}.$$

Weak convergence then implies that F_n converges to F pointwise where F is discontinuous (which only happens at countably many points because F is monotone) points of discontinuity. So for all ω where X is continuous, X_n converges to X. (Here's one way to show that: if X is continuous at $X(\omega) = x$, then for all $\varepsilon > 0$ there is some $\delta > 0$ such that for all $\omega' \in (\omega - 3\delta, \omega + 3\delta)$, $X(\omega') \in (x - \frac{\varepsilon}{2}, x + \frac{\varepsilon}{2})$. Picking some $\varepsilon' \in (\frac{\varepsilon}{2}, \varepsilon)$ such that F is continuous at both $x - \varepsilon'$ and $x + \varepsilon'$, we must have $F(x - \varepsilon') < \omega - 2\delta$ and $F(x + \varepsilon') > \omega + 2\delta$ (by definition of X and that F is monotone). Because F_n converges to F pointwise at $x \pm \varepsilon'$, for all sufficiently large n we thus have $F_n(x - \varepsilon') < \omega - \delta$ and $F_n(x + \varepsilon') > \omega + \delta$, meaning $X_n(\omega) \in (x - \varepsilon, x + \varepsilon)$ because $\varepsilon' < \varepsilon$. Taking $\varepsilon \to 0$ shows the result.) Since X itself is also monotone, the discontinuity points are of measure zero, and thus X_n converges to X almost surely. Since we've already shown that the laws of X_n and X are μ_n and μ respectively, we've proven the desired result. □

(The proof of Skorohod's theorem in general is kind of related to what we've done here, but it is more complicated because we don't have the simple mapping using F anymore.)

16 November 4, 2019

Class started with another attendance quiz today:

Problem 138

Let X, X_i be iid exponential random variables of density $1\{x > 0\}e^{-x}dx$. Find a sequence b_n and a random variable Y such that $(\max_{1 \le i \le n} X_i) - b_n \xrightarrow{d} Y$.

Solution. We can write the distribution function for the maximum of independent random variables in terms of the distribution functions of the individual variables: specifically,

$$\mathbb{P}\left(\max_{1 \le i \le n} X_i \le t\right) = \mathbb{P}(X \le t)^n = (1 - e^{-t})^n.$$

Substituting in $t = \log n + x$ to eliminate the dependence of n on the right-hand side, we have

$$\mathbb{P}\left(\max_{1 \le i \le n} X_i \le \log n + x\right) = (1 - \exp(-(\log n + x)))^n,$$

which converges to $\exp(-e^{-x})$ as $n \to \infty$. Thus, we can take $\boxed{b_n = \log n}$, and $Y = \max X_i - b_n$ converges in distribution to a random variable Y with distribution function $\boxed{\mathbb{P}(Y \le y) = e^{-e^{-y}}}$, which is called the **Gumbel distribution**. □

Remark 139. *The terms* **convergence in law** *and* **weakly convergent** *mean the same thing.*

The problem above is an example of an **extreme value statistic**, where we want to study the largest of a bunch of samples. Let's do another example of this type:

Example 140

Let $Z, Z_i \sim N(0, 1)$ be standard Gaussians with density $g(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{dz}\right) dz$. We wish to study the distribution of $\max(Z_1, \cdots, Z_n)$.

Solution. To use a similar strategy as the previous problem, we need to similarly study $\mathbb{P}(Z \le x)$ but for a standard Gaussian. This time, the function

$$\Psi(x) = \mathbb{P}(Z \ge x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$

has no closed form expression, but we can get a nice lower bound for it when x is large:

$$\Psi(x) \le \frac{\mathbb{E}(Z; Z \ge x)}{x} = \frac{1}{\sqrt{2\pi}x} \int_x^\infty z \exp\left(-\frac{z^2}{2}\right) dz,$$

which we can integrate with a u-substitution $u = -\frac{z^2}{2}$ to find that $\boxed{\Psi(x) \le \frac{1}{\sqrt{2\pi}x} \exp\left(-\frac{x^2}{2}\right) = \frac{g(x)}{x}}$. This is only an upper bound, and it's not very useful for small x. But we can show that it's a pretty tight bound for large x – write

$$\sqrt{2\pi}\Psi(x) = \int_x^\infty \frac{1}{z} \cdot -z \exp\left(-\frac{z^2}{2}\right) dz,$$

and now integrate by parts (with $u = \frac{1}{z}$ and $dv = -z \exp\left(-\frac{z^2}{2}\right)$) to simplify to

$$\sqrt{2\pi}\Psi(x) = -\frac{1}{z} \exp\left(-\frac{z^2}{2}\right)\Big|_x^\infty - \int_x^\infty \frac{1}{z^2} \exp\left(-\frac{z^2}{2}\right) dz.$$

The first term evaluates to $\frac{1}{x}\exp\left(-\frac{x^2}{2}\right)$, and the second can be integrated again by parts – letting $u = \frac{1}{z^3}$ and $dv = z\exp\left(-\frac{z^2}{2}\right)$, we have

$$\sqrt{2\pi}\Psi(x) = \frac{1}{x}\exp\left(-\frac{x^2}{2}\right) - \left[-\frac{1}{z^3}\exp\left(-\frac{z^2}{2}\right)\right]_x^\infty + \int_x^\infty \frac{3}{z^4}\exp\left(-\frac{z^2}{2}\right) dz.$$

We want a lower bound, so we can toss the last term because the integrand is always positive. We thus get the bound

$$\sqrt{2\pi}\Psi(x) \ge \frac{1}{x}\exp\left(-\frac{x^2}{2}\right) - \frac{1}{x^3}\exp\left(-\frac{x^2}{2}\right) \implies \boxed{\Psi(x) \ge \frac{g(x)}{x}\left(1 - \frac{1}{x^2}\right)}.$$

This implies that for x large and for some $t \in \mathbb{R}$ with $|t| \ll x$, we have the ratio

$$\frac{\Psi(x+t)}{\Psi(x)} \sim \frac{\frac{1}{\sqrt{2\pi}(x+t)}\exp\left(-\frac{(x+t)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}x}\exp\left(-\frac{x^2}{2}\right)} \sim \exp\left(-xt - \frac{t^2}{2}\right),$$

and the leading order behavior is proportional to e^{-tx} – more specifically, we can change variables $t \mapsto \frac{t}{x}$ to get

$$\lim_{x\to\infty} \frac{\Psi\left(x + \frac{t}{x}\right)}{\Psi(x)} = \exp(-t).$$

With this, we can go back to our original question: we have Z_is that are iid standard Gaussians, and we want to understand the distribution of $M_n = \max Z_i$. Just like in Problem 138,

$$\mathbb{P}(M_n \le x) = (1 - \Psi(x))^n,$$

and we want to pick $\Psi(x)$ to be approximately $\frac{1}{n}$ to again get rid of the n-dependence. So choose $x = b_n$ such that this is true (this is a deterministic real number for each n, because Ψ is monotone). We then have, as $b_n \to \infty$,

$$\mathbb{P}\left(M_n \le b_n + \frac{t}{b_n}\right) = \left(1 - \Psi\left(b_n + \frac{t}{b_n}\right)\right)^n = \left(1 - \frac{e^{-t}}{n}(1 - o_n(1))\right)^n$$

by our calculation above, and just like before this converges to $e^{-e^{-t}}$ as $n \to \infty$! So the only difference from last time is that our variables need to be rescaled:

$$\boxed{\left(\max_{1\le i\le n} Z_i - b_n\right) b_n \xrightarrow{d} \text{Gumbel}}.$$

□

Remark 141. *Not all random variables behave in this way – the result depends on the tail behavior of the distribution for X – but there is a large class of variables for which we get convergence to the Gumbel distribution. (And that's why Gumbel is an extreme value statistic.)*

We can also say a bit more about the values of b_n we are picking for the standard Gaussian – if we want $\Psi(b_n) = \frac{1}{n}$, then the rough behavior we're looking for is that

$$\frac{1}{n} = \frac{g(b_n)}{b_n} = \frac{1}{\sqrt{2\pi} b_n} \exp\left(-\frac{b_n^2}{2}\right).$$

Based on just this equation, it looks like we want $b_n^2 = 2\log n + (\text{lower order terms})$ – the next correction should account for the b_n in the numerator. Taking logs on both sides of the preivous equation, we find that

$$\frac{b_n^2}{2} = \log n - \frac{1}{2}\log\log n + O(1) \implies \boxed{b_n = \sqrt{2\log n}\left(1 - \frac{\log\log n + O(1)}{4\log n}\right)}.$$

And the estimate of $\Psi(b_n) \approx \frac{g(b_n)}{b_n}$ is good enough, because adjusting by the factor of $\left(1 - \frac{1}{b_n^2}\right)$ only changes b_n by a multiplicative factor of $\left(1 - \frac{O(1)}{\log n}\right)$.

We'll do some more examples of weak convergence on our homework – it's useful in general to be able to show convergence of a sequence of random variables, so there will definitely be something about that on our next exam. But for now, we'll move on and spend a bit of time talking about weak convergence on $\mathbb{R}^d$. (We should also read section 3.10 in the textbook, but we'll go over some of the main points here.)

Definition 142

Let μ be a probability measure on $\mathbb{R}^d$. The **generalized cdf** of μ is defined by $F(x) = \mu\left(\prod_{i=1}^d (-\infty, x_i]\right)$ for $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$, and the **characteristic function** of μ is defined by $\chi_\mu(t) = \int \exp\left(i\langle t, x\rangle\right) d\mu(x)$ for $t \in \mathbb{R}^d$.

We showed on our homework that F is monotone and right-continuous, and we also found that the measure of any d-dimensional rectangle can be found by an inclusion-exclusion argument as

$$\mu\left(\prod_{i=1}^d (a_i, b_i]\right) = \sum_{v\in V} (-1)^{\{\#i : v_i = a_i\}} F(v),$$

where V is the set of vertices $\{a_1, b_1\} \times \cdots \times \{a_d, b_d\}$. Then it turns out that $\mu_n \implies \mu$ (with the general definition from Definition 114) if and only if F_n converges weakly to F, meaning that $F_n(x) \to F(x)$ for all x where F is continuous. And for the characteristic function, notice that the integrand still takes values on the unit circle, so $\chi_\mu(t)$ still takes values inside the unit disk. We do have a Fourier inversion theorem in higher dimensions as well, but it's easier to state if we assume the boundary has zero measure:

Theorem 143 (Fourier inversion for probability measures on $\mathbb{R}^d$)
Let μ be a probability measure on $\mathbb{R}^d$, and suppose we have $A = \prod_{i=1}^d [a_i, b_i]$ with $\mu(\partial A) = 0$. Then

$$\mu(A) = \lim_{T\to\infty} \int_{[-T,T]^d} \phi_\mu(t) \prod_{j=1}^d \left(\frac{e^{-it_j a_j} - e^{-it_j b_j}}{2\pi i t_j} \right) dt.$$

So again knowing ϕ determines μ even in higher dimensions (because we can find a dense set of reals to pick our endpoints a_i, b_i from such that $\mu(\partial A)$ is always zero). With this, the one-dimensional continuity theorem also generalizes directly with a similar proof (showing tightness):

Theorem 144 (Continuity theorem in higher dimensions)
Let μ_n be probability measures on $\mathbb{R}^d$ with characteristic functions ϕ_n. Then

- If $\mu_n \implies \mu$, then ϕ_n converge pointwise to ϕ_μ.
- If ϕ_n converge pointwise to ϕ and ϕ is continuous at $t = 0$, then ϕ is the characteristic function of some μ and $\mu_n \implies \mu$.

Next, we can relate convergence in distribution in higher dimensions to that in one dimension:

Theorem 145 (Cramér–Wold)
If X_n, X are $\mathbb{R}^d$-valued random variables, then $X_n \xrightarrow{d} X$ if and only if the one-dimensional distributions converge, meaning that $\langle \theta, X_n \rangle \xrightarrow{d} \langle \theta, X \rangle$ for all $\theta \in \mathbb{R}^d$.

Proof. First suppose we know that $\langle \theta, X_n \rangle \xrightarrow{d} \langle \theta, X \rangle$ for all θ. Because $f(x) = e^{ix}$ is a bounded continuous function, we know that (applying the definition of convergence in distribution to the random variable $\langle \theta, X_n \rangle$)

$$\mathbb{E}[\exp(i\langle \theta, X_n \rangle])] \to \mathbb{E}[\exp(i\langle \theta, X \rangle)].$$

But now treating θ as a free parameter, we see that the characteristic functions ϕ_n of X_n converge pointwise to the characteristic function ϕ of X, and $\mathbb{E}[\exp(i\langle \theta, X \rangle)]$ is continuous in θ (by the same argument as Proposition 106), so by the continuity theorem we indeed have $X_n \xrightarrow{d} X$, as desired.

Similarly for the other direction, suppose $X_n \xrightarrow{d} X$. The distribution of $\langle \theta, X_n \rangle$ on $\mathbb{R}$ is the pushforward measure of μ_n under the mapping $f(x) = \langle \theta, x \rangle$, so the change of variables formula tells us that the characteristic function of $\langle \theta, X_n \rangle$ is

$$\phi_{\langle \theta, X_n \rangle}(t) = \int e^{itx} d(\mu_n \circ f^{-1}) = \int e^{it\langle \theta, X_n \rangle} d\mu_n.$$

But the right-hand side is the characteristic function of X_n evaluated at $t\theta$, and we have pointwise convergence $\chi_{X_n}(t\theta) \to \chi_X(t\theta)$ for any t because $X_n \xrightarrow{d} X$. Thus we also have pointwise convergence of the characteristic functions $\phi_{\langle \theta, X_n \rangle}$ to that of $\phi_{\langle \theta, X \rangle}$, and thus $\langle \theta, X_n \rangle \xrightarrow{d} \langle \theta, X \rangle$. □

Corollary 146 (Central limit theorem for iid sequences in $\mathbb{R}^d$)
Let $X, X^{(i)}$ be iid $\mathbb{R}^d$-valued random variables such that $\mathbb{E}(||X||^2)$ is finite. Define the **mean vector** and **covariance matrix** of X as

$$\mu = \mathbb{E}[X] \in \mathbb{R}^d, \quad \Sigma = \text{Cov}(X) = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right] \in \mathbb{R}^{d\times d}.$$

(Unpacking the notation, $\Sigma = \text{Cov}(X)$ is the $d \times d$ matrix with entries equal to the covariances $\Sigma_{ij} = \text{Cov}(X_i, X_j)$ between the different coordinates.) Then we have $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} N(\vec{0}, \Sigma)$ (that is, convergence to the multivariate normal distribution).

Proof. Much like in previous proofs, we're going to show that the characteristic functions converge, so let's start by figuring out the characteristic function for the multivariate Gaussian. Because Σ is a $d \times d$ positive semidefinite matrix, it has a **Cholesky decomposition** $\Sigma = AA^T$ for some $A \in \mathbb{R}^{d\times k}$. Then having $Y \sim N(0, \Sigma)$ is equivalent to saying that $Y \overset{d}{=} AZ$ where Z is a standard multivariate Gaussian $N(0, I_{k\times k})$ (by plugging into the definition of the covariance matrix). We thus find that the characteristic function for Y is

$$\boxed{\phi_\Sigma(\theta)} = \mathbb{E}\left[\exp(i\langle\theta, Y\rangle)\right] = \mathbb{E}\left[\exp(i\langle\theta, AZ\rangle)\right] = \mathbb{E}\left[\exp\left(i\langle A^T\theta, Z\rangle\right)\right].$$

But by definition, we know that $Z = (Z_1, Z_2, \cdots, Z_k)$ with all entries iid standard Gaussian, so the above expression simplifies to (here using that $c_1Z_1 + \cdots + c_kZ_k$ has the same distribution as $\sqrt{c_1^2 + \cdots + c_k^2}Z'$ for a standard Gaussian Z' and that the characteristic function for the standard Gaussian is $\mathbb{E}[e^{itZ'}] = e^{-t^2/2}$)

$$\exp\left(-\frac{||A^T\theta||^2}{2}\right) = \exp\left(-\frac{1}{2}(\theta^T AA^T\theta)\right) = \boxed{\exp\left(-\frac{\langle\theta, \Sigma\theta\rangle}{2}\right)}.$$

Our goal is to show that the left-hand side's characteristic function converges to this function $\phi_\Sigma(\theta)$, meaning that it suffices to show that for all $\theta \in \mathbb{R}^d$ that

$$\mathbb{E}\left(\exp\left(i\left\langle\theta, \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu)\right\rangle\right)\right) \xrightarrow{?} \phi_\Sigma(\theta).$$

But we know by the one-dimensional central limit theorem that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle\theta, X_i - \mu\rangle \xrightarrow{d} N(0, \text{Var}(\langle\theta, X\rangle)) = N(0, \langle\theta, \Sigma\theta\rangle)$$

(the last equality can be verified by expanding out the formula for variance). Thus we can take the characterstic functions of those variables and find that for any fixed θ,

$$\mathbb{E}\left(\exp\left(i\frac{1}{\sqrt{n}}\sum_i \langle\theta, X_i - \mu\rangle\right)\right) \to \exp\left(-\frac{\langle\theta, \Sigma\theta\rangle}{2}\right),$$

and thus we have convergence of the characteristic functions and therefore convergence to the desired distribution. □

This concludes the content that we're covering from chapter 3 of Durrett, and now we're going to move on to our next topic, **martingales and conditional expectation**. We'll start with a motivating discussion:

Definition 147

Suppose we're on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and suppose we have two events $A, B \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$. Then we can define the **conditional probability** $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$. (In particular, $\mathbb{P}(B|A) = \mathbb{P}(B)$ if A, B are independent.) If $\mathbb{P}(A) > 0$ and $X \in L^1$ is an integrable random variable, we can also define the **conditional expectation** $\mathbb{E}(X|A) = \frac{\mathbb{E}(X;A)}{\mathbb{P}(A)}$.

Next, suppose X, Y are both random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. For any $y \in \Omega$ with $\mathbb{P}(Y = y) > 0$, we can define $g(y) = \mathbb{E}[X|Y = y] = \frac{\mathbb{E}[X;Y=y]}{\mathbb{P}(Y=y)}$. Then if Y satisfies $\mathbb{P}(Y = y) > 0$ for all $y \in \operatorname{supp} \mathcal{L}_Y$, then we can define the **conditional expectation** (random variable) $\mathbb{E}[X|Y] = g(Y)$.

Because of the division by $\mathbb{P}(A)$ and $\mathbb{P}(Y = y)$ above, it doesn't directly make sense to condition on events of probability zero. However, there are often situations where we do want to condition on an event of that sort. For example, suppose we have random variables distributed as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, \Sigma) = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right),$$

where $\det \Sigma > 0$. In general, a Gaussian of the form $N(0, \Sigma = AA^T)$ doesn't have a density on $\mathbb{R}^d$ unless A is actually an invertible $d \times d$ matrix. But if we assume that $\det \Sigma > 0$ for the matrix above, then $\mathcal{L}_{(X,Y)}$ will be a measure on $\mathbb{R}^2$ with some density g_Σ (the details can be worked out using a calculus argument). But the point is that even though the probability that Y takes on any particular value is zero, we would still like to be able to say that

$$\mathbb{E}[X|Y = y] = \frac{\int x g_\Sigma(x, y) dx}{\int g_\Sigma(x, y) dx}.$$

So the ordinary notion of conditioning from an introductory probability class isn't quite enough, and we'll discuss this more next time!

17 November 6, 2019

Our second midterm exam will be of similar difficulty to the first one (and weighted equally). To give us more time to finish the exam, it will be scheduled between 5–9pm on December 4th, depending on our availability.

Today, we'll discuss conditional expectation and martingales. Recall that if X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}|X| < \infty$, then we can define $\mathbb{E}[X|A] = \frac{\mathbb{E}[X;A]}{\mathbb{P}(A)}$ for any event A with positive probability. In particular, if Y is another random variable with $\mathbb{P}(Y = y)$ positive, we can condition on the event $Y = y$ and write $\mathbb{E}[X|Y = y] = \frac{\mathbb{E}[X;Y=y]}{\mathbb{P}(Y=y)}$. However, this doesn't really work as generally as we'd like, because we want to define $\mathbb{P}(X|Y)$ even when $\mathbb{P}(Y = y) = 0$, particularly when Y is completely nonatomic. So the key idea is that it doesn't make sense to condition on events of measure zero, so we shouldn't think about individual events $\{Y = y\}$. Instead, we should think about Y as a whole and try to define $\mathbb{E}[X|Y]$ as a random variable. This random variable should be $\sigma(Y)$-measurable (because it should depend only on the value of Y), and it should satisfy the key property that

$$\mathbb{E}[\mathbb{E}[X|Y]h(Y)] = \mathbb{E}[Xh(Y)]$$

for a function h, as long as both sides are well-defined. Let's formalize this now:

Definition 148

Suppose X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}[|X|] < \infty$, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. We say that Y is a **version of** $\mathbb{E}[X|\mathcal{G}]$ if $Y \in \mathcal{G}$ (that is, Y is measurable with respect to $\mathcal{G}$) and $\mathbb{E}[X; A] = \mathbb{E}[Y; A]$ for all $A \in \mathcal{G}$.

It's an exercise for us to reconcile this definition with the "conditional probability" that we learned in introductory probability. We need to show that this random variable Y actually exists and that it's unique in some way. Also, the notation suggests that Y should be integrable, so we should check that as well. But eventually, we will use this to define conditional expectation via $\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y))$.

Lemma 149 (Integrability)

If Y is a version of $\mathbb{E}[X|\mathcal{G}]$, then $\mathbb{E}[|Y|] \leq \mathbb{E}[|X|] < \infty$ (so Y is integrable).

Proof. Because Y is $\mathcal{G}$-measurable, $\{Y > 0\}$ is in $\mathcal{G}$, meaning that $\mathbb{E}[Y_+] = \mathbb{E}[Y; Y > 0] = \mathbb{E}[X; Y > 0]$. Similarly, $\mathbb{E}[Y_-] = \mathbb{E}[-Y; Y < 0] = \mathbb{E}[-X; Y < 0]$. So now

$$\mathbb{E}[|Y|] = \mathbb{E}[Y_+] + \mathbb{E}[Y_-] = \mathbb{E}[X; Y > 0] + \mathbb{E}[-X; Y < 0] \leq \mathbb{E}[|X|; Y > 0] + \mathbb{E}[|X|; Y < 0] \leq \mathbb{E}[|X|],$$

as desired. □

Lemma 150 (Uniqueness)

If Y and Y' are both versions of $\mathbb{E}[X|\mathcal{G}]$, then $Y = Y'$ almost surely.

Proof. By definition, both random variables are $\mathcal{G}$-measurable, so define the event $A_\varepsilon = \{Y - Y' \geq \varepsilon\}$ (which is also $\mathcal{G}$-measurable) for any $\varepsilon > 0$. We have that

$$0 = \mathbb{E}[Y; A_\varepsilon] - \mathbb{E}[Y'; A_\varepsilon] = \mathbb{E}[Y - Y'; A_\varepsilon] \geq \varepsilon \mathbb{P}(A_\varepsilon),$$

so $\mathbb{P}(A_\varepsilon) = 0$. Taking $\varepsilon \to 0$, we conclude that $Y \leq Y'$ almost surely, and flipping the roles of Y and Y' shows that $Y' \leq Y$ almost surely as well. Thus $Y = Y'$ almost surely. □

Definition 151

Let μ and ν be measures on $(\Omega, \mathcal{F})$. We say that ν is **absolutely continuous with respect to μ** (denoted $\nu \ll \mu$) if for all $A \in \mathcal{F}$ with $\mu(A) = 0$, we also have $\nu(A) = 0$.

Theorem 152 (Radon-Nikodym)

Suppose μ, ν are σ-finite measures on $(\Omega, \mathcal{F})$ with $\nu \ll \mu$. Then there is an $\mathcal{F}$-measurable function f such that $\nu(A) = \int_A f \, d\mu$ for all $A \in \mathcal{F}$. (The function f is often written as $\frac{d\nu}{d\mu}$).

This result basically says that if ν is absolutely continuous respect to μ, then ν has a density with respect to μ. Recall that we had a similar situation in the proof of Cramér's theorem, in which we had an exponential tilting with $\frac{d\mathbb{P}_\theta}{d\mathbb{P}} = \frac{\exp(\theta X)}{m(\theta)}$ (here $\mathbb{P}_\theta$ was absolutely continuous with respect to $\mathbb{P}$).

We won't prove Radon-Nikodym right now, but we will return to it later in the course (see Theorem 215). For now, we will use it to show **existence of the conditional expectation**.

Proof of existence of conditional expectation. As in the definition, let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then the restriction of $\mathbb{P}$ to $\mathcal{G}$, which we write as $\mu = \mathbb{P}|_{\mathcal{G}}$ is a probability measure on $(\Omega, \mathcal{G})$. We can now define a measure ν on $(\Omega, \mathcal{G})$ by setting

$$\nu(A) = \mathbb{E}(X_+; A) \text{ for all } A \in \mathcal{G}.$$

This is a finite measure, because $\nu(\Omega) = \mathbb{E}[X_+] < \infty$ by assumption. Also, ν is absolutely continuous with respect to μ, because if A is an event of probability zero, then $\mathbb{E}(X_+; A) = 0$. Thus the Radon-Nikodym theorem tells us that there is a $\mathcal{G}$-measurable function $Y = \frac{d\nu}{d\mu}$ such that $\nu(A) = \int_A Y \, d\mu$ for all $A \in \mathcal{G}$.

But now notice that the left-hand side is $\nu(A) = \mathbb{E}[X_+; A]$, and the right hand side is $\int_A Y \, d\mu = \mathbb{E}[Y; A]$. Thus, Y is a version of $\mathbb{E}[X_+|\mathcal{G}]$. Similarly constructing a version of $\mathbb{E}[X_-|\mathcal{G}]$ and then subtracting the two random variables gives us the desired conditional expectation. □

We should read the textbook for some basic properties of the conditional expectation – in particular, it's linear and monotone, it satisfies a version of Jensen's inequality, and so on. But there's a few properties that are important to know which are not analogous to the ordinary expectation:

Proposition 153
Let $\mathcal{G}^{\text{small}}$ be a sub-σ-algebra of $\mathcal{G}$ (which is a sub-σ-algebra of $\mathcal{F}$). If $\mathbb{E}[X|\mathcal{G}] \in \mathcal{G}^{\text{small}}$, then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}^{\text{small}}]$.

Proof. Let $Y = \mathbb{E}[X|\mathcal{G}]$; we wish to show that Y is a version of $\mathbb{E}[X|\mathcal{G}^{\text{small}}]$. By assumption, Y is measurable with respect to $\mathcal{G}^{\text{small}}$, so the first condition is satisfied. Also, $\mathbb{E}[Y; A] = \mathbb{E}[X; A]$ for any $A \in \mathcal{G}$ by definition of Y, so in particular it holds for all $A \in \mathcal{G}^{\text{small}}$, verifying the second condition. □

Proposition 154 (Tower property)
Suppose we have sub-σ-algebras $\mathcal{G}^{\text{small}} \subseteq \mathcal{G} \subseteq \mathcal{F}$ as before. Then

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}^{\text{small}}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}^{\text{small}}] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{G}^{\text{small}}].$$

Proof. The proof is very similar to the one above. For the first equality, let $Y = \mathbb{E}[X|\mathcal{G}^{\text{small}}]$; we wish to show that Y is a version of $\mathbb{E}[Y|\mathcal{G}]$. Indeed, it is measurable with respect to $\mathcal{G}^{\text{small}}$ and thus $\mathcal{G}$, and the other condition is asking us to show that $\mathbb{E}[Y; A] = \mathbb{E}[Y; A]$ for $A \in \mathcal{G}$, which is true. For the second equality, define $Z = \mathbb{E}[X|\mathcal{G}]$; we wish to show that Y is a version of $\mathbb{E}[Z|\mathcal{G}^{\text{small}}]$. Again Y is measurable with respect to $\mathcal{G}^{\text{small}}$, and the other condition asks us to show that $\mathbb{E}[Y; A] = \mathbb{E}[Z; A]$ for any $A \in \mathcal{G}^{\text{small}}$, which is true because both of these are equal to $\mathbb{E}[X; A]$ by the definitions of Y and Z. □

Proposition 155
Let X and Y be random variables such that $\mathbb{E}[|Y|]$ and $\mathbb{E}[|XY|]$ are finite and $X \in \mathcal{G}$. Then

$$\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}].$$

In other words, because X is a known constant with respect to $\mathcal{G}$, we can pull it out of the conditional expectation.

Proof. First, we show that the equality holds if $X = 1_B$ for some $B \in \mathcal{G}$, and we'll do this by showing that the right-hand side is a version of $\mathbb{E}[XY|\mathcal{G}]$. It's measurable because both X and $\mathbb{E}[Y|\mathcal{G}]$ are $\mathcal{G}$-measurable, and the product of

two measurable functions is also measurable. For the other condition, we must check that $\mathbb{E}[XY;A] = \mathbb{E}[X\mathbb{E}[Y|\mathcal{G}];A]$ for all $A \in \mathcal{G}$, and this is true because

$$\boxed{\mathbb{E}[XY;A]} = \mathbb{E}[1_B Y;A] = \mathbb{E}[Y;A\cap B] = \mathbb{E}[\mathbb{E}[Y|\mathcal{G}];A\cap B]$$

(in the last step we use that $A\cap B \in \mathcal{G}$ because $A, B \in \mathcal{G}$), which simplifies to $\mathbb{E}[1_B\mathbb{E}[Y|\mathcal{G}];A] = \boxed{\mathbb{E}[X\mathbb{E}[Y|\mathcal{G}];A]}$, as desired. Finally, by linearity of conditional expectation, this identity then holds for simple functions and thus general measurable functions. □

One last note is that we can sometimes interpret conditional expectation as an **orthogonal projection**. To set that up, notice that if $\mathcal{G} \subseteq \mathcal{F}$, then the space of random variables $L^2(\mathcal{G})$ is a subspace of $L^2(\mathcal{F})$.

Proposition 156

For any random variable $X \in L^2(\mathcal{F})$, the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ is an orthogonal projection of X onto $L^2(\mathcal{G})$.

Proof. First of all, conditional Jensen's inequality tells us that $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]^2] \leq \mathbb{E}[\mathbb{E}[X^2|\mathcal{G}]] = \mathbb{E}[X^2]$ (which is finite because $X \in L^2(\mathcal{F})$), so $\mathbb{E}[X|\mathcal{G}]$ is integrable and thus indeed in $L^2(\mathcal{G})$. To show orthogonality, we need to show that the difference between X and $\mathbb{E}(X|\mathcal{G})$ is orthogonal to any element of $L^2(\mathcal{G})$. In other words, we must show that for any $Z \in L^2(\mathcal{G})$, we have $\mathbb{E}[Z(X - \mathbb{E}[X|\mathcal{G}])] = 0$. By linearity of expectation, we can write the left-hand side as $\mathbb{E}[XZ] - \mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]]$. But since $Z \in L^2(\mathcal{G})$ is $\mathcal{G}$-measurable, we have by Proposition 155 and then Proposition 154 that

$$\mathbb{E}[XZ] - \mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[XZ] - \mathbb{E}[\mathbb{E}[XZ|\mathcal{G}]] = \mathbb{E}[XZ] - \mathbb{E}[XZ] = 0,$$

since an ordinary expectation is like conditioning on the trivial σ-algebra. □

There's one special topic in our textbook about regular probability distributions, which we'll skip for now. Instead, we'll move on to **martingales**, which is a topic with many interesting applications. Throughout the following discussion, we're working on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 157

A **filtration** is a sequence $(\mathcal{F}_n)_{n\geq 0}$ of σ-fields such that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$. A sequence of random variables $(X_n)_{n\geq 0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is **adapted** to a filtration $(\mathcal{F}_n)_{n\geq 0}$ if $X_n \in \mathcal{F}_n$ for all n. The **natural filtration** of the sequence (X_n) is defined by setting $\mathcal{F}_n = \sigma(X_1, \cdots, X_n)$.

(Notice that the definition of a filtration doesn't involve any probability – it's just a statement about families of sets.) In words, we "gradually reveal more information" at each step in a filtration.

Definition 158

Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$. We say that $(X_n)_{n\geq 0}$ is a **martingale** with respect to $\mathcal{F}_n$ if X_n is adapted to $\mathcal{F}_n$, $\mathbb{E}[|X_n|] < \infty$ for all n (though we do not need to have a uniform bound across all n), and $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$.

In other words, given everything we know about the random variables up to the nth step, the next step's expected value is equal to the current value.

Definition 159

In the last part of the definition above, $(X_n)_{n\geq 0}$ is a **supermartingale** if we only have $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$, and $(X_n)_{n\geq 0}$ is a **submartingale** if we only have $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n$.

(For notational convenience from here on, we will write that X_n is a martingale instead of (X_n).)

Example 160

Let X_n be a simple random walk on $\mathbb{Z}$, meaning that we start at $X_0 = 0$ and define $X_n = \sum_{i=1}^n Y_i$, where Y_i are iid and are each 1 or -1 with probability $\frac{1}{2}$ each. This sequence of random variables is a martingale, because each step has probability $\frac{1}{2}$ of adding 1 to X_n and probability $\frac{1}{2}$ of subtracting 1 (so $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \frac{1}{2}(X_n+1)+\frac{1}{2}(X_n-1) = X_n$.

The first observation we can make about martingales in general is that

$$\mathbb{E}[X_n] = \mathbb{E}\left[\mathbb{E}[X_n|\mathcal{F}_{n-1}]\right] = \mathbb{E}[X_{n-1}],$$

where we've applied the tower property (Proposition 154) and then the martingale property. Applying this iteratively tells us that

$$\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}] = \cdots = \mathbb{E}[X_0].$$

But we'll now show that we also have $\mathbb{E}[X_0] = \mathbb{E}[X_\tau]$ for certain random times τ, and this will have many applications.

Definition 161

Let τ be a nonnegative integer-valued random variable. We say that τ is a **stopping time** with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$ if $\{\tau = n\} \in \mathcal{F}_n$ for all n.

The intuitive reason for the name "stopping time" is that we can think of a martingale as modeling our wealth in a stock market, and we can only choose to stop our random process at time n (and hence deciding that we're in the event $\{\tau = n\}$) based on the information that what we already know so far (which is the σ-algebra $\mathcal{F}_n$). The statement $\mathbb{E}[X_0] = \mathbb{E}[X_\tau]$ is then saying that we "cannot make money off of the stock market," but things aren't quite so simple:

Example 162

Continuing our example from above, let X_n be a simple random walk on $\mathbb{Z}$ started at $X_0 = 0$, and define

$$\tau = \inf\{n : X_n = 1\}.$$

Then τ is a valid stopping time, because it's a nonnegative integer-valued random variable, it's finite almost surely (exercise), and at time n the event $\{\tau = n\}$ is measurable with respect to what we already know (it's the same as the event $X_n = 1$). But notice that $0 = \mathbb{E}[X_0] \neq \mathbb{E}[X_\tau] = 1$.

On the other hand, it does make sense to expect that $\mathbb{E}[X_0] = \mathbb{E}[X_\tau]$. The key observation is that if X_n is a martingale and τ is a stopping time, then the **stopped process** defined by

$$Y_n = X_{n\wedge\tau}, \qquad \text{where } n\wedge\tau = \min(n,\tau),$$

is also a martingale. Indeed, $Y_n = X_n$ unless we have "already stopped" (meaning $\tau < n$), so we can write

$$Y_n = \sum_{i=1}^{n-1} (1\{\tau = i\}X_i) + 1\{\tau \geq n\}X_n.$$

We can check that Y_n is integrable for any n (exercise). From there, notice that the first sum is measurable with respect to $\mathcal{F}_{n-1}$, and $1\{\tau \geq n\} = 1 - \sum_{i=0}^{n-1} 1\{\tau = i\}$ is also in $\mathcal{F}_{n-1}$. So conditioning the equation in $\mathcal{F}_{n-1}$, we get

$$\mathbb{E}[Y_n|\mathcal{F}_{n-1}] = \sum_{i=1}^{n-1} 1\{\tau = i\}X_i + 1\{\tau \geq n\}\mathbb{E}[X_n|\mathcal{F}_{n-1}].$$

But now by the martingale property of X, the conditional expectation on the right is X_{n-1}, so the entire right-hand side reduces to Y_{n-1} and **the stopped process is also a martingale**! In particular, the earlier calculation shows that $\mathbb{E}[Y_0] = \mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \cdots$, meaning that

$$\mathbb{E}[Y_n] = \mathbb{E}[Y_0] \implies \mathbb{E}[X_0] = \mathbb{E}[X_{n\wedge\tau}].$$

But $n \wedge \tau \to \tau$ almost surely (in other words, because τ is finite almost surely, there is some $n(\omega)$ such that $\tau(\omega) \wedge n = \tau(\omega)$ for all sufficiently large $n \geq n(\omega)$). This means that $X_{n\wedge\tau}$ converges to X_τ almost surely, so it's reasonable to expect that with sufficient integrability conditions, we also have convergence of expectation and thus $\mathbb{E}[X_0] = \mathbb{E}[X_\tau]$. So a large part of this section of the class is figuring out conditions under which we have $\mathbb{E}[X_{n\wedge\tau}] \to \mathbb{E}[X_\tau]$. We'll continue this discusion next time, but for now we'll describe a nice application:

Example 163

Let $G = (V, E)$ be a finite connected graph, and let $Z \subseteq V$ be its "boundary." Let $f : Z \to \mathbb{R}$ be a function (a "boundary condition,"), and suppose we perform a simple random walk on the graph G stopped at the boundary (meaning that we travel along the edges and always move to a random neighbor of v at each step). Let τ be the first time we hit the boundary Z, and define

$$M_n = \mathbb{E}[f(X_\tau)|\mathcal{F}_n],$$

where $\mathcal{F}_n$ is the filtration $\sigma(X_0, \cdots, X_n)$ of the walk. Then M_n is a martingale (exercise), and its value depends only on the current position X_n. It turns out that $M_n = h(X_n)$ is the **harmonic interpolation** of f (which has the property that the value at a vertex of the graph is the average of the neighboring values). And as an extension, if we replace "harmonic function" with "subharmonic function" in $M_n = h(X_n)$, then we'll get a submartingale instead of a martingale.

18 November 13, 2019

Last time, we defined martingales: given a **filtration** on a probability space, which is a nested sequence of σ-fields $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$, a **martingale** is a sequence of integrable random variables $(X_n)_{n\geq 0}$ where $X_n \in \mathcal{F}_n$ (the sequence is **adapted** to the filtration), and the martingale property $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$ is satisfied. (If we replace these conditions with an inequality $\geq$ or $\leq$, we get a submartingale or supermartingale, respectively.)

Remark 164. *Last time, we showed that $\mathbb{E}[X_0] = \mathbb{E}[X_1] = \cdots$ for a martingale. By extension, if X_n is a submartingale, then $\mathbb{E}X_0 \leq \mathbb{E}X_1 \leq \cdots$, and we have the reverse inequality for a supermartingale.*

We'll prove some martingale convergence theorems today, but we'll start with some preliminary facts:

Lemma 165
Let X_n be a martingale, and let ϕ be a convex function such that $\phi(X_n) \in L^1$ for all n. Then $\phi(X_n)$ is a submartingale.

As we'll see in the proof, we can replace "convex" with "concave" and we'll get a supermartingale instead.

Proof. Convex functions are measurable and $X_n \in \mathcal{F}_n$, so $\phi(X_n)$ will be measurable with respect to $\mathcal{F}_n$. Thus, it suffices to check the submartingale condition, and indeed by conditional Jensen's inequality we have

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] \geq \phi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) = \phi(X_n),$$

as desired. □

We can also generalize this to submartingales:

Lemma 166
Let X_n be a submartingale, and let ϕ be a nondecreasing convex function such that $\phi(X_n) \in L^1$ for all n. Then $\phi(X_n)$ is a submartingale.

(Again, it is valid to replace "submartingale" and "convex" with "supermartingale" and "concave," respectively.)

Proof. This is almost identical to the previous proof – again $\phi(X_n)$ is measurable with respect to $\mathcal{F}_n$, and we need to check the submartingale condition again. This time, we have

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] \geq \phi[\mathbb{E}(X_{n+1}|\mathcal{F}_n)] \geq \phi(X_n),$$

where we've used conditional Jensen's in the first equality and both the submartingale condition and ϕ being nondecreasing in the second equality. □

Definition 167
A sequence of random variables $(H_n)_{n\geq 0}$ is **predictable** (also **previsible**) with respect to a filtration $(\mathcal{F}_n)$ if $H_n \in \mathcal{F}_{n-1}$ for all n. If $X = (X_n)$ is another sequence of random variables, then the ***H*-transform** of X is defined via $(H \cdot X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1})$.

To understand this definition, suppose we play a betting game, where at each positive integer time k betting 1 dollar means that we make net winnings of $\Delta X_k = X_k - X_{k-1}$ dollars (meaning that we win $1 + \Delta X_k$ dollars from the game). Betting exactly one dollar at each time gives us net winnings of $\sum_{k=1}^{n} \Delta X_k = X_n - X_0$ (by a telescoping sum), but more generally we can choose to bet different amounts of money each time. If we instead bet H_k dollars at time k, we do indeed make net winnings of $(H \cdot X)_n$ at time n.

The reason that we require $H_n \in \mathcal{F}_{n-1}$ is then because the amount we bet at time n should not depend on what we win at time n, only what has happened up to time $(n-1)$ (this also motivates the name "previsible"). And if we're in a situation where $\mathbb{E}[\Delta X_k|\mathcal{F}_{k-1}] \leq 0$ for all k (because casinos tend to make the player lose in expectation), then $(H \cdot X)_n$ will be a supermartingale.

Example 168

Recall that a **stopping time** τ is a nonegative integer-valued random variable with $\{\tau = n\} \in \mathcal{F}_n$ for all n. If we define $H_n = 1\{\tau \geq n\}$ (in other words, betting only if τ didn't happen yet), notice that $H_n = 1 - \sum_{\ell<n} 1\{\tau = \ell\} \in \mathcal{F}_{n-1}$, so H is previsible. We can then check that $(H \cdot X)_n = X_{n\wedge\tau} - X_0$.

Lemma 169

Let X_n be a submartingale, and let H_n be nonnegative and previsible. Suppose that $H_n \in L^\infty$ for all n, meaning that each H_n is uniformly bounded by a (potentially different) constant almost surely. Then $(H \cdot X)_n$ is a submartingale as well.

Again, the analogous statement holds replacing "submartingales" with "supermartingales."

Proof. Let $Y_n = (H \cdot X)_n = \sum_{k=1}^n H_k(X_k - X_{k-1})$; we want to show that Y_n is a submartingale. First of all, Y_n is integrable, because it's the sum of finitely many terms which are each integrable (because each H_k is bounded by a constant, and each X_k is integrable because X is a submartingale). Also, $Y_n \in \mathcal{F}_n$ because all of the individual terms are in $\mathcal{F}_n$. So it suffices to show the submartingale condition, which can be written as

$$\mathbb{E}[Y_{n+1}|\mathcal{F}_n] \overset{?}{\geq} Y_n \iff \mathbb{E}[Y_{n+1} - Y_n|\mathcal{F}_n] \overset{?}{\geq} 0,$$

since $\mathbb{E}[Y_n|\mathcal{F}_n] = Y_n$ because Y_n is $\mathcal{F}_n$-measurable. And this indeed holds: notice that

$$\mathbb{E}[Y_{n+1} - Y_n|\mathcal{F}_n] = \mathbb{E}[H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n] = H_{n+1}\mathbb{E}[(X_{n+1} - X_n)|\mathcal{F}_n]$$

(where we use that H_{n+1} is measurable with respect to $\mathcal{F}_n$), and this conditional expectation is nonnegative because X_n is a submartingale (and H_{n+1} is nonnegative), as desired. □

Corollary 170

If X_n is a submartingale and τ is a stopping time, then $Y_n = X_{n\wedge\tau}$ is also a submartingale (because the stopped process can be expressed as an H-transform by Example 168).

One goal of today's class (which we should keep in mind) is to prove that any nonnegative supermartingale X_n converges almost surely to some integrable limit X_∞ with $\mathbb{E}[X_\infty] \leq \mathbb{E}[X_0]$. Because $\mathbb{E}[X_0] \geq \mathbb{E}[X_1] \geq \cdots$ for a supermartingale, but all expectations are bounded from below by 0, this fact shouldn't be too surprising. But the main technical component of this proof is to show a bound on how often a submartingale can "go up," which is equivalent to a bound on how often supermartingales can go down.

Definition 171

Fix constants $-\infty < a < b < \infty$ and let X_n be a martingale. Define $\tau_0 = -1$, and for all positive integers k, define $\tau_{2k-1} = \inf\{n > \tau_{2k-2} : X_n \leq a\}$ and $\tau_{2k} = \inf\{n > \tau_{2k-1} : X_n \geq b\}$.

In other words, the τ_is are the times where X_n goes below a, then above b, then below a, and so on. (And if X never goes below a, we have $\tau_1 = \infty$.) By definition, we have $\tau_0 < \tau_1 < \tau_2 < \cdots$, and we want to keep track of the ranges $(\tau_{2k-1}, \tau_{2k}]$ – these are known as **upcrossings** because we're going from a up to b. For any n, the number of upcrossings $U_n(a, b)$ of $[a, b]$ completed by time n is then $U_n(a, b) = \sup\{k : \tau_{2k} \geq n\}$.

Theorem 172 (Doob's upcrossing inequality)

Let X_n be a submartingale. Then for any $-\infty < a < b < \infty$,

$$\mathbb{E}[U_n(a,b)] \leq \frac{\mathbb{E}[(X_n - a)_+ - (X_0 - a)_+]}{b-a} \leq \frac{\mathbb{E}[(X_n - a)_+]}{b-a}.$$

Proof. Let H_n be the indicator variable for time n being inside an upcrossing (in other words, $H_n = 1$ if $n \in (\tau_{2k-1}, \tau_{2k}]$ for some k). Equivalently, we may write

$$H_n = \sum_{k \geq 1} 1\{n \in (\tau_{2k-1}, \tau_{2k}]\} = \sum_{k \geq 1} 1\{n \leq \tau_{2k}\} - 1\{n \leq \tau_{2k-1}\}$$

because the different upcrossings are disjoint. (If we're worried about having an infinite sum, notice that we really only need to sum up to n because there can't be more than n upcrossings by time n.) But because each τ_i is a stopping time, each indicator $1\{n \leq \tau_i\}$ is measurable with respect to $\mathcal{F}_{n-1}$ (again by the same logic as Example 168), so H_n is previsible. (Intuitively, we know at time $(n-1)$ whether we're in an upcrossing at time n, because the upcrossing interval includes τ_{2k}.) Since H_n is bounded between 0 and 1 almost surely, we can also define $K_n = 1 - H_n$, and then Lemma 169 tells us that $(H \cdot X)_n$ and $(K \cdot X)_n$ are both submartingales.

The idea now is that betting during each upcrossing corresponds to an increase in X of $(b-a)$, but there is a minor problem: we may start an upcrossing and then have X become arbitrarily negative, which is bad for our bound. So we will define

$$Y_n = \max(a, X_n) = a + (X_n - a)_+,$$

which is also a submartingale by Lemma 166 because $\max(a, x)$ is a convex function. Lemma 169 still applies, so $(H \cdot Y)_n$ and $(K \cdot Y)_n$ are both submartingales as well. Now notice that

$$(H \cdot Y)_n \geq (b-a) U_n(a,b),$$

because we can interpret this H-transform as "only betting during the upcrossings of Y," in which we win the increment $(b-a)$ from each upcrossing (because at the start of each upcrossing we have $X_n \leq a$, so $Y_n = a$, and at the end of each upcrossing we have $X_n \geq b$, so $Y_n \geq b$) and may even win an extra amount at the end. (Here is where using Y instead of X is important – we wouldn't necessarily know that the extra amount won is nonnegative with X.) Taking expectations on both sides, we thus have

$$\boxed{\mathbb{E}[U_n(a,b)] \leq \frac{\mathbb{E}[H \cdot Y]_n}{b-a}},$$

and we just want to upper bound this last quantity in terms of X. By definition, we know that

$$Y_n - Y_0 = (1 \cdot Y)_n = (H \cdot Y)_n + (K \cdot Y)_n,$$

and rearranging and taking expectations tells us that

$$\mathbb{E}[(H \cdot Y)_n] = \mathbb{E}[Y_n - Y_0 - (K \cdot Y)_n].$$

But $Y_n - Y_0 = (X_n - a)_+ - (X_0 - a)_+$ by definition, and $(K \cdot Y)_n$ is a submartingale started at zero, so $\mathbb{E}(K \cdot Y)_n \geq 0$. Plugging both of these into the boxed expression gives us the desired result. □

Intuitively, the key inequality here is that $\mathbb{E}[(K \cdot Y)_n] \geq 0$. In words, it's hard to have many upcrossings because it's hard for the submartingale $(K \cdot Y)_n$ (the amount of money we make betting outside of upcrossings) to be very

negative.

Theorem 173 (Submartingale convergence theorem)
If X_n is a submartingale with $\sup_n \mathbb{E}[(X_n)_+] < \infty$, then X_n converges almost surely to some $X_\infty \in L^1$.

Proof. For any real numbers $a, x \in \mathbb{R}$, we have $(x-a)_+ \leq x_+ + |a|$, so Theorem 172 also tells us that

$$\mathbb{E}[U_n(a,b)] \leq \frac{\mathbb{E}[(X_n)_+ + |a|]}{b-a}.$$

By assumption, $\mathbb{E}[(X_n)_+]$ is uniformly bounded in n, so the right-hand side is some finite value and thus the expected number of upcrossings across $[a,b]$ is uniformly bounded. By the bounded convergence theorem, the limit $\lim_{n\to\infty} U_n(a,b)$ will be some random variable $U_\infty(a,b)$ with finite expectation, and any random variable with finite expectation is finite almost surely. This means that $U_\infty(a,b)$ will be finite for **all** rational $-\infty < a < b < \infty$ almost surely (by countable subadditivity). This means that our sequence $X_1, X_2, \cdots$ does not cross any rational interval infinitely many times, so $X_\infty = \lim_{n\to\infty} X_n$ indeed exists almost surely.

It remains to show that X_∞ is finite almost surely and that it is integrable. By Fatou's lemma (because $(X_n)_+$ converges to $(X_\infty)_+$), we have

$$\mathbb{E}[(X_\infty)_+] \leq \liminf_{n\to\infty} \mathbb{E}[(X_n)_+] < \infty$$

(by assumption), so $(X_\infty)_+$ does have finite expectation. Fatou's lemma also tells us that $\mathbb{E}[(X_\infty)_-] \leq \liminf_{n\to\infty} \mathbb{E}[(X_n)_-]$, but we haven't stated explicitly in our assumptions that this quantity is finite. However, we can write

$$\liminf_{n\to\infty} \mathbb{E}[(X_n)_-] = \liminf_{n\to\infty} \mathbb{E}[(X_n)_+ - X_n].$$

Now $\mathbb{E}[(X_n)_+]$ is bounded, and because X_n is a submartingale we know that $\mathbb{E}[-X_n] \leq \mathbb{E}[-X_0]$. Thus we do indeed have

$$\mathbb{E}[(X_\infty)_-] \leq \liminf_{n\to\infty} \mathbb{E}[(X_n)_-] \leq \liminf_{n\to\infty} \mathbb{E}[(X_n)_+] - \mathbb{E}[X_0] < \infty.$$

So X_∞ is in L^1 because both its positive and negative parts have finite expectation, as desired (which also implies that it is finite almost surely). □

Theorem 174
If X_n is a nonnegative supermartingale, then it converges almost surely to some X_∞ with $\mathbb{E}[X_\infty] \leq \mathbb{E}[X_0]$.

Proof. Since $Y_n = -X_n$ is a submartingale, and $(Y_n)_+ = 0$ almost surely for all n (in particular, $\mathbb{E}[(Y_n)_+]$ is uniformly bounded), the submartingale convergence theorem tells us that we have almost-sure convergence $Y_n \to Y_\infty$, so X_n converges almost surely to $X_\infty = -Y_\infty$ Since X_ns are all nonnegative, so is X_∞, and Fatou's lemma then tells us that

$$\mathbb{E}[X_\infty] \leq \liminf_{n\to\infty} \mathbb{E}[X_n] \leq \mathbb{E}[X_0],$$

where the last inequality comes from X_n being a supermartingale (so $\mathbb{E}[X_0] \geq \mathbb{E}[X_1] \geq \cdots$). □

However, we should be careful – it is not true in general that we will have $\mathbb{E}[X_n] \to \mathbb{E}[X_\infty]$:

Example 175

Let X_n be a simple random walk on $\mathbb{Z}$, and let $X_0 = 1$ and $\tau = \inf\{n \geq 0 : X_n = 0\}$. Then $Y_n = X_{n\wedge\tau}$ is a nonnegative martingale (which is also a supermartingale), so it converges almost surely. Specifically, it converges to $Y_\infty = X_\tau = 0$, so we have $\mathbb{E}[Y_\infty] = 0$ but $\mathbb{E}[Y_n] = 1$ for any n.

To conclude the lecture, we'll see an application of the convergence theorem:

Example 176 (Branching process)

Let $X_{n,i}$ be an array of iid nonnegative integer random variables that have the same distribution as X. The law of X is some probability measure p supported on the nonnegative integers called the **offspring law**. In a **Galton-Watson tree**, we start with a root vertex which gets a random number of children (possibly zero) $X_{1,1}$, which form the first level of the tree. From there, the next level is formed by giving the first child $X_{2,1}$ children, the second child $X_{2,2}$ children, and so on.

This tree may be finite or infinite, and the **Galton-Watson process** Z_n is a summary of this tree, where Z_n denotes the number of vertices at level n. If we now define

$$\mathcal{F}_n = \sigma\left(X_{\ell,i} : i \geq 1, 1 \leq \ell \leq n\right),$$

then the process Z_n is adapted to the filtration $\mathcal{F}_n$ (because we only need the $X_{n,i}$s up to level n to determine the number of children at level n). And in fact, $\mathcal{F}_n$ encodes more than just $\sigma(Z_0, Z_1, \cdots, Z_n)$, because the Z_is only tell us the total population in each generation and not which parents they come from. One way to write this process more mathematically is that

$$Z_0 = 1, \quad Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i}.$$

In particular, if $Z_{n-1} = 0$ (there is no population at level $(n-1)$), then $Z_n = 0$ (all future populations are dead). More generally,

$$\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = Z_n\mathbb{E}[X],$$

because each of the Z_n children generate $\mathbb{E}[X]$ children on average. So this means that $\frac{Z_n}{\mathbb{E}[X]^n}$ is a nonnegative martingale, and next time, we'll use this fact to study what happens to the random tree as n gets large.

19 November 18, 2019

As a reminder, the next homework assignment is due on Wednesday (there are two extra problems from the original assignment). Last time, we proved the submartingale convergence theorem: if $\mathbb{E}[X_n)_+]$ is uniformly bounded, then X_n converges almost surely to some finite random variable X_∞. In particular, this implies that any nonnegative supermartingale X_n converges almost surely to some integrable limit X_∞. However, neither result implies that $\mathbb{E}[X_n] \to \mathbb{E}[X_\infty]$ converges, and it is useful to know when that statement is true. For example, if $X_n = M_{n\wedge\tau}$ is a stopped process, then we want to find conditions under which $\mathbb{E}[M_{n\wedge\tau}] \to \mathbb{E}[M_\tau]$ (in particular, this would imply that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$).

So the main result we'll be proving today is the L^p martingale convergence theorem, which states that if X_n is a martingale with $\sup_n ||X_n||_p < \infty$ for some $1 < p < \infty$, then $X_n \to X_\infty$ almost surely and in L^p. It will turn out that this also implies that $X_n \to X_\infty$ in L^1, so $\mathbb{E}[X_n] \to \mathbb{E}[X_\infty]$ (which is what we actually care about). And notably, this result would not be true if we used $p = 1$ instead.

Remark 177. *Last lecture, we mentioned that many results that hold for submartingales also hold for supermartingales (with reversed inequalities and assumptions, and thus with equality for martingales). This will hold for many of the results we prove in the next few lectures, as we can see in the proofs.*

Lemma 178 (Doob's maximal inequality)
Let X_n be a submartingale, and define the nondecreasing process $\overline{X_n} = \max\{(X_i)_+ : 0 \le i \le n\}$. Then for any $\lambda > 0$,

$$\lambda \mathbb{P}\left(\max_{0\le i\le n} X_i \ge \lambda\right) = \lambda \mathbb{P}(\overline{X_n} \ge \lambda) \le \mathbb{E}(X_n; \overline{X_n} \ge \lambda) \le \mathbb{E}[(X_n)_+; \overline{X_n} \ge \lambda] \le \mathbb{E}[(X_n)_+].$$

Proof. The first equality is the definition of $\overline{X}_n$, the second inequality holds because $X_n \le (X_n)_+$, and the last inequality is true because $(X_n)_+$ is always nonnegative (so only counting the expectation on $\{\overline{X}_n \ge \lambda\}$ can only decrease the expectation). Thus, we only need to check the first inequality (which would be Markov's inequality if X_n were replaced with $\overline{X_n}$.

First of all, because X_n is a submartingale, if τ is a stopping time with $0 \le \tau \le k$ almost surely, then $\mathbb{E}[X_0] \le \mathbb{E}[X_\tau] \le \mathbb{E}[X_k]$ (exercise). So defining $\sigma = \inf\{i \ge 0 : X_i \ge \lambda\}$ and $\tau = \sigma \wedge n$ (so that $0 \le \tau \le n$), we have $\mathbb{E}[X_\tau] \le \mathbb{E}[X_n]$. But we can only have $X_\tau \ne X_n$ if $\sigma < n$ (meaning that we've already hit at least λ before n), which means that $\overline{X_n} \ge \lambda$ whenever $X_\tau \ne X_n$. This means that we also have

$$\mathbb{E}[X_n; \overline{X_n} \ge \lambda] \ge \mathbb{E}[X_\tau; \overline{X_n} \ge \lambda] \ge \lambda \mathbb{P}(\overline{X_n} \ge \lambda)$$

(first step because $\mathbb{E}[X_\tau] \le \mathbb{E}[X_n]$ and outside of the event $\overline{X_n} \ge \lambda$ the two variables are equal, and second step because $X_\tau \ge \lambda$ whenever $\overline{X_n} \ge \lambda$). This shows the desired result. □

Corollary 179 (Kolmogorov maximal inequality)
Let M_n be a martingale with $M_0 = 0$, and let $X_n = M_n^2$ (which is a submartingale because $f(x) = x^2$ is convex). Then applying Lemma 178 to X_n, we have

$$\mathbb{P}\left(\max_{0\le i\le n} |M_i| \ge x\right) = \mathbb{P}\left(\max_{0\le i\le n} |X_i| \ge x^2\right) \le \frac{\mathbb{E}[(X_n)_+]}{x^2} = \frac{\mathrm{Var}(M_n)}{x^2},$$

where we've used that $\mathbb{E}[M_n] = \mathbb{E}[M_0] = 0$.

This result may look similar to Chebyshev's inequality, which tells us directly that

$$\mathbb{P}(|M_n| \ge x) \le \frac{\mathrm{Var}(M_n)}{x^2},$$

but this version is stronger because we're able to bound multiple M_is at once.

Theorem 180 (L^p maximal inequality)
Let X_n be a submartingale, and define $\overline{X_n} = \max\{(X_i)_+ : 0 \le i \le n\}$. Then for all $p \in (1, \infty)$, we have

$$\left|\left|\overline{X_n}\right|\right|_p \le \frac{p}{p-1} \left|\left|(X_n)_+\right|\right|_p.$$

In other words, the maximum of the first n values of the submartingale can be controlled by just the final one. (However, note that this is really only a statement about the positive part of X_n – it doesn't prohibit X_n from being very negative.)

Proof. One annoying detail is that we don't know in advance that $\overline{X_n}$ is in L^p, so we will truncate and work with $\overline{X_n} \wedge M$ for some finite constant M. We claim that

$$\lambda \mathbb{P}(\overline{X_n} \wedge M \geq \lambda) \leq \mathbb{E}[(X_n)_+; \overline{X_n} \wedge M \geq \lambda].$$

Indeed, this is trivially true when $M < \lambda$ (because both sides are zero), and otherwise this reduces to Lemma 178. Thus we have (with our usual integration trick)

$$\mathbb{E}\left[(\overline{X_n} \wedge M)^p\right] = \int_0^\infty p y^{p-1} \mathbb{P}(\overline{X_n} \wedge M \geq y) dy \leq \int_0^\infty p y^{p-2} \mathbb{E}\left[(X_n)_+; \overline{X_n} \wedge M \geq \lambda\right].$$

By Tonelli's theorem (everything in the integrand is nonnegative), we may swap the order of integration, and the right-hand side becomes

$$\mathbb{E}\left[(X_n)_+ \int_0^\infty p y^{p-2} 1\{\overline{X_n} \wedge M \geq y\} dy\right] = \mathbb{E}\left[(X_n)_+ \cdot (\overline{X_n} \wedge M)^{p-1} \cdot \frac{p}{p-1}\right].$$

So by Hölder's inequality applied to $(X_n)_+$ and $(\overline{X}_n \wedge M)^{p-1}$ (where $\frac{1}{p} + \frac{1}{p'} = 1 \implies p' = \frac{p}{p-1}$), we have

$$\mathbb{E}\left[(\overline{X_n} \wedge M)^p\right] \leq \frac{p}{p-1} ||(X_n)_+||_p \left|\left|(\overline{X_n} \wedge M)^{p-1}\right|\right|_{p'} = \frac{p}{p-1} ||(X_n)_+||_p \left|\left|\overline{X_n} \wedge M\right|\right|_p^{p-1}.$$

Since the left-hand side is $\left|\left|\overline{X}_n \wedge M\right|\right|_p^p$, rearranging this inequality yields

$$||\overline{X_n} \wedge M||_p \leq \frac{p}{p-1} ||(X_n)_+||_p.$$

Now if the right-hand side is infinite, the theorem is automatically true, and otherwise we can take $M \to \infty$ and use the monotone convergence theorem to finish. □

Theorem 181 (L^p martingale convergence theorem)
Let X_n be a martingale. If there is some $p \in (1, \infty)$ such that $\sup_n ||X_n||_p < \infty$, then $X_n \to X_\infty$ almost surely and in L^p.

Proof. We have $\sup_n \mathbb{E}[(X_n)_+] \leq \sup_n \mathbb{E}[|X_n|] < \infty$, because $||X_n||_p$ is uniformly bounded and $||X_n||_1 \leq ||X_n||_p$ (by Corollary 75). So X_n satisfies the conditions of the submartingale convergence theorem (Theorem 173), and thus $X_n \to X_\infty$ almost surely.

To show convergence in L^p, we apply the L^p maximal inequality, which tells us that

$$\left|\left|\max_{1 \leq i \leq n} (X_i)_+\right|\right|_p \leq \frac{p}{p-1} ||(X_n)_+||_p,$$

and also that

$$\left|\left|\max_{1 \leq i \leq n} (X_i)_-\right|\right|_p \leq \frac{p}{p-1} ||(X_n)_-||_p$$

(because both X_n and $-X_n$ are submartingales if X_n is a martingale). By assumption, both right-hand sides are uniformly bounded in n, and the left hand sides are nondecreasing in n, so $\sup_{i \geq 0} |X_i| = Y$ is an L^p-integrable random variable (by the monotone convergence theorem applied to the variables $|\max_{1 \leq i \leq n} X_i|^p$).

So now $X_n \to X_\infty$ almost surely, meaning $|X_n - X_\infty|^p \to 0$ almost surely. But $|X_n - X_\infty|^p$ is dominated by $(2Y)^p$, which is integrable, so the dominated convergence theorem tells us that $\mathbb{E}[|X_n - X_\infty|^p] \to 0$, which is the desired L^p convergence. □

We're now ready to apply these results to **branching processes**. Recall the definition of a **Galton-Watson tree**: we have a double infinite array of random variables $(X_{n,i})_{n\geq 1, i\geq 1}$ all iid to an **offspring law** X which is supported on $\mathbb{Z}_{\geq 0}$ and has finite mean $\mu = \mathbb{E}[X]$. We start with a single root vertex, which generates $X_{1,1}$ children at depth 1. Then the ith child at depth $(n-1)$ generates $X_{n,i}$ children at depth n, and the Galton-Watson process Z_n is just the number of vertices at depth n, and we have the equation $Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i}$. Last time, we showed that if $\mathcal{F}_n = \sigma(X_{\ell,i} \text{ for all } i \geq 1, 1 \leq \ell \leq n)$ (that is, the filtration tells us the structure of the tree up to depth n), then

$$\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = \mu Z_n \implies M_n = \frac{Z_n}{\mu^n} \text{ is a martingale.}$$

Furthermore, if M_n is a nonnegative martingale, we know that M_n converges almost surely to some limit M_∞ with $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_0] = 1$. But from here, the behavior depends on the mean μ:

- One case that is easy to understand is the **subcritical Galton-Watson**, where $\mu < 1$. By Markov's inequality (using that Z_n is nonnegative-valued),

  $$\mathbb{P}(Z_n > 0) = \mathbb{P}(Z_n \geq 1) \leq \mathbb{E}[Z_n] = \mu^n,$$

  which is exponentially small in n. Thus Z_n must converge to 0 almost surely, so the population goes extinct with probability 1. In fact, we even have that $M_n = \frac{Z_n}{\mu^n} \to 0$ almost surely **in this case**, because Z_n is integer-valued so it must eventually be zero. So the limit M_∞ of the martingale is identically zero, and in fact this is an example where $\mathbb{E}[M_\infty] = 0 < 1 = \mathbb{E}[M_0]$.
- The other cases are more interesting. $\mu = 1$ gives us the the **critical Galton-Watson**, and for this case we will assume that $\mathbb{P}(X = 1) < 1$ (because otherwise we just have one child at each level and nothing interesting happens). It turns out that extinction will occur here again, but the proof is less straightforward:

Proposition 182

In the critical Galton-Watson process, we have $Z_n \to 0$ almost surely.

Proof. Because $\mu = 1$, Z_n is a nonnegative martingale (and thus also a nonnegative supermartingale), so it converges almost surely to some limit Z_∞. Since Z_n is integer-valued, Z_∞ will also be integer-valued, and in fact (to have almost-sure convergence) we must have $Z_n(\omega) = Z_\infty(\omega)$ for all sufficiently large $n \geq n(\omega)$. In other words, we have that

$$\mathbb{P}\left(\bigcup_{k\geq 0}\bigcup_{m\geq 0}\{Z_n = k \text{ for all } n \geq m\}\right) = 1$$

(where k corresponds to $Z_\infty(\omega)$ and m corresponds to $n(\omega)$). It suffices to show that this event cannot occur with positive probability for any $k > 0$. Intuitively, this is because we're asking the process to stay at k children forever, but X is a nondegenerate random variable so this cannot always happen. To make that rigorous, note that for any $k > 0$, there is some constant $c_k < 1$ (independent of n) such that $\mathbb{P}(Z_{n+1} = k | Z_n = k) = c_k$. Then

$$\mathbb{P}(Z_n = k \text{ for all } n \in \{m, \cdots, m+\ell\}) \leq c_k^\ell,$$

which decreases exponentially with ℓ, so the probability that $Z_n = k$ for all $n \geq m$ is zero. The countable union of all such events over all $k > 0$ and all m thus has probability zero,so we indeed have

$$\mathbb{P}\left(\bigcup_{m\geq 0}\{Z_n = 0 \text{ for all } n \geq m\}\right) = 1.$$

This is exactly the same as having Z_n converge to 0 almost surely. □

- The most interesting case is the **supercritical Galton-Watson**, where $\mu > 1$. This time, if the probability p_0 of producing zero children is zero, then $Z_n > 0$ almost surely for all n (because every vertex produces a positive number of children). But if $p_0 > 0$, then the tree does go extinct with some positive probability (for example, if all vertices at a given level produce zero children). Intuitively, though, once the tree has survived for 100 generations, it will typically be pretty big, so we expect the tree to survive with some sizable probability. So the interesting question here is **whether the probability of nonextinction is positive**.

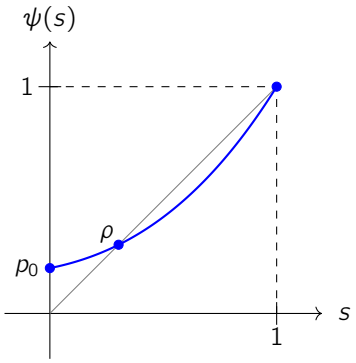
  To answer this question, we condition on the first level of the tree. If the first level has (for example) 3 children, then the $(n+1)$th level of the tree dies out if and only if the 3 level-n trees rooted at those children all die out. This yields the recursive formula

$$\mathbb{P}(Z_{n+1} = 0) = \sum_{k \geq 0} p_k \mathbb{P}(Z_n = 0)^k = \psi(\mathbb{P}(Z_n = 0)),$$

  where

$$\psi(s) = \sum_{k \geq 0} p_k s^k = \mathbb{E}[s^X].$$

  is a reparameterization of the moment generating function. Because X is nonnegative, $0 \leq s^X \leq 1$ when $0 \leq s \leq 1$. Thus $\mathbb{E}[s^X]$ is finite for $s \in [0, 1]$ (which is the range in which we are applying ψ), meaning this function is indeed defined. We can also check that ψ is nondecreasing and convex in s, and $\psi'(s) \uparrow \mu$ as $s \uparrow 1$ (by differentiating term-by-term and using the dominated convergence theorem on the partial sums). Since $\psi(1) = 1$ and $\psi(0) = p_0 > 0$, the graph of ψ will have the general shape shown below:



  Specifically, because the slope of ψ as $s \to 1$ is $\mu > 1$, our function will intersect $\psi(x) = x$ at some unique point $0 < \rho < 1$ (because $\psi(x) - x$ is convex, has negative derivative at $s = 0$, and has positive derivative at $s = 1$). We can now write

$$\mathbb{P}(Z_n = 0) = \psi(\mathbb{P}(Z_{n-1} = 0)) = \psi^n(\mathbb{P}(Z_0 = 0)) = \psi^n(0).$$

  Repeatedly applying a convex function will bring us towards a fixed point (which cannot be 1 because $\psi(x) < x$ for x near 1), and thus $\mathbb{P}(Z_n = 0) \to \rho$ as $n \to \infty$. Finally, because the events $\{Z_n = 0\}$ are nested (if we're extinct at time n, we're extinct at time $(n+1)$), their probabilities are nondecreasing and increase to the event {tree goes extinct}. Thus by continuity from below, the probability that the Galton-Watson tree goes extinct is indeed $\rho < 1$.

For a more complicated question, suppose we have a supercritical Galton-Watson tree and condition on the event that we do not go extinct. Then we may ask about the limiting distribution of $M_\infty = \lim_{n \to \infty} \frac{Z_n}{\mu^n}$. Knowing the answer

tells us a lot about the process – if we knew that M_n converged to some finite limit in $(0, \infty)$, then we'd know that Z_n grows like μ^n up to a constant factor (which is much stronger than just knowing for example that $Z_n = \mu^{n+o(n)}$). It turns out that we have an exact characterization of when this happens – the same recursion as before tells us that

$$\psi(\mathbb{P}(M_\infty = 0)) = \mathbb{P}(M_\infty = 0).$$

The only fixed points of ψ are $0, \rho$, and 1, and we can't have $\mathbb{P}(M_\infty = 0)$ because extinction implies $M_\infty = 0$ and occurs with positive probability (here we use that $p_0 \neq 0$, so $\rho \neq 0$). But it turns out that the other two possibilities both occur:

Theorem 183 (Kesten-Stigum $L \log L$ criterion)
In a supercritical Galton-Watson tree, $M_\infty = \lim \frac{Z_n}{\mu^n}$ is not identically zero if and only if the offspring law X satisfies $\mathbb{E}[X \log X] < \infty$.

We'll see a special case of this on our homework, and we'll also discuss this topic more in a later lecture.

20 November 20, 2019

Our last homework assignment is already posted on Stellar, and it's due on Monday, December 2 (the class day before the exam). There will be lecture that Monday but not on Wednesday, and we should make sure we have enough time to both study for the exam and finish the problem set.

We've been discussing martingale convergence theorems in the last few lectures. We'll start with a brief review: if X_n is a submartingale satisfying a moment condition, then it converges almost surely (by Doob's upcrossing inequality). As a consequence, any nonnegative supermartingale converges almost surely. However, we have no guarantee that there is convergence in L^1, and indeed we do not always have $\mathbb{E}[X_n] \to \mathbb{E}[X_\infty]$. Last time, we showed a sufficient condition with the L^p martingale convergence theorem: if there is some $p \in (1, \infty)$ such that $\sup_n ||X_n||_p < \infty$, then $X_n \to X_\infty$ almost surely and in L^p. Then if $X_n \to X_\infty$ in L^p, then it converges in L^1, so the expectations also converge.

But it turns out that we can get an exact characterization for convergence in L^1, which we'll discuss today.

Definition 184
Let $(X_i)_{i \in I}$ be a family of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then (X_i) is **uniformly integrable** (also **u.i.**) if

$$\lim_{M \to \infty} \sup_{i \in I} \mathbb{E}[|X_i|; |X_i| \geq M] = 0.$$

This is related to the idea of tightness of a set of measures (but isn't exactly equivalent). Recall that for any integrable function X, the dominated convergence theorem tells us that $\mathbb{E}[|X|; |X| \geq M] \to 0$ as $M \to \infty$, but uniform integrability is a uniform condition on the whole set of random variables.

Remark 185. *Note that if $(X_i)_{i \in I}$ is uniformly integrable, then for any X_i we can write*

$$\mathbb{E}[|X_i|] = \mathbb{E}[|X_i|; |X_i| < M] + \mathbb{E}[|X_i|; |X_i| \geq M] \leq M + \mathbb{E}[|X_i|; |X_i| \geq M].$$

Then we can make the second term uniformly small (say smaller than 1) by taking M to be sufficiently large. In particular, this shows that $\boxed{\sup_{i \in I} \mathbb{E}[|X_i|] < \infty}$*, so uniform integrability is stronger than having a uniform bound on the L^1 norm. In particular, this means that for any fixed M, the supremum in Definition 184 is finite.*

We'll describe one natural way to construct a family of uniformly integrable random variables:

Lemma 186
For any random variable $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, $\{\mathbb{E}[X|\mathcal{G}] : \mathcal{G} \subseteq \mathcal{F} \text{ sub-}\sigma\text{-field}\}$ is uniformly integrable.

Proof. First, we claim that for all $\varepsilon > 0$, there is some $\delta > 0$ so that for all $A \in \mathcal{F}$,

$$\mathbb{P}(A) \leq \delta \implies \mathbb{E}[|X|; A] \leq \varepsilon.$$

(This was on our homework – the idea is that if this didn't hold, then we can find a series of events A_n with $\mathbb{P}(A_n) \downarrow 0$ but $\mathbb{E}[X; A_n] \neq 0$. But X is in L^1, so this is a violation of the dominated convergence theorem.) Turning to the proof of the lemma, by Markov's inequality, conditional Jensen's inequality, and the tower property, we have

$$\mathbb{P}(|\mathbb{E}[X|\mathcal{G}]| \geq M) \leq \frac{\mathbb{E}[|\mathbb{E}[X|\mathcal{G}]|]}{M} \leq \frac{\mathbb{E}[\mathbb{E}[|X|\,|\mathcal{G}]]}{M} = \frac{\mathbb{E}[|X|]}{M}.$$

We will be using the various $\{|\mathbb{E}[X|\mathcal{G}]| \geq M\}$s (for sub-$\sigma$-fields $\mathcal{G}$) as the potential events A, and our goal is to bound

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]; |\mathbb{E}(X|\mathcal{G})| \geq M] = \mathbb{E}[|\mathbb{E}[X|\mathcal{G}] \cdot 1\{|\mathbb{E}(X|\mathcal{G})| \geq M\}|]$$

by writing it in the form $\mathbb{E}[|X|; A]$. But since $1\{|\mathbb{E}[X|\mathcal{G}]| \geq M\}$ is $\mathcal{G}$-measurable, and thus we can put it inside the expectation as well and use conditional Jensen's and the tower property again, simplifying the expression above to

$$\mathbb{E}\left[|\mathbb{E}[X \cdot 1\{|\mathbb{E}(X|\mathcal{G})| \geq M\}|\mathcal{G}]|\right] \leq \mathbb{E}\left[\mathbb{E}[|X| \cdot 1\{|\mathbb{E}(X|\mathcal{G})| \geq M\}|\mathcal{G}]\right] = \mathbb{E}\left[|X|; |\mathbb{E}(X|\mathcal{G})| \geq M\right].$$

By the claim at the beginning of our proof, we can make this expression at most ε (for all $\mathcal{G}$ simultaneously) if we make $\mathbb{P}(|\mathbb{E}(X|\mathcal{G})| \geq M)$ sufficiently small. But we have shown that those probabilities are bounded by $\frac{\mathbb{E}[|X|]}{M}$, so taking M large enough yields the result. □

Theorem 187
Let X_n be integrable random variables. If $X_n \to X$ converges in probability, then the following are equivalent:

1. $\{X_n\}_{n \geq 0}$ is uniformly integrable,
2. $X_n \to X$ in L^1,
3. $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$.

This result is true for random variables in general, but we'll apply it to martingales.

Remark 188. *We will use the bounded convergence theorem in our proof, so we will mention now that our measure theory convergence theorems can be strengthened. Instead of proving that $\int f_n d\mu \to \int f\, d\mu$, both results (with slight modifications) actually prove the stronger result that $\int |f_n - f| d\mu \to 0$. Additionally (using that convergence in probability implies almost-sure convergence along a subsequence), the dominated convergence theorem only requires convergence in probability.*

Proof. First, we show that (1) $\implies$ (2). By Remark 185, uniform integrability implies that $\sup_n \mathbb{E}[|X_n|] < \infty$. Again recalling that convergence in probability implies existence of a subsequence $X_{n_k} \to X$ converging almost surely, Fatou's lemma shows that $\mathbb{E}[|X|] \leq \liminf_{k \to \infty} \mathbb{E}[|X_{n,k}|] < \infty$, so $X \in L^1$. To show convergence in L^1, define the "truncated

identity function"

$$\phi(x) = \begin{cases} M & x \geq M \\ -M & x \leq M \\ x & \text{otherwise.} \end{cases}$$

Then by the triangle inequality, we can write

$$\mathbb{E}[|X_n - X|] \leq \mathbb{E}[|\phi(X_n) - \phi(X)|] + \mathbb{E}[|\phi(X) - X|] + \mathbb{E}[|X_n - \phi(X_n)|].$$

Because ϕ is continuous and uniformly bounded, and $X_n \to X$ in probability, we also have $\phi(X_n) \to \phi(X)$ in probability (for example by using the characterization that "all subsequences have a further subsequence converging almost surely"). Then $\mathbb{E}[|\phi(X_n) - \phi(X)|] \to 0$ by the (slightly stronger, as in Section 20) bounded convergence theorem, so the first term goes to zero.

Next, because $|\phi(X) - X| \leq |X| 1\{|X| \geq M\}$, the second term can be upper bounded by $\mathbb{E}(|X|; |X| \geq M)$. Finally, the third term is bounded by $\mathbb{E}[|X_n|, |X_n| \geq M]$. Thus, we can make the right-hand side arbitarily small by first fixing M and taking $n \to \infty$ to make the first term small, and then take $M \to \infty$ (using uniform integrability here) to make the last two small. Thus $\mathbb{E}[|X_n - X|] \to 0$, showing the desired L^1 convergence.

To show that (2) $\implies$ (3), we use Jensen's inequality to find that

$$|\mathbb{E}[|X_n|] - \mathbb{E}[|X|]| = |\mathbb{E}[|X_n| - |X|]| \leq \mathbb{E}[||X_n| - |X||] \leq \mathbb{E}[|X_n - X|]$$

(where the last step is just a property of real numbers). Since the right-hand side goes to zero by assumption, so does the left-hand side.

Finally, for (3) $\implies$ (1), it suffices to show that

$$\lim_{M \to \infty} \left(\limsup_{n \to \infty} \mathbb{E}(|X_n|; |X_n| \geq M) \right) = 0$$

(the usual definition uses sup instead of lim sup, but to get to sup, we can deal with finitely many exceptions for X_n by picking a large enough M for each one and then taking the maximum). Consider the function

$$\psi_M(x) = \begin{cases} x & |x| \leq M - 1 \\ 0 & |x| \geq M \\ \text{linear interpolation} & \text{otherwise.} \end{cases}$$

Notice that ψ_M is continuous and that $|X| 1\{|X| \geq M\} \leq |X| - |\psi(X)| \leq |X| 1\{|X| \geq M - 1\}$. Thus,

$$\limsup_{n \to \infty} \mathbb{E}[|X_n|; |X_n| \geq M] \leq \limsup_{n \to \infty} \mathbb{E}[|X_n| - |\psi(X_n)|].$$

now using the bounded convergence theorem (because ψ is uniformly bounded by M) on the second term and assumption(3) on the first term means that

$$\limsup_{n \to \infty} \mathbb{E}[|X_n|; |X_n| \geq M] \leq \mathbb{E}[|X| - |\psi(X)|] \leq \mathbb{E}[|X|; |X| \geq M - 1].$$

As we send $M \to \infty$, the right-hand side goes to zero (because X is integrable), so the left-hand side does as well, proving uniform integrability. □

We can now apply these results to martingales:

Theorem 189 (Submartingale L^1 convergence theorem)
For a submartingale $(X_n)_{n\geq 0}$, the following are equivalent:

1. Uniform integrability,
2. Convergence almost surely and in L^1,
3. Convergence in L^1.

Proof. To show that (1) $\implies$ (2), we again have that $\sup_n \mathbb{E}|X_n| < \infty$ from our earlier observations. Thus (by the submartingale convergence theorem, Theorem 173), X_n converges almost surely to some limit X_∞, meaning $X_n \to X_\infty$ in probability as well. Thus we can apply Theorem 187 to show that uniform integrability also implies L^1 convergence. Finally, (2) $\implies$ (3) is trivial, and (3) $\implies$ (1) is a consequence of (2) $\implies$ (1) in Theorem 187. □

Theorem 190 (Martingale L^1 convergence theorem)
For a martingale $(X_n)_{n\geq 0}$, the following are equivalent:

1. Uniform integrability,
2. Convergence almost surely and in L^1,
3. Convergence in L^1,
4. The existence of a random variable $X \in L^1$ such that $X_n = \mathbb{E}[X|\mathcal{F}_n]$ for all n.

The first three conditions are the same as above (since martingales are submartingales), but this theorem also tells us that if we have a uniformly integrable martingale, then there is some integrable random variable for which that martingale is just gradually "exposing information."

Proof. (1), (2), and (3) are equivalent by Theorem 189, and (4) $\implies$ (1) is Lemma 186, so we just need to show that (3) $\implies$ (4). By repeatedly applying the martingale property, we know that $X_n = \mathbb{E}[X_\ell|\mathcal{F}_n]$ for all $\ell \geq n$, which is equivalent to saying that

$$\mathbb{E}[X_n; A] = \mathbb{E}[X_\ell; A] \text{ for all } A \in \mathcal{F}_n \text{ and } \ell \geq n.$$

Taking $\ell \to \infty$, we have $\mathbb{E}[X_n; A] = \mathbb{E}[X_\infty; A]$ because $X_\ell \to X_\infty$ in L^1 (so if $\mathbb{E}[|X_\ell - X_\infty|] \to 0$, we also have $\mathbb{E}[|1_A \cdot (X_\ell - X_\infty)|] \to 0$). Because everything is integrable here and we have satisfied the conditional expectation identity, $\mathbb{E}[X_\infty|\mathcal{F}_n] = X_n$ for all n, and X_∞ is our random variable X. □

Definition 191
Let $\mathcal{F}_n$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$. For any $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, the sequence $X_n = \mathbb{E}[X|\mathcal{F}_n]$ is the **Doob martingale** of X with respect to $\mathcal{F}_n$.

As mentioned, the idea of such a sequence is that we reveal the randomness of X "a little bit at a time," and what we've just shown is that any martingale converging in L^1 is exactly a Doob martingale for some X.

Theorem 192 (Lévy convergence theorem)
Let $\mathcal{F}_n$ be a filtration of $(\Omega, \mathcal{F}, \mathbb{P})$, and define $\mathcal{F}_\infty = \sigma\left(\bigcup_n \mathcal{F}_n\right)$. Then for any $X \in L^1$, we have $\mathbb{E}[X|\mathcal{F}_n] \to \mathbb{E}[X|\mathcal{F}_\infty]$ almost surely and in L^1.

Remark 193. *The union of sigma-algebras is not itself a sigma-algebra, but here we are taking the sigma-algebra of the union. This is sometimes denoted $\mathcal{F}_n \uparrow \mathcal{F}_\infty$.*

This is a statement about the limit of the Doob martingale. We should compare this with Theorem 190 – in that case, we explicitly constructed a random variable X which satisfied the properties we want. But $\mathcal{F}_\infty$ can be strictly smaller than $\mathcal{F}$, and the Lévy convergence theorem holds even if X is some arbitrary measurable function on $\mathcal{F}$ (even if it's not in $\mathcal{F}_\infty$).

Proof. Define $X_n = \mathbb{E}[X|\mathcal{F}_n]$. We know that $\{X_n\}_{n\geq 0}$ is a uniformly integrable family by Lemma 186, so it converges to some X_∞ almost surely and in L^1 by the L^1 martingale convergence theorem (Theorem 190). So we just need to show that $X_\infty = \mathbb{E}[X|\mathcal{F}_\infty]$, which is an exercise with the definition of conditional expectation. We know that $X_\infty \in \mathcal{F}_\infty$, because it is the almost-sure limit of the X_ns, each of which is measurable with respect to $\mathcal{F}_n$ and thus to $\mathcal{F}_\infty$. For the conditional identity, notice that

$$X_n = \mathbb{E}[X|\mathcal{F}_n] \implies \mathbb{E}[X_n; A] = \mathbb{E}[X; A] = E[X_\infty; A] \text{ for all } A \in \mathcal{F}_n,$$

where the last equality comes from $X_n = \mathbb{E}[X_\infty|\mathcal{F}_n]$ (again by the proof of Theorem 190). Applying this for an arbitrary n, we find that $\mathbb{E}[X; A] = \mathbb{E}[X_\infty; A]$ for all $A \in \bigcup_{n\geq 0} \mathcal{F}_n$. To finish, we extend this to the sigma-algebra $\sigma\left(\bigcup_{n\geq 0} \mathcal{F}_n\right)$ with the standard $\pi - \lambda$ argument. □

Applying this result to the indicator random variable 1_A (for any $A \in \mathcal{F}_\infty$) yields the following useful result:

Corollary 194 (Lévy 0-1 law)
Let $\mathcal{F}_n \uparrow \mathcal{F}_\infty$. Then for any $A \in \mathcal{F}_\infty$, we have $\mathbb{P}(A|\mathcal{F}_n) \to 1_A$ (so in particular, the conditional probability converges to either 0 or 1).

We'll finish with a few helpful hints for the homework. The **Azuma-Hoeffding bound** says that if X_n is a martingale with bounded increments, then $X_n - \mathbb{E}[X_n]$ is well-concentrated. One consequence is the **concentration of the independence number** in $G_{n,p}$ (the random graph of n vertices where each pair is connected with probability p). Basically, if $G = (V, E)$ is a graph, then $S \subseteq V$ is an **independent set** if no two vertices in S are neighbors in G. The **independence number** $\alpha(G)$ is then maximal cardinality of such a set $|S|$ (which is some number between 1 and $|V|$). We are asked to determine the behavior of $X = \alpha(G)$ as a random variable. One useful way to define a filtration is

$$\mathcal{F}_\ell = \sigma\left(\text{edges restricted to the first } \ell \text{ vertices}\right).$$

This is known as an **edge-revealing filtration**, in which we're told at each step how a new vertex is connected to the currently revealed graph. There are variations on this idea, but the point is that the random variables $\mathbb{E}[X|\mathcal{F}_\ell]$ have no integrability issues (because G is a finite graph) and form a Doob martingale, so using the Azuma-Hoeffding bound will give us a concentration bound! And this kind of strategy also works for the clique number and chromatic number of a random graph, too. (This general type of argument comes from a paper by Shamir and Spencer [9].) But as we'll see on our homework, our asymptotic bound will only be good for some values of p.

21 November 25, 2019

In the last few lectures, we've shown some results about submartingale convergence: specifically, if X_n is a submartingale with $\sup_n \mathbb{E}[(X_n)_+]$ finite, then $X_n \to X_\infty$ converges almost surely with $\mathbb{E}[|X_\infty|] < \infty$. Also, uniform integrability, convergence almost surely and in L^1, and convergence in L^1 are all equivalent for a submartingale.

Today, we'll talk about the **optional stopping theorem**. Recall that a **stopping time** τ with respet to a filtration $\mathcal{F}_n$ is a random variable such that $\{\tau = n\} \in \mathcal{F}_n$ for all n. As discussed previously, for any submartingale X_n and any stopping time that is almost surely bounded by $0 \leq \tau \leq k$, we have $\mathbb{E}[X_0] \leq \mathbb{E}[X_\tau] \leq \mathbb{E}[X_k]$. More generally, this tells us that $\mathbb{E}[X_0] \leq \mathbb{E}[X_{\tau \wedge n}] \leq \mathbb{E}[X_n]$ for any finite n, but we can't just replace n with ∞ here. So today's discussion will explain what we can do when we do take this limit.

Lemma 195

Let X_n be a uniformly integrable submartingale and τ be any stopping time (potentially with $\mathbb{P}(\tau = \infty) > 0$). Then $\mathbb{E}[|X_\tau|] < \infty$.

Proof. Because X_n is uniformly integrable, it converges to some limit X_∞ almost surely, so the limit exists and X_τ is well-defined (even when $\tau = \infty$). Since $f(x) = \max(x, 0)$ is a nondecreasing convex function, $(X_n)_+$ is also a submartingale, so we also have $\mathbb{E}[(X_{n\wedge\tau})] \leq \mathbb{E}[(X_n)_+]$ for any n. The right-hand side is uniformly bounded (by uniform integrability) by some constant, which means that $\sup_n \mathbb{E}[(X_{n\wedge\tau})_+] < \infty$. But $(X_{n\wedge\tau})$ is a stopped process of (X_n), so it is a submartingale as well. Thus by the submartingale convergence theorem, $X_{n\wedge\tau}$ converges to a limit in L^1. But that limit is exactly X_τ (whether τ is finite or infinite). □

Corollary 196

If X_n is a uniformly integrable submartingale, and τ is any stopping time, then $Y_n = X_{n\wedge\tau}$ is also a uniformly integrable submartingale.

Proof. We wish to show that

$$\lim_{M\to\infty} \left(\sup_n \mathbb{E}[|X_{n\wedge\tau}|; |X_{n\wedge\tau}| \geq M] \right) = 0.$$

We can split up the expectation into two cases, based on whether n or τ is smaller, to get

$$\begin{aligned} \sup_n \mathbb{E}[|X_{n\wedge\tau}|; |X_{n\wedge\tau}| \geq M] &= \sup_n \mathbb{E}[|X_n| \cdot 1\{\tau > n\} + |X_\tau| 1\{\tau \leq n\}; |X_{n\wedge\tau}| \geq M] \\ &\leq \sup_n \left(\mathbb{E}[|X_n| 1\{|X_n| \geq M\}] + \mathbb{E}[|X_\tau|; |X_\tau| \geq M] \right). \end{aligned}$$

But because the X_n are uniformly integrable, the first term goes to zero as $M \to \infty$, and because $X_\tau \in L^1$ (by Lemma 195), the second term goes to zero as $M \to \infty$ by the dominated convergence theorem. □

Theorem 197

If X_n is a uniformly integrable submartingale and τ is a stopping time, then $\mathbb{E}[X_0] \leq \mathbb{E}[X_\tau] \leq \mathbb{E}[X_\infty]$.

As we will see in the proof, the first inequality $\mathbb{E}[X_0] \leq \mathbb{E}[X_\tau]$ holds even if we only know that the stopped process $X_{n\wedge\tau}$ is a uniformly integrable submartingale.

Proof. We have previously shown that $\mathbb{E}[X_0] \leq \mathbb{E}[X_{\tau\wedge n}] \leq \mathbb{E}[X_n]$ for all finite n. Since $X_{\tau\wedge n}$ is uniformly integrable (by Corollary 196), it converges to X_τ in L^1, so $\mathbb{E}[X_0] \leq \mathbb{E}[X_\tau]$.

For the other bound, since X_n is uniformly integrable, X_n converges to X_∞ almost surely and in L^1. So $X_{\tau\wedge n}$ and X_n converge almost surely and in L^1 to X_τ and X_∞, respectively, and because $\mathbb{E}[X_{\tau\wedge n}] \leq \mathbb{E}[X_n]$ for each n, taking limits on both sides yields $\mathbb{E}[X_\tau] \leq \mathbb{E}[X_\infty]$, as desired. □

However, uniform integrability can often be difficult to check, so here's an easier criterion:

Theorem 198

Suppose X_n is a submartingale such that $\mathbb{E}[|X_{n+1} - X_n| \,|\mathcal{F}_n] \leq B$ for all n. Then if τ is a stopping time with $\mathbb{E}[\tau] < \infty$, then $X_{n\wedge\tau}$ is uniformly integrable (so in particular we do have $\mathbb{E}[X_0] \leq \mathbb{E}[X_\tau]$).

Proof. For any $n \geq 0$, we can write

$$|X_{n\wedge\tau}| \leq |X_0| + \sum_{i\geq 1} 1\{\tau > i\}|X_{i+1} - X_i|$$

by the triangle inequality. Call the right-hand side Y, and notice that it is independent of n. Now $1\{\tau > i\}$ is measurable with respect to $\mathcal{F}_i$, so the expectation of Y can be written (by the tower property) as

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[|X_0|] + \sum_{i\geq 0} \mathbb{E}\left[\mathbb{E}\left[1\{\tau > i\} \cdot |X_{i+1} - X_i| \,|\mathcal{F}_i\right]\right] \\ &= \mathbb{E}[|X_0|] + \sum_{i\geq 0} \mathbb{E}\left[1\{\tau > i\} \cdot \mathbb{E}\left[|X_{i+1} - X_i| \,|\mathcal{F}_i\right]\right].\end{aligned}$$

But now using our assumption, we can show that Y is integrable:

$$\mathbb{E}[Y] \leq \mathbb{E}[|X_0|] + \sum_{i\geq 0} \mathbb{E}\left[1\{\tau > i\} \cdot B\right] = \mathbb{E}[|X_0|] + B\mathbb{E}[\tau] < \infty.$$

Because all $|X_{n\wedge\tau}|$s are dominated by the integrable random variable Y, we have $\mathbb{E}[|X_{n\wedge\tau}|; |X_{n\wedge\tau}| \geq M] \leq \mathbb{E}[|Y|; |Y| \geq M]$ for all n. Since the right-hand side goes to zero as $M \to \infty$, we get the desired uniform integrability condition. □

Next, we have a result that's a bit disconnected from the previous ones, in that we don't even need uniform integrability at all:

Theorem 199

If X_n is a nonnegative supermartingale, and τ is any stopping time, then $\mathbb{E}[X_0] \geq \mathbb{E}[X_\tau]$.

Proof. By Theorem 174, we have $X_n \to X_\infty$ almost surely, so again X_τ is well-defined. But because $\mathbb{E}[X_0] \geq \mathbb{E}[X_{n\wedge\tau}]$ for all n, Fatou's lemma tells us that

$$\mathbb{E}[X_\tau] \leq \liminf \mathbb{E}[X_{n\wedge\tau}] \leq \mathbb{E}[X_0],$$

as desired. □

We will now generalize Theorem 197 by looking at multiple stopping times at once:

Definition 200

If τ is a stopping time with respect to $\mathcal{F}_n$, then the **stopping time σ-algebra** associated to τ is

$$\mathcal{F}_\tau = \{A \in \mathcal{F} : A \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n\}.$$

This definition essentially says that an event A is in the stopping time σ-algebra if we "know if we're in A at time τ," since any event in $\mathcal{F}_n$ is "known" at stage n. We can check that if $\tau = k$ almost surely, then $\mathcal{F}_\tau = \mathcal{F}_k$, and (with more work) if we have a process $Y_n \in \mathcal{F}_n$, then $Y_\tau \in \mathcal{F}_\tau$.

Theorem 201

Let $X_{n\wedge\tau}$ be a uniformly integrable submartingale. If σ, τ are both stopping times and $\sigma \leq \tau$ almost surely, then $\mathbb{E}[X_\tau | F_\sigma] \geq X_\sigma$.

(When $\sigma = 0$, this is very similar to the inequality $\mathbb{E}[X_\tau] \geq \mathbb{E}[X_0]$ that we've proved above.)

Proof. Since $Y_n = X_{n\wedge\tau}$ is a uniformly integrable submartingale and σ is a stopping time, we know that $\mathbb{E}[Y_0] \leq \mathbb{E}[Y_\sigma] \leq \mathbb{E}[Y_\infty]$ (by Theorem 197), which we can rewrite as $\mathbb{E}[X_0] \leq \mathbb{E}[X_\sigma] \leq \mathbb{E}[X_\tau]$ because $\sigma \leq \tau$ almost surely. Now fix $A \in \mathcal{F}_\sigma$ and define the random variable

$$\xi(\omega) = \sigma 1_A + \tau 1_{A^c}.$$

We always have either $\xi = \sigma$ or $\xi = \tau$, and we claim that ξ is a stopping time. Indeed, we can write

$$\{\xi = n\} = (A \cap \{\sigma = n\}) \sqcup (A^c \cap \{\tau = n\}).$$

Because $A \in \mathcal{F}_\sigma$, the first term $(A \cap \{\sigma = n\})$ is in $\mathcal{F}_n$ (by the definition of the stopping time σ-algebra). And the second term can be written as

$$A^c \cap \{\tau = n\} = \left(\bigcup_{k=0}^{n} A^c \cap \{\sigma = k\} \cap \{\tau = n\} \right)$$

(since $\sigma \leq \tau$ by assumption), at which point we can notice that $A^c \cap \{\sigma = k\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$ and $\{\tau = n\} \in \mathcal{F}_n$, so the entire right-hand side is also in $\mathcal{F}_n$. Putting this all together, $\{\xi = n\} \in \mathcal{F}_n$, so we do have a stopping time. Since ξ is bounded by τ almost surely, we have $\mathbb{E}[Y_\xi] \leq \mathbb{E}[Y_\infty] \implies \mathbb{E}[X_\xi] \leq \mathbb{E}[X_\tau]$ by the same logic as for σ. But writing out the definition of ξ, this tells us that

$$\mathbb{E}[X_\sigma; A] + \mathbb{E}[X_\tau; A^c] \leq \mathbb{E}[X_\tau] \implies \boxed{\mathbb{E}[X_\sigma; A]} \leq \mathbb{E}[X_\tau; A] = \boxed{\mathbb{E}[\mathbb{E}[X_\tau | \mathcal{F}_\sigma]; A]}$$

by the tower property. Since A is an arbitrary event in $\mathcal{F}_\sigma$, this shows the desired result. □

We can now work with a quantitative example to see the optional stopping theorem in action:

Example 202 (Asymmetric random walk)

Consider a random walk on the integers defined by $S_n = \sum_{i=1}^n \xi_i$, where each ξ_i is 1 with probability p and -1 with probability $q = 1 - p$. Assume that $p \in (0.5, 1)$ (so that there is a drift in the upward direction).

We can rewrite S_n as

$$S_n = \sum_{i=1}^{n} (\xi_i - \mathbb{E}[\xi]) + n \cdot \mathbb{E}[\xi].$$

The first term here is the sum of n iid terms of mean zero, and it has standard deviation on the order of $\sqrt{n}$. But the second term is of order n, so it will dominate for large n and we will eventually stop returning to position 0. To quantify this, notice that the equation $1 = \mathbb{E}[s^\xi] = ps + q\frac{1}{s}$ has solutions at $s = 1$ and $\frac{q}{p}$, so $M_n = \left(\frac{q}{p}\right)^{S_n}$, which is a product of iid terms of mean 1, will be a nonnegative martingale. Now for any integer x, let

$$\tau_x = \inf\{n \geq 0 : S_n = x\}$$

be the first hitting time of x, which can be infinite if the walk never reaches x. Now fix some $a, b > 0$ and define $\tau = \tau_{-a} \wedge \tau_b$. The stopped martingale M_τ is uniformly bounded (between $\left(\frac{q}{p}\right)^{-a}$ and $\left(\frac{q}{p}\right)^{b}$), so $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$. (Here, we are essentially using Theorem 197 and its analogous result for supermartingales.) Also, τ is finite almost surely because the probability of not escaping $(-a, b)$ is exponentially decaying in n. But now if we let π be the probability that the walk hits $-a$ before b, we have

$$1 = \mathbb{E}[M_0] = \mathbb{E}[M_\tau] = \pi \cdot \left(\frac{q}{p}\right)^{-a} + (1-\pi) \cdot \left(\frac{q}{p}\right)^{b} \implies \pi = \frac{1 - \left(\frac{q}{p}\right)^{b}}{\left(\frac{q}{p}\right)^{-a} - \left(\frac{q}{p}\right)^{b}}.$$

Taking $a \to \infty$, we find that the probability of never hitting b is $\mathbb{P}(\tau_b = \infty) = 0$, and taking $b \to \infty$, we find that the probability of hitting a is $\mathbb{P}(\tau_{-a} < \infty) = \left(\frac{q}{p}\right)^{a}$. So in words, any positive level will be reached almost surely, and any negative level has a positive probability of never being reached.

Next, we ask about the expected amount of time it will take to hit any $b > 0$. For this, define a new martingale

$$X_n = S_n - n\mathbb{E}[\xi] = S_n - n(p-q).$$

Since $S_{\tau_b} = b$, it makes sense to expect that $b - \mathbb{E}[\tau_b](p-q) = 0$. To justify this, let S_n be the minimum value of S_n over all $n \geq 0$. This is a nonpositive random variable, and it is integrable because

$$\mathbb{E}[-S_{\min}] = \sum_{a \geq 1} \mathbb{P}(-S_{\min} \geq a) = \sum_{a \geq 1} \mathbb{P}(\tau_{-a} < \infty) = \sum_{a} \left(\frac{q}{p}\right)^{a} < \infty$$

because we have a geometric series with common ratio $\frac{q}{p} < 1$. So $S_{\min} \in L^1$, and because $S_{\min} \leq S_{n \wedge \tau_b} \leq b$ for all n, $S_{n \wedge \tau_b}$ is uniformly integrable (for example because it is dominated by $|S_{\min}| + b$). This means that $S_{n \wedge \tau_b}$ converges almost surely and in L^1 to its limit S_{τ_b}, which is almost surely b by definition. So now turning back to our martingale X_n, this convergence in L^1 tells us that (because $0 \leq \tau_b \wedge n \leq n$)

$$0 = \mathbb{E}[X_0] = \mathbb{E}[X_{n \wedge \tau}] = \mathbb{E}[S_{n \wedge \tau_b}] - \mathbb{E}[n \wedge \tau_b](p-q) \overset{n \to \infty}{\implies} 0 = b - \mathbb{E}[\tau_b](p-q),$$

because $\mathbb{E}[n \wedge \tau_b]$ converges to $\mathbb{E}[\tau_b]$ by the monotone convergence theorem. Thus $\mathbb{E}[\tau_b] = \frac{b}{p-q} = \frac{b}{2p-1}$.

Example 203 (Patterns in a random string)
Sample $(\sigma_1, \sigma_2, \cdots)$ iid from the alphabet $\mathcal{A} = \{A, B, \cdots, Z\}$ (here $|\mathcal{A}| = 26$). Let τ be the first time we see a particular sequence of letters w, which we'll take to be "ABRACADABRA." We wish to study $\mathbb{E}[\tau]$.

We have $\tau \geq \text{length}(w) = 11$, and we can check that $\mathbb{E}[\tau] < \infty$ (because within any block of time there is a positive chance to see the word, so we have geometric decay). So τ is indeed integrable. We construct a martingale – suppose that before each time n, a new gambler G_n enters and bets 1 dollar on the event $\{\sigma_n = w_1\}$. If $\sigma_n \neq w_1$, then G_n loses and exits; otherwise G_n wins \$26. In the latter case, G_n bets all of the money on the event $\{\sigma_{n+1} = w_2\}$, either losing it or winning $\$26^2$. This betting continues for G_n until there is a mismatch (that is, the event $\{\sigma_{n+i-1} \neq w_i\}$), or the entire string w is correctly predicted, in which case the gambler wins $\$26^{11}$.

Now, let M_n be the total winnings up to time n from the point of view of the casino. This is integrable for any n (since any gambler's winnings are bounded), and all games are fair, so M_n is a martingale (with respect to $\mathcal{F}_n$, which encodes both the sequence and the bets). Also, $\mathbb{E}[|M_{n+1} - M_n| \,|\mathcal{F}_n]$ is uniformly bounded almost surely, because only the 11 most recent gamblers can be making bets (in particular we have a bound of $26^{11} \cdot 11$). Thus by Theorem 198,

if τ is the stopping time where some gambler wins for the first time,

$$0 = \mathbb{E}[M_0] = \mathbb{E}[M_\tau] = \mathbb{E}\left[\tau - 26^{11} - 26^4 - 26\right],$$

because when the game stops, the casino has won one dollar from each of the τ gamblers, but at the stopping point, there are three gamblers who have won $26, 26^4, 26^{11}$ dollars respectively. So $\mathbb{E}[\tau] = 26^{11} + 26^4 + 26$ for this example (and this generalizes to any sequence w), giving us our answer.

Next lecture, we'll move on to discussing **reverse martingales** and their applications to **zero-one laws**.

22 November 27, 2019

We'll study **reverse martingales** today, starting with two different (equivalent) definitions:

Definition 204

A **reverse martingale** is a martingale "indexed by the nonpositive integers $\mathbb{Z}_{\leq 0}$." In other words, a reverse martingale is a sequence of random variables $(M_n)_{n\leq 0}$ such that (1) $M_n \in L^1(\mathcal{F}_n)$ for all n, where $\cdots \subseteq \mathcal{F}_{-2} \subseteq \mathcal{F}_{-1} \subseteq \mathcal{F}_0 \subseteq \mathcal{F}$, and (2) we have the usual martingale property $M_n = \mathbb{E}[M_{n+1}|\mathcal{F}_n]$.

Equivalently, we can index a reverse martingale in the usual way $(M_n)_{n\geq 0}$, but now we require $\mathcal{F} \supseteq \mathcal{F}_0 \supseteq \mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \cdots$, and our martingale property is now $\mathbb{E}[M_n|\mathcal{F}_{n+1}] = M_{n+1}$. **We'll use this latter notation.**

Notice that if we have this nesting of σ-fields $\mathcal{F} \supseteq \mathcal{F}_0 \supseteq \cdots$, then $\mathcal{F}_n \downarrow \bigcap_{n\geq 0} \mathcal{F}_n$, which is itself a σ-field which we call $\mathcal{F}_\infty$. (So unlike with ordinary martingales, the the limit of the σ-fields is also a σ-field because we're taking intersections instead of unions.)

The theory of reverse martingales is actually easier than normal martingales: for example, we have $M_n = \mathbb{E}[M_0|\mathcal{F}_n]$ for all n, so we have "less and less" information as we progress.

Theorem 205

If $(M_n)_{n\geq 0}$ is a reverse martingale with respect to a filtration $\mathcal{F}_n \downarrow \mathcal{F}_\infty$, then $M_n \to M_\infty = \mathbb{E}[M_0|\mathcal{F}_\infty]$ almost surely and in L^1.

Proof. Let $U_n(a, b)$ be the number of upcrossings of $[a, b]$ traversed by the process $(M_n, \cdots, M_0)$ (this is going backwards in time, so this is a normal martingale). By Doob's upcrossing inequality, we have $\mathbb{E}[U_n(a, b)] \leq \frac{\mathbb{E}[M_0 - a]_+}{b-a}$. But the right-hand side is independent of n and finite (because M_0 is integrable), so $\mathbb{E}[U_n(a, b)]$ is bounded. Thus $\mathbb{E}[U_\infty(a, b)] < \infty$ by the monotone convergence theorem, so by the same logic as in Theorem 173, M_n must converge almost surely to M_∞. Because $M_n = \mathbb{E}[M_0|\mathcal{F}_n]$ for all n, $(M_n)_{n\geq 0}$ is a collection of conditional expectations of a M_0 and is thus uniformly integrable. Convergence in L^1 thus follows from Theorem 187.

To check that the limit is indeed $\mathbb{E}[M_0|\mathcal{F}_\infty]$, notice that for any $A \in \mathcal{F}_\infty \subset \mathcal{F}_n$, we have $\mathbb{E}[M_0; A] = \mathbb{E}[M_n; A]$ because $M_n = \mathbb{E}[M_0|\mathcal{F}_n]$, and then convergence $M_n \to M_\infty$ in L^1 implies that $\lim_{n\to\infty} \mathbb{E}[M_n; A] \to \mathbb{E}[M_\infty; A]$ as well. So $\mathbb{E}[M_0; A] = \mathbb{E}[M_\infty; A]$ for all $A \in \mathcal{F}_\infty$, which is the conditional expectation identity for $M_\infty = \mathbb{E}[M_0|\mathcal{F}_\infty]$. □

Corollary 206

For any random variable $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and any σ-fields $\mathcal{F}_n \downarrow \mathcal{F}_\infty$ (where $\mathcal{F}_n \subseteq \mathcal{F}$ for all n), $\mathbb{E}[Y|\mathcal{F}_n]$ converges to $\mathbb{E}[Y|\mathcal{F}_\infty]$ almost surely and in L^1.

Proof. $M_n = \mathbb{E}[Y|\mathcal{F}_n]$ is a reverse martingale, so it converges almost surely and in L^1 to $M_\infty = \mathbb{E}[M_0|\mathcal{F}_\infty] = \mathbb{E}[\mathbb{E}[Y|\mathcal{F}_0]|\mathcal{F}_\infty]$, which is indeed $\mathbb{E}[Y|\mathcal{F}_\infty]$ by the tower property. □

We'll spend the rest of the lecture on various **zero-one laws**, which tell us that certain types of events in a probability space must have probability either 0 or 1.

Definition 207

Let $(X_i)_{i\geq 1}$ be independent (but not necessarily iid) random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and define $\mathcal{F}_{n+} = \sigma(X_n, X_{n+1}, \cdots)$ for all n. The **tail σ-algebra** of (X_i) is the σ-field $\mathcal{T} = \bigcap_{n\geq 1} \mathcal{F}_{n+}$.

Example 208

The event $A = \{X_n \geq 3 \text{ i.o.}\}$ is in the tail sigma-algebra of the X_ns, because it does not depend on the value of the first k random variables for any k.

Formally, we can prove this by writing

$$A = \bigcap_{n\geq 1} \bigcup_{m\geq n} \{X_m \geq 3\}$$

(in words, $X_n \geq 3$ occurring infinitely often is the same as having $X_m \geq 3$ for some $m \geq n$ no matter what n we pick). But because the inner union is decreasing as a function of n, we can rewrite

$$A = \bigcap_{n\geq \ell} \bigcup_{m\geq n} \{X_m \geq 3\}$$

for any ℓ. But this event is now in $\mathcal{F}_{\ell_+}$ for all ℓ (because it doesn't depend on $X_1, \cdots, X_{\ell-1}$), so A is also in the intersection $\mathcal{T}$.

Theorem 209 (Kolmogorov 0-1 law)

Let $(X_i)_{i\geq 1}$ be independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{T}$ be their tail σ-algebra. Then $\mathcal{T}$ is trivial under $\mathbb{P}$, meaning that $\mathbb{P}(A) = 0$ or 1 for all $A \in \mathcal{T}$.

In words, if an event doesn't change when we change only finitely many coordinates, it either always happens or never happens.

Proof without martingales. Define $\mathcal{F}_k = \sigma(X_1, \cdots, X_k)$ and $\mathcal{F}_{m,n} = \sigma(X_m, \cdots, X_n)$ – because all of the X_is are independent, $\mathcal{F}_k$ is independent of $\mathcal{F}_{m,n}$ as long as $k < m \leq n$. Now fix $k < m$, and notice that $\bigcup_{n\geq m} \mathcal{F}_{m,n}$ is a π-system generating the σ-algebra $\mathcal{F}_{m+}$, so $\mathcal{F}_k$ **is independent of** $\mathcal{F}_{m+}$ by a π-λ argument (since the collection of subsets independent of $\mathcal{F}_k$ is a λ-system, and all of the π-system $\bigcup_{n\geq m} \mathcal{F}_{m,n}$ is is independent of $\mathcal{F}_k$).

Next, $\mathcal{F}_{m+}$ contains the tail sigma-algebra, so $\mathcal{F}_k \perp\!\!\!\perp \mathcal{T}$ for all k. Now $\bigcup_k \mathcal{F}_k$ is a π-system (with all elements independent of $\mathcal{T}$) which generates $\mathcal{F}_{1+}$, so $\mathcal{F}_{1+}$ **is independent of** $\mathcal{T}$ by another π-λ argument. But $\mathcal{F}_{1+}$ contains $\mathcal{T}$, so this really implies that $\mathcal{T}$ **is independent of itself**. Thus, for all $A \in \mathcal{T}$,

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2,$$

meaning that $\mathbb{P}(A) = 0$ or 1 for all A, as desired. □

Proof with forward martingales. We may argue as above to show that $\mathcal{F}_k$ is independent of $\mathcal{T}$ for all k. For any $A \in \mathcal{T}$, consider the martingale $M_n = \mathbb{E}[1_A|\mathcal{F}_n]$. By Theorem 192 M_n converges almost surely and in L^1 to $\mathbb{E}[1_A|\mathcal{F}_\infty] = 1_A$

(because $T \subseteq \mathcal{F}_\infty$, so $1_A \in \mathcal{F}_\infty$). But A is independent of $\mathcal{F}_n$ for all n, so $M_n = \mathbb{E}[1_A] = \mathbb{P}(A)$ for all n and thus we also have $M_\infty = \mathbb{P}(A)$. Setting the two limits equal, $\mathbb{P}(A) = 1_A$, so the probability can only be 0 or 1. □

The Kolmogorov 0-1 law is stated in terms of random variables, but we can also state it slightly differently. Suppose our probability space has a product structure

$$(\Omega, \mathcal{F}, \mathbb{P}) = \left(\prod_{i=1}^{\infty} S_i, \bigoplus_{i=1}^{\infty} \mathcal{S}_i, \bigoplus_{i=1}^{\infty} \mu_i\right),$$

and define the random variables $X_i : \Omega \to S_i$ via

$$X_i(\omega = (\omega_j)_{j=1}^{\infty}) = \omega_i$$

(that is, returning the ith coordinate of ω), so that the law of X_i is exactly μ_i. The next result requires the X_is to be iid, but it will be stated similarly to the previous setting:

Definition 210

Suppose we have a product space $(\Omega, \mathcal{F}, \mathbb{P}) = \left(\prod_{i=1}^{\infty} S, \bigoplus_{i=1}^{\infty} \mathcal{S}, \bigoplus_{i=1}^{\infty} \mu\right)$ (equivalently, a sequence of iid random variables). The **exchangeable σ-algebra** $\mathcal{E}$ is the set of events invariant under permutation of finitely many coordinates (that is, a bijection $\pi : \mathbb{N} \to \mathbb{N}$ with fixed points at all but finitely many coordinates). Formally, for any $\omega \in \Omega = S^{\mathbb{N}}$, define $\omega_\pi = (\omega_{\pi(i)})_{i=1}^{\infty}$, and for any $A \in \mathcal{F}$, define $A_\pi = \{\omega_\pi : \omega \in A\}$. Call A **permutable** if $A = A_\pi$ for all finite permutations, and let $\mathcal{E} = \{A : A \text{ permutable}\}$.

The set of events invariant under permutations of the first n coordinates is often written

$$\mathcal{E}_n = \{A \in \mathcal{F} : A = A_\pi \text{ for all } \pi : \mathbb{N} \to \mathbb{N} \text{ with } \pi(i) = i \ \forall i > n\}.$$

In particular, notice that $\mathcal{E}_n \downarrow \mathcal{E}$.

Example 211

We have $\mathcal{T} \subseteq \mathcal{E}$, because any event in the tail σ-algebra is invariant under exchange of the first n coordinates for any n and thus invariant under all permutations of finitely many coordinates. However, $\mathcal{T} \neq \mathcal{E}$ – for example, $A = \left\{\sum_{i=1}^{n} X_i \in [9, 10] \text{ i.o.}\right\}$ is exchangeable but not in $\mathcal{T}$ in general.

Theorem 212 (Hewitt-Savage 0-1 law)

On a product space $(\Omega, \mathcal{F}, \mathbb{P}) = \left(\prod_{i=1}^{\infty} S, \bigoplus_{i=1}^{\infty} \mathcal{S}, \bigoplus_{i=1}^{\infty} \mu\right)$, the exchangeable σ-field $\mathcal{E}$ is trivial under $\mathbb{P}$ (meaning that any event has probability 0 or 1).

Proof without martingales. We start with a useful fact:

Fact 213

On any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, if $\mathcal{F}_i \uparrow \mathcal{F}_\infty \subseteq \mathcal{F}$, then for any $A \in \mathcal{F}_\infty$, we can find $A_i \in \mathcal{F}_i$ such that $\mathbb{P}(A \triangle A_i) \to 0$ (where Δ is the symmetric difference).

(This can be proven by a π-λ argument, because the set of A for which this is true forms a λ-system, and $\bigcup_i \mathcal{F}_i$

is a π-system.) We will apply this to our product space by defining (for all n)

$$\mathcal{F}_n = \left\{ E_1 \times E_2 \times \cdots \times E_n \times \prod_{i>n} S : E_1, \cdots, E_n \in \mathcal{S} \right\}.$$

We have $\mathcal{F}_n \uparrow \mathcal{F}$ by definition of the product σ-field, so for all $A \in \mathcal{F}$, we can find $A_n \in \mathcal{F}_n$ such that $\mathbb{P}(A \triangle A_n) \to 0$ (by Fact 213). Now define $\pi_n : \mathbb{N} \to \mathbb{N}$ to swap k and $n+k$ for all $1 \le k \le n$ and keep the other coordinates fixed, and let $B_n = (A_n)_{\pi_n}$. For any event $A \in \mathcal{E}$, we have

$$\mathbb{P}(A \triangle B_n) = \mathbb{P}((A \triangle B_n)_{\pi_n}) = \mathbb{P}(A \triangle A_n),$$

where the first equality comes from the measure $\mathbb{P}$ being invariant under permutations of indices (because it is a product iid measure), and the second equality comes from A being permutable (we can check the details ourselves). On the other hand, A_n depends only on the first n coordinates, while B_n depends on coordinates $n+1$ through $2n$, so $A_n \perp\!\!\!\perp B_n$. Thus,

$$\mathbb{P}(A_n \cap B_n) = \mathbb{P}(A_n)\mathbb{P}(B_n) \implies \mathbb{P}(A) = \mathbb{P}(A)^2,$$

because $\mathbb{P}(A_n), \mathbb{P}(B_n), \mathbb{P}(A_n \cap B_n) \to \mathbb{P}(A)$ if the symmetric differences go to zero. □

Proof with reverse martingales. Again, we start with a useful result:

Lemma 214
If $\mathcal{G}, \mathcal{H}$ are any two sub-σ-fields of $\mathcal{F}$, and W is a random variable with $W \perp\!\!\!\perp \mathcal{H}$ and $\mathbb{E}[W|\mathcal{G}] \in \mathcal{H}$, then $\mathbb{E}[W|\mathcal{G}] = \mathbb{E}[W]$ is constant.

Proof of lemma. The conditional expectation $Y = \mathbb{E}[W|\mathcal{G}]$ is $\mathcal{G}$-measurable by definition and also $\mathcal{H}$-measurable by assumption. Define the event

$$A = \{Y - \mathbb{E}[W] \ge \varepsilon\} \in \mathcal{G} \cap \mathcal{H}.$$

Since $A \in \mathcal{G}$, the definition of conditional expectation tells us that

$$\mathbb{E}[W; A] = \mathbb{E}[Y; A] \ge (\mathbb{E}[W] + \varepsilon) \cdot \mathbb{P}(A).$$

But we also have $A \in \mathcal{H}$, and W is independent of $\mathcal{H}$, so we also have

$$\mathbb{E}[W; A] = \mathbb{E}[W]\mathbb{P}(A).$$

Putting these together, $\mathbb{E}(W)\mathbb{P}(A) \ge (\mathbb{E}(W) + \varepsilon)\mathbb{P}(A)$, which is a contradiction unless $\mathbb{P}(A) = 0$. We can similarly show that $Y - \mathbb{E}[W] \le -\varepsilon$ with probability zero for any $\varepsilon > 0$. Thus we indeed have $Y = \mathbb{E}[W]$ as desired. □

Turning to the proof, we will show that $\mathcal{E} \perp\!\!\!\perp \mathcal{F}_k$ for all k (with $\mathcal{F}_k$ as defined in the previous proof), which implies that $\mathcal{E} \perp\!\!\!\perp \mathcal{F}$ by a π-λ argument. Since $\mathcal{F}$ contains $\mathcal{E}$, this shows that $\mathcal{E}$ is independent of itself, so $\mathbb{P}(A)^2 = \mathbb{P}(A)$ for all $A \in \mathcal{E}$ just like before.

To show that $\mathcal{E} \perp\!\!\!\perp \mathcal{F}_k$, it suffices to show that for any bounded measurable function $\phi : S^k \to \mathbb{R}$, if we let $W = \phi(X_1, \cdots, X_k) \in \mathcal{F}_k$, then we have $\mathbb{E}[W|\mathcal{E}] = \mathbb{E}[W]$. (This is indeed sufficient because we can choose W to be the indicator for an arbitrary event in $\mathcal{F}_k$.) Because $\mathcal{E}_n \downarrow \mathcal{E}$, we know that $\mathbb{E}[W|\mathcal{E}_n] \to \mathbb{E}[W|\mathcal{E}]$ almost surely and in L^1 by Corollary 206. Now for an alternate way of studying $\mathbb{E}[W|\mathcal{E}_n]$, consider the random variable $A_n(\phi)$ (for $n \ge k$) defined by

$$A_n(\phi) = \frac{1}{(n)_k} \sum_{\underline{i} \in [n]_k} \phi(X_{i_1}, \cdots, X_{i_k}),$$

in which we take an average of ϕ over all permutations of the indices $\{1, \cdots, n\}$ (and use the first k of them). Because of this averaging, permuting the first n indices does not change the value of $A_n(\phi)$, so $A_n(\phi) \in \mathcal{E}_n$ and thus

$$\boxed{A_n(\phi)} = \mathbb{E}[A_n(\phi)|\mathcal{E}_n] = \frac{1}{(n)_k} \sum_{\underline{i} \in [n]_k} \mathbb{E}\left[\phi(X_{i_1}, \cdots, X_{i_k})|\mathcal{E}_n\right].$$

But because the different X_is are iid, all of these expectations are the same, so this reduces to $\mathbb{E}[\phi(X_1, \cdots, X_k)|\mathcal{E}_n] = \boxed{\mathbb{E}[W|\mathcal{E}_n]}$. Thus we can study convergence of $A_n(\phi)$ instead of $\mathbb{E}[W|\mathcal{E}_n]$ directly. We claim that $\lim_{n\to\infty} \mathbb{E}[W|\mathcal{E}_n] \in \mathcal{F}_{(k+1)+}$. Indeed, consider the terms in $A_n(\phi)$ as $n \to \infty$. The probability that $(i_1, \cdots, i_k)$ has any intersection with $(1, \cdots, k)$ goes to 0 as $n \to \infty$ (it's at most $\frac{k^2}{n}$ by a union bound). So the **limiting** expectation is independent of $\sigma(X_1, \cdots, X_k)$ (since the total contribution to the expectation from terms including any of $X_1, \cdots, X_k$ is then at most $\frac{k^2}{n} \sup(\phi)$, which goes to zero as $n \to \infty$), and thus $\lim_{n\to\infty} \mathbb{E}[W|\mathcal{E}_n] \in \mathcal{F}_{(k+1)+}$.

However, remembering that this limit is also $\mathbb{E}[W|\mathcal{E}]$, we can apply Lemma 214 with $\mathcal{G} = \mathcal{E}$ and $\mathcal{H} = \mathcal{F}_{(k+1)+}$. Specifically, W is indeed independent of $\mathcal{F}_{(k+1)+}$ by construction (because it is $\mathcal{F}_k$-measurable), and $\mathbb{E}[W|\mathcal{E}] \in \mathcal{F}_{k+1}$, so $\mathbb{E}[W|\mathcal{E}] = \mathbb{E}[W]$, finishing the proof. □

For a final application of the reverse martingale, recall the strong law of large numbers, which says that for any iid random variables X, X_i with $\mathbb{E}[X] = \mu$, we have $\frac{1}{n}\sum_{i=1}^n X_i \to \mu$ almost surely. Here's a shorter proof than the one we last showed:

Alternate proof of Theorem 88. Let $S_n = X_1 + \cdots + X_n$ be the usual partial sum, and define $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \cdots)$ and $\mathcal{G}_n \downarrow \mathcal{G}$ (as usual, this is a σ-algebra because it's the intersection of σ-algebras). Define $M_n = \frac{S_n}{n}$, and notice that

$$\mathbb{E}[M_n|\mathcal{G}_{n+1}] = \mathbb{E}\left[\frac{S_n}{n}\middle| S_{n+1}, S_{n+2}, S_{n+3}, \cdots,\right] = \mathbb{E}\left[\frac{S_n}{n}\middle| S_{n+1}, X_{n+2}, X_{n+3}, \cdots\right].$$

But if we know the value of S_{n+1}, also knowing $X_{n+2}, X_{n+3}, \cdots$ do not give us any additional information about $\frac{S_n}{n}$, so

$$\mathbb{E}[M_n|\mathcal{G}_{n+1}] = \mathbb{E}\left[\frac{S_n}{n}\middle| S_{n+1}\right] = \frac{S_{n+1}}{n+1} = M_{n+1}.$$

So M_n is a reverse martingale with respect to $\mathcal{G}_n$, meaning that M_n converges almost surely to $M_\infty = \mathbb{E}[M_1|\mathcal{G}_\infty] = \mathbb{E}[X_1|\mathcal{G}_\infty]$ by Theorem 205. But $\mathcal{G}_n \subseteq \mathcal{E}_n$ for all n (because the values of $S_n, S_{n+1}, \cdots$ don't change when we permute $X_1, \cdots, X_n$), so $\mathcal{G}_\infty \subseteq \mathcal{E}$. So by Theorem 212, $\mathbb{E}[X_1|\mathcal{G}_\infty]$ is a conditional expectation under a σ-algebra in which all probabilities are 0 or 1, so it must be $\mathbb{E}[X_1] = \mu$ almost surely. This is the desired result. □

23 December 2, 2019

Today is our last lecture before the second exam, so we'll cover some new material and have quite a bit of review. (There is no class on Wednesday – our exam is in the evening instead.)

Recall that if we have two finite measures μ, ν on $(\Omega, \mathcal{F})$, we say that ν is **absolutely continuous** with respect to μ (denoted $\nu \ll \mu$), if for all $A \in \mathcal{F}$ with $\mu(A) = 0$, we also have $\nu(A) = 0$. Then a measurable function $f : \Omega \to [0, \infty)$ satisfying

$$\nu(A) = \int_A f\, d\mu = \int_\Omega f(\omega) 1\{\omega \in A\} d\mu(\omega)$$

for all $A \in \mathcal{F}$ is called the **Radon-Nikodym derivative** of ν with respect to μ and denoted $f = \frac{d\nu}{d\mu}$. (An example of such a derivative from our proof of Cramér's theorem is the exponential tilting $\frac{d\mathbb{P}_\theta}{d\mathbb{P}} = \frac{\exp(\theta\sum_{i=1}^n X_i)}{\mathbb{E}(e^{\theta X})^n}$.) The Radon-Nikodym derivative doesn't exist for all μ, ν, but if it does, ν must be absolutely continuous with respect to μ, because

given any set A with $\mu(A) = 0$, $\nu(A) = \int_A f\,d\mu = 0$. It turns out that the converse is also true (we stated this as Theorem 152 earlier in the course):

Theorem 215 (Radon-Nikodym)
If μ, ν are σ-finite measures on $(\Omega, \mathcal{F})$ and $\nu \ll \mu$, then there is a measurable function f such that $\nu(A) = \int_A f\,d\mu$ for all $A \in \mathcal{F}$.

Notice that f is unique μ-almost-surely, because otherwise there would be some set A of positive measure such that the integrals over A differ (if f_1 and f_2 were two different Radon-Nikodym derivatives, consider $A = \{f_1 - f_2 \geq \varepsilon\}$ or $A = \{f_1 - f_2 \leq -\varepsilon\}$). Earlier in the class, we used this theorem (without proof) to construct conditional expectation. Specifically, if $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and we have a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, then we can assume without loss of generality that X is nonnegative and define $\mathbb{E}[X|\mathcal{G}] = \frac{d\nu}{d\mu}$, where $\mu = \mathbb{P}|_{\mathcal{G}}$ and $\nu(A) = \mathbb{E}_\mu[X; A]$ for all $A \in \mathcal{G}$. We can do this because ν is absolutely continuous with respect to μ (if A has measure zero, then $\mathbb{E}[X; A] = 0$ because any simple function has integral zero over A). This conditional expectation is $\mathcal{G}$-measurable and satisfies the conditional expectation property by plugging in the definitions of μ and ν into $\nu(A) = \int_A f\,d\mu$.

To start, we'll prove a few basic properties of absolutely continuous measures:

Lemma 216
If we have finite measures $\pi \ll \nu \ll \mu$ on $(\Omega, \mathcal{F})$, then $\frac{d\pi}{d\mu} = \frac{d\pi}{d\nu}\frac{d\nu}{d\mu}$, μ-almost-surely.

Proof. Notice that we do indeed have $\pi \ll \mu$, because $\mu(A) = 0 \implies \nu(A) = 0 \implies \pi(A) = 0$. We claim that for all nonnegative measurable functions h, we have

$$\int h\,d\nu = \int h\frac{d\nu}{d\mu}d\mu.$$

Indeed, this identity holds for any indicator $h = 1_A$ (by definition of the Radon-Nikodym derivative), so it holds for simple functions by linearity, and then we can approximate from below and use the monotone convergence theorem in general. Next, because π is absolutely continuous with respect to ν, we have

$$\pi(A) = \int_A \frac{d\pi}{d\nu}d\nu$$

for any event A. But $\frac{\partial d\pi}{d\nu}$ is a nonnegative measurable function, so by the claim above, we have

$$\pi(A) = \int_A \frac{d\pi}{d\nu}\frac{d\nu}{d\mu}d\mu.$$

Because the Radon-Nikodym derivative is unique almost-surely under μ, $\frac{d\pi}{d\nu}\frac{d\nu}{d\mu}$ must agree almost surely with $\frac{d\pi}{d\mu}$, as desired. □

Plugging in μ for π, or doing so while swapping the roles of μ and ν, yields the following result:

Lemma 217
If $\nu \ll \mu$ and $\mu \ll \nu$, then $\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu}\right)^{-1}$ almost surely under both μ and ν.

The full proof of Radon-Nikodym can be found in appendix A.4 of Durrett, and it's a bit outside the scope of the class. However, there is one case which is easy to prove, which we'll discuss now. Suppose we can partition our

probability space into countably many pieces $\Omega = \bigsqcup_{i=1}^{\infty} \Omega_i$, and $\mathcal{F} = \sigma(\Omega_1, \Omega_2, \cdots)$. Then for any finite measures μ, ν on Ω with $\nu \ll \mu$, we can define

$$\frac{d\nu}{d\mu}(\omega) = f(\omega) = \begin{cases} \frac{\nu(\Omega_i)}{\mu(\Omega_i)} & \text{when } \omega \in \Omega_i \text{ with } \mu(\Omega_i) > 0, \\ 0 & when \omega \in \Omega_i \text{ with } \mu(\Omega_i) = 0. \end{cases}$$

(We can replace 0 in the second case with any other number – that case occurs with probability zero, so it doesn't contribute to any relevant calculations.) We can indeed check both properties of the Radon-Nikodym derivative in this simple case, and thus we prove the theorem when we have a countable partition.

More generally, suppose our σ-algebra can be approximated with sets of the above form. In other words, consider a space $(\Omega, \mathcal{F})$, where $\mathcal{F}_n \uparrow \mathcal{F}$ and each $\mathcal{F}_n$ is generated by a countable partition of Ω. For example, if we take $\Omega = \mathbb{R}$ and $\mathcal{F} = \sigma\left(\frac{[i,i+1)}{2^n} : i \in \mathbb{Z}\right)$, then $\mathcal{F}_n$ increases to the entire Borel σ-algebra $\mathcal{B}_{\mathbb{R}}$ (because the Borel σ-algebra is generated by open intervals, and we can approximate open intervals with a countable union of these dyadic sets). So if we have two finite measures μ, ν on $(\Omega, \mathcal{F})$ such that $\nu \ll \mu$, we can define μ_n (resp. ν_n) be the restriction of μ (resp. ν) to $\mathcal{F}_n$. We have $\mu_n \ll \nu_n$ for all n, so $\frac{d\nu_n}{d\mu_n}$ exists by the above explicit construction, and it's natural to ask whether $\frac{d\nu_n}{d\mu_n} \to \frac{d\nu}{d\mu}$.

Remark 218. *It's possible to have $\mu_n \ll \nu_n$ but not $\mu \ll \nu$. For example, take μ to be the infinite product of Bernoulli(p) variables, and take ν to be the infinite product of Bernoulli(q) variables (for $p \neq q$). Then we do have $\mu_n \ll \nu_n$ for any n, but by the strong law of large numbers we do not have $\mu \ll \nu$ (the event that $\frac{S_n}{n} \to p$ has probability zero in ν but not in μ).*

Lemma 219

Let μ, ν be probability measures on $(\Omega, \mathcal{F})$ with $\mathcal{F}_n \uparrow \mathcal{F}$, and let $\mu_n = \mu|_{\mathcal{F}_n}$, $\nu_n = \nu|_{\mathcal{F}_n}$ for all n. Suppose that $\nu_n \ll \mu_n$ for all n, so $X_n = \frac{d\nu_n}{d\mu_n}$ is well-defined. (However, we're not assuming that $\nu \ll \mu$.) Then X_n is a martingale with respect to $\mathcal{F}_n$ on $(\Omega, \mathcal{F}, \mu)$.

Proof. Let $\mathbb{E}$ denote expectation with respect to μ. Because X_n depends only on ν_n and μ_n, it is in $\mathcal{F}_n$ by definition, so

$$\mathbb{E}[X_n] = \int \frac{d\nu_n}{d\mu_n} d\mu = \int \frac{d\nu_n}{d\mu_n} d\mu_n = 1$$

(middle equality because $\frac{d\nu_n}{d\mu_n} \in \mathcal{F}_n$ so restricting μ to μ_n is sufficient, and last equality by the Radon-Nikodym property for $A = \Omega$). Thus $X_n \in L^1(\mathcal{F}_n)$, showing integrability. For the conditional expectation property, let A be any event in $\mathcal{F}_n$. We have

$$\mathbb{E}[X_{n+1}; A] = \int_A X_{n+1} d\mu = \int_A X_{n+1} d\mu_{n+1} = \nu_{n+1}(A),$$

where the middle equality comes from X_{n+1} being in $\mathcal{F}_{n+1}$ and the last equality comes from the Radon-Nikodym definition. But $\nu_{n+1}(A) = \nu(A) = \nu_n(A)$ (all of these measures are restrictions of ν), so

$$\mathbb{E}[X_{n+1}; A] = \nu_n(A) = \int_A X_n d\mu_n = \mathbb{E}[X_n; A],$$

showing the conditional expectation property and verifying that we have a martingale. □

With this, we can answer our convergence question:

Theorem 220

Let μ, ν be probability measures on $(\Omega, \mathcal{F})$ (again not assuming absolute continuity), such that $\mathcal{F}_n \uparrow \mathcal{F}$, and again define $\mu_n = \mu|_{\mathcal{F}_n}$ and $\nu_n = \nu|_{\mathcal{F}_n}$ for all n. If $\nu_n \ll \mu_n$ with $X_n = \frac{d\nu_n}{d\mu_n}$ for all n, then $X = \limsup_{n\to\infty} X_n$ satisfies

$$\nu(A) = \int_A X\, d\mu + \nu(A \cap \{X = \infty\}).$$

The first term $\int_A X\, d\mu$ is also called $\nu_{\text{continuous}}(A)$ (because it is absolutely continuous with respect to μ), while the second term $\nu(A \cap \{X = \infty\})$ is called $\nu_{\text{singular}}(A)$ (meaning its support is disjoint from ν). Basically, X_n is a nonnegative martingale under the measure μ, so it converges almost surely to a finite limit under μ. However, it may not do so under the measure ν, explaining the "boundary" second term. (Specifically, even though $\mu(X = \infty) = 0$, we may have $\nu(X = \infty) > 0$.)

Once we prove this result, we'll have proved the Radon-Nikodym theorem in the special case where we have countable partitions $\mathcal{F}_n \uparrow \mathcal{F}$ increasing to the full σ-algebra (since ν_{singular} is identically zero if we already know that $\nu \ll \mu$, because then $\nu(X = \infty) = 0$).

Proof. The key idea is to introduce a measure that interpolates between μ and ν. Let $\rho = \frac{\mu+\nu}{2}$, and let $\rho_n = \rho|_{\mathcal{F}_n} = \frac{\mu_n+\nu_n}{2}$ for all n. By assumption, $\nu_n \ll \mu_n \ll \rho_n$ (because ρ_n is an average of μ_n and ν_n, so any event with measure zero under ρ_n must have measure zero under both μ_n and ρ_n), so the Radon-Nikodym derivatives

$$X_n = \frac{d\mu_n}{d\nu_n}, \quad Y_n = \frac{d\mu_n}{d\rho_n}, \quad Z_n = \frac{d\nu_n}{d\rho_n}$$

are well-defined. We have $Y_n + Z_n = 2$ ρ_n-almost-surely, and by Lemma 219, Y_n and Z_n are $\mathcal{F}_n$-martingales on $(\Omega, \mathcal{F}, \mathbb{P})$. Furthermore, because they are nonnegative, they converge to finite limits Y, Z ρ-almost-surely, and thus $Y + Z = 2$ ρ-almost-surely. (In particular, Y and Z are both bounded, so we have strong convergence results.) We claim that $Y = \frac{d\mu}{d\rho}$. Indeed, fix some positive integer ℓ; for all $n \geq \ell$ and any $A \in \mathcal{F}_\ell \subseteq \mathcal{F}_n$, we have (by the definition of Y_n, which is in $\mathcal{F}_n$)

$$\boxed{\mu(A)} = \mu_n(A) = \int_A Y_n d\rho_n = \int_A Y_n d\rho,$$

which converges to $\boxed{\int_A Y\, d\rho}$ by the bounded convergence theorem (notice that this last step doesn't work for X). Thus, for any $A \in \mathcal{F}_\ell$, we have $\mu(A) = \int_A Y\, d\rho$, so by a π-λ argument we have $\mu(A) = \int_A Y\, d\rho$ for all $A \in \mathcal{F} = \sigma\left(\bigcup_\ell \mathcal{F}_\ell\right)$, so $Y = \frac{d\mu}{d\rho}$. Similarly, $Z = \frac{d\nu}{d\rho}$. Now $\mu_n \ll \mu_n \ll \rho_n$, so by Lemma 216 we have $\frac{d\nu_n}{d\rho_n} = \frac{d\nu_n}{d\mu_n}\frac{d\mu_n}{d\rho_n}$ almost surely – that is, we have $Z_n = X_n Y_n$ almost surely, so $X_n = \frac{Z_n}{Y_n} \in [0, \infty]$ (where $X_n = 0$ only if $Z_n = 0$, and $X_n = \infty$ only if $Y_n = 0$). Thus, $X = \limsup_{n\to\infty} X_n = \frac{Z}{Y}$ (we have lim sup instead of lim in case of division-by-zero issues).

Turning back to the statement we are trying to prove, we know that $\nu \ll \rho$ with Radon-Nikodym derivative Z, so we have

$$\nu(A) = \int_A Z\, d\rho = \int_A Z\left(WY + 1\{Y = 0\}\right) d\mathbb{P} = \int_A ZWY\, d\rho + \int_A Z 1\{Y = 0\} d\rho,$$

where $W = \frac{1}{Y}$ when $Y > 0$ and 0 otherwise. (We break up the integral in this way because $X = \infty$ corresponds to $Y = 0$, and we're trying to get X into the integral.) But now the first term is $\int_A XY\, d\rho = \int_A X\frac{d\mu}{d\rho} d\rho = \int_A X\, d\mu$, because $ZW = X$ on the event that Y is positive (when $Y = 0$ there is no contribution to the integral), and we don't need to worry about the event $\{Y = 0\} = \{X = \infty\}$ in this term now that we're integrating under μ-measure because X is finite μ-almost-surely. And because $Z = \frac{d\nu}{d\rho}$, the second term simplifies to $\int_A 1\{Y = 0\}\frac{d\nu}{d\rho} d\rho = \int_A 1\{X = \infty\} d\nu$, which is indeed $\nu(A \cap \{X = \infty\})$, as desired. □

24 December 9, 2019

There won't be any more assignments for this class – in this last week, we'll cover a few topics of general interest. Today, we'll go over the proof of the Kesten-Stigum $L \log L$ criterion – the initial work comes from various papers by Kesten and Stigum (including [5]), and many of the ideas covered in this lecture come from [7].

We'll start with a review of the setup: say that $\xi \sim p$ is a probability distribution supported on $\mathbb{Z}_{\geq 0}$ (ξ is the L in the theorem statement). Define an array $\xi_{n,i}$ of iid random variables with law p, and define the Galton-Watson tree by starting from a root vertex o (at depth 0) and letting the ith vertex at depth $(n-1)$ have $\xi_{n,i}$ children at depth n. This generates a (finite or infinite) tree, and $\mathcal{F}_n = \sigma(\xi_{\ell,i} : i \geq 1, 1 \leq \ell \leq n)$ encodes all the information about the tree.

Let Z_n be the population of the tree at depth n, which satisfies the recursive equation $Z_n = \sum_{i=1}^{Z_{n-1}} \xi_{n,i}$. If we assume that $\mathbb{E}[\xi] = m \in (1, \infty)$ (so we're in the supercritical case), then we know from previous discussion that the probability of extinction (that is, the probability that $Z_n = 0$ eventually) is some $\rho \in [0, 1)$, meaning that there is a positive probability that the population never dies out.

To understand the behavior of Z_n in more detail, we know that $W_n = \frac{Z_n}{m^n}$ is a nonnegative martingale, so it converges almost surely to a limit $W \in [0, \infty)$. Because $\mathbb{P}(W = 0)$ satisfies the same fixed point equation as the extinction probability, it is either ρ or 1. Then if $\mathbb{P}(W = 0) = 1$, then $Z_n \ll m^n$ almost surely, and if $\mathbb{P}(W = 0) = \rho$, we have $\mathbb{P}(W = 0 \,|\, \text{nonextinction}) = 0$ (because whenever $Z = 0$, we also have $W = 0$), so W is positive and finite almost surely. So conditioned on nonextinction, $Z_n \asymp m^n$ and we have exponential growth.

Kesten-Stigum then tells us when each of these cases occurs (this is a restatement of Theorem 183)

Theorem 221 (Kesten-Stigum)
Suppose $(Z_n)_{n \geq 0}$ is a supercritical Galton-Watson tree with offspring law $\xi \sim p$, where $\mathbb{E}[\xi] = m \in (1, \infty)$. Let $W_n = \frac{Z_n}{m^n}$ converge to W almost surely. Then the following are equivalent:

- $\mathbb{P}(W = 0) = \rho$,
- $\mathbb{E}[W] = 1$,
- $\mathbb{E}[\xi \log^+ \xi]$ is finite (where $\log^+(0) = 0$ but otherwise we have the usual log).

On our homework, we showed that if $\mathbb{E}[\xi^2] < \infty$, then the martingale W_n is bounded uniformly in L^2 norm, so the L^2 martingale convergence theorem says that W_n converges in L^2 to W, and hence also in L^1. This means $\mathbb{E}[W] = 1$, so we can't have $\mathbb{P}(W = 0) = 1$ and thus must have $\mathbb{P}(W = 0) = \rho$. This theorem is a stronger version of that result.

One fact we'll use in the proof of Kestin-Stigum also comes from our homework: we showed that if Y, Y_i are iid nonnegative random variables, then

$$\limsup_{n \to \infty} \frac{Y_n}{n} = \begin{cases} 0 & \text{if } \mathbb{E}[Y] < \infty, \\ \infty & \text{if } \mathbb{E}[Y] = \infty. \end{cases}$$

(The proof is that for any $x > 0$, $\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{Y_n}{n} \geq x\right)$ is sandwiched between $\mathbb{E}\lfloor \frac{Y}{x} \rfloor$ and $\mathbb{E}\lceil \frac{Y}{x} \rceil$, which either both go to 0 or both go to ∞, and we can use Borel-Cantelli in either case.) As an easy consequence, we also have that if Y, Y_i are nonnegative integer valued random variables, then for any $c > 1$,

$$\limsup_{n \to \infty} \frac{Y_n}{c^n} = \begin{cases} 0 & \mathbb{E}[\log^+ Y] < \infty \\ \infty & \mathbb{E}[\log^+ Y] = \infty. \end{cases}$$

(This is the same proof as before but rewriting $\mathbb{P}(\frac{Y_n}{c^n} \geq x) = \mathbb{P}\left(\log_c\left(\frac{Y_n}{x}\right) \geq n\right)$ – we have to deal with the case $Y_n = 0$ separately, but that changes the expectation by at most 1.) This allows us to define a new process:

Definition 222

In a **branching process with immigration**, the population X_n at level n is given by

$$X_n = \sum_{i=1}^{X_{n-1}} \xi_{n,i} + Y_n,$$

where $\xi_{n,i}$ are iid as before and Y, Y_i are iid (corresponding to the additional immigrating population).

Fact 223

On our homework, we showed that if $\mathbb{E}[\log^+ Y] = \infty$, then $\limsup_{n\to\infty} \frac{X_n}{c^n}$ is infinite for all $c > 1$, and if $\mathbb{E}[\log^+ Y] < \infty$, then $\limsup_{n\to\infty} \frac{X_n}{m^n}$ converges to a finite limit $\in (0, \infty)$. The first case is easy by the previous discussion, and the second can be shown with submartingale convergence theorem.

We'll spend the rest of today linking these two results together. One key component comes from an exam question and was also proved in last lecture as Theorem 220: basically, if μ, ν are finite measures on $(\Omega, \mathcal{F})$, and we have a filtration $\mathcal{F}_n \uparrow \mathcal{F}$, then we can write $\mu_n = \mu|_{\mathcal{F}_n}$ and $\nu_n = \nu|_{\mathcal{F}_n}$. The result is that if $\nu_n \ll \mu_n$ for all n, then we can write

$$\nu(A) = \int_A W d\mu + \nu(A \cap \{W = \infty\}) = \nu_{\text{continuous}} + \nu_{\text{singular}},$$

where $W_n - \frac{d\nu_n}{d\mu_n}$ and $W = \limsup W_n$. There are two extreme cases of this: having $W = 0$ μ-almost-surely is the same as having $\nu_{\text{continuous}} = 0$ and $\nu = \nu_{\text{singular}}$ singular with respect to μ, which is the same as having $\nu(W = \infty) = 1$ by setting $A = \Omega$. On the other hand, if $\int_\Omega W d\mu = 1$, meaning that $\nu_{\text{continuous}}(\Omega) = 1$, that's equivalent to having absolute continuity $\nu \ll \mu$ and $\nu(W = \infty) = 0$.

To prove Theorem 221, we'll apply this decomposition result on $\Omega = \{\text{rooted graphs}\}$ (which contains the rooted trees) and $\mathcal{F}_n = \sigma(\text{trees up to depth } n)$. Ω can be made into a Polish space under the isomorphism distance (also on our homework), and $\mathcal{F}_n$ increases to $\mathcal{F}$, the Borel sigma-algebra on the metric space. The idea is to let $\mu =$ Galton-Watson measure with offspring law p (meaning that we just generate a Galton-Watson tree in the ordinary way under μ) and define ν so that $\frac{d\nu_n}{d\mu_n} = W_n = \frac{Z_n}{m^n}$; we'll then proceed by looking at this process under the measure ν. In principle, ν is already well-defined, since we know μ_n (and thus ν_n) for every n. But ν itself is a bit confusing, and we don't really know how to analyze a statement like $\nu(W = \infty) = 0$ yet. So we'll do some preliminary work:

- First of all, if T_n is the Galton-Watson tree T truncated at depth n, then for any event $B \in \mathcal{F}_n$,

  $$\nu(T_n \in B) = \nu_n(T_n \in B) = \int 1\{T_n \in B\} \frac{Z_n}{m^n} d\mu_n$$

  by the definition of the Radon-Nikodym derivative. In particular, letting B be the event that T dies by level n, this integral is just $\int 0\, d\mu_n = 0$ (because $Z_n = 0$ on B). So the tree survives forever under ν (because $\nu(\text{extinction}) = 0$ – this is not true under μ.

- We can define ν_1 explicitly: we have (remembering that $m = \mathbb{E}[\xi]$)

  $$\frac{d\nu_1}{d\mu_1} = W_1 = \frac{\xi_{1,1}}{m},$$

  meaning that the original measure is biased by the number of offspring at the first level. In other words, ν_1 is the law of T_1 with a **size bias**, and the probability of having k offspring is $\hat{p}_k = \frac{kp_k}{m}$ for any $k \geq 1$. More generally,

here's a description of ν_n: let ν_n^* be a probability measure on pairs (T_n, x_n), where T_n is a tree of depth n and x_n is a vertex at depth n in T_n (meaning that we pick out a distinguished vertex at level n), given by

$$\nu_n^*(T_n, x_n) = \frac{\mu_n(T_n)}{\text{normalization}},$$

where the normalization factor is

$$\sum_{T_n, x_n} \mu_n(T_n) = \sum_{T_n} \mu_n(T_n) \sum_{x_n} 1 = \sum_{T_n} \mu_n(T_n) Z_n(T_n) = m^n,$$

because this is the expression for the expected number of children $\mathbb{E}[Z_n]$ under the normal Galton-Watson measure. Now if $\tilde{\nu_n}$ is the marginal of ν_n^* on T_n (meaning that we forget which distinguished vertex we choose), we have

$$\tilde{\nu_n}(T_n) = \sum_{x_n} \nu_n^*(T_n, x_n) = \sum_{x_n} \frac{\mu_n(T_n)}{m^n} = \mu_n(T_n) \frac{Z_n(T_n)}{m^n}.$$

In particular, this means that $\frac{d\tilde{\nu_n}}{d\mu_n} = \frac{Z_n}{m^n} = W_n$, so $\tilde{\nu_n}$ is indeed the ν_n that we've been looking for – ν_n can be obtained from μ_n by sampling a tree with a marked vertex and then forgetting the position of that vertex.

Nothing here seems to actually help yet (we still don't have a way to describe ν), but the key observation is the following: we can sample $(T_n, x_n) \sim \nu_n^*$ for each n (which we can think of as T_n with a (unique) marked path ζ_n from the root o to x_n). We can also sample a tree (T, ξ) with a marked ray in the following way:

- Start from the root and have it generate $\hat{\zeta} \sim \hat{p}$ children (from the size-biased law). Choose one of them, x_1, to be on the marked path.
- For all other children except x_1, generate children according to the original law p.
- Generate children for x_2 according to the size-biased law $\hat{p}$, and choose one of them, x_2, to be on the marked path.
- Repeat this indefinitely, generating children according to p for any vertex not on the marked path and according to $\hat{p}$ for any vertex on it.

We can continue this process indefinitely, producing an infinite tree with a marked ray – importantly, this is guaranteed to be an infinite tree because the size-biased law is supported on the **positive** integers, so an x_i will always exist. The infinite tree (T, ζ) has some law ν_*, and we claim that this agrees in law with (T_n, x_n) – in other words, we claim that $\nu_*|_{\mathcal{F}_n} = \nu_n^*$ for all n, where the right hand side is defined (as above) to be $\frac{\mu_n(T_n)}{m^n}$. Write the left-hand side as $\nu_*|_{\mathcal{F}_n} = \nu_n^\circ$ – by our sampling process, we have

$$\nu_n^\circ(T_n, \zeta_n) = \frac{k p_k}{m} \cdot \frac{1}{k} \cdot \left(\prod_{x \neq x_1} \mu_{n-1}(T_{n-1}^{(x)}) \right) \cdot \nu_{n-1}^\circ(T_{n-1}^{(x_1)}),$$

because we generate k children for the root vertex under the size-bias law, pick one of them at random, construct a regular Galton-Watson tree rooted at x for all children $x \neq x_{n-1}$, and perform the special sampling procedure up to level $(n-1)$ for the marked vertex x_1. We can compare this recursion to the one for ν_n^*, which is

$$\nu_n^*(T_n, \zeta_n) = \frac{\mu(T_n)}{m^n} = \frac{p_k}{m^n} \prod_{x \text{ at depth } 1} \mu_{n-1}(T_{n-1}^{(x)}),$$

because other than the m^n factor, we're just doing the normal Galton-Watson sampling. This can be rewritten as

$$\nu_n^*(T_n, \zeta_n) = \frac{p_k}{m}\left(\prod_{x \neq x_1} \mu_{n-1}(T_{n-1}^{(x)})\right)\frac{\mu_{n-1}(T_{n-1}^{(x_1)})}{m^{n-1}}.$$

Noticing that this is the same recursion as for ν_n°, and the two measures agree at $n = 1$, induction tells us that $\nu_n^\circ = \nu_n^*$ for all n. So now we have a measure ν^* on (T, ζ), and we can just forget ζ by letting ν be the marginal of ν^* on T. Then ν restricted to $\mathcal{F}_n$ is the marginal of $\nu^*|_{\mathcal{F}_n}$, which is the marginal of $\nu_n^*|_{\mathcal{F}_n}$, which is ν_n by our explicit calculations above, so we have the desired $\nu|_{\mathcal{F}_n} = \nu_n$.

Now note that if $(T, \zeta) \sim \nu^*$, then $T \setminus \zeta$ is a branching process with immigration because the unmarked children coming off of ζ (that is, the siblings of the marked children) are "immigrating" into the tree as $Y_1, Y_2, \cdots$. Specifically, if X_n is the number of children at level n of the tree except for the marked ray, we have

$$X_n = \sum_{i=1}^{X_{n-1}} \xi_{n,i} + Y_n,$$

where the Y_ns are iid with law $\hat{\xi} - 1$ (since we have the size-biased distribution but don't count the marked vertex). And by Fact 223, $\limsup \frac{X_n}{m^n}$ depends on $\mathbb{E}[\log^+ Y]$, and

$$\mathbb{E}[\log^+ Y] = \mathbb{E}[\log^+(\hat{\xi} - 1)] = \frac{\mathbb{E}[\xi \log(\xi - 1)]}{m},$$

which is finite if and only if $\mathbb{E}[\xi \log \xi]$ is finite! So we can now put everything together: Ω is the space of rooted graphs, $\mathcal{F}_n$ are the sigma-algebras of the tree up to level n, μ is the Galton-Watson law with offspring law $\xi \sim p$, ν^* is the law on pairs (T, ζ) as defined above, and ν is the marginal of ν^* on T. For all n, μ_n, ν_n are the restrictions of μ, ν to $\mathcal{F}_n$, and $\frac{d\nu_n}{d\mu_n} = W_n = \frac{Z_n}{m^n}$. And now if $(T, \zeta) \sim \nu^*$, then $T \setminus \zeta$ is a branching process with immigration law $Y = \hat{\xi} - 1$, so

$$\mathbb{E}[\xi \log^+ \xi] < \infty \iff \mathbb{E}[\log^+ Y] < \infty \iff \nu^*\left(\lim \frac{X_n}{m^n} < \infty\right) = 1$$

by Fact 223. Since X_n is the number of children at the nth level except for one of the vertices, we can replace $\frac{X_n}{m^n}$ with $W_n = \frac{Z_n}{m^n}$ (because the extra $\frac{1}{m^n}$ is negligible). And if an event occurs with probability 1 under ν^*, it also occurs with probability 1 under the marginal ν. Thus we can continue the chain of statements and write

$$\mathbb{E}[\xi \log^+ \xi] < \infty \iff \nu\left(\lim_{n\to\infty} W_n < \infty\right) = 1 \iff \nu \ll \mu \text{ and } \int W d\mu = 1,$$

where the last step comes from the "extreme decomposition case" we mentioned above where $\mathbb{P}(W = \infty) = 0$. So because $\mathbb{P}(W = 0)$ is either ρ or 1, but the latter case would yield $\int W d\mu = 0$, $\mathbb{E}[\xi \log^+ \xi]$ being finite is indeed equivalent to $\mathbb{P}(W = 0) = \rho$. And finally, writing a similar chain in terms of the notation from before (but now assuming infinite $\mathbb{E}[\xi \log^+ \xi]$), we have

$$\mathbb{E}[\xi \log^+ \xi] = \infty \iff \mathbb{E}[\log^+ Y] = \infty \iff \nu^*\left(\limsup \frac{X_n}{m^n} = \infty\right) = 1 \iff \nu(\limsup W_n = \infty) = 1.$$

But $W = \limsup W_n$ being infinite almost surely corresponds to the other extreme case where ν is singular with respect to μ, meaning $W = 0$ μ-almost surely and thus $\mathbb{E}[W]$ cannot be 1. Therefore $\mathbb{E}[\xi \log^+ \xi]$ being finite is also equivalent to $\mathbb{E}[W] = 1$, as desired.

25 December 11, 2019

Today, we'll cover the $\zeta(2)$ limit in the **random assignment problem**. We'll use some of the techniques we've seen in this class, and the results are very beautiful! There are two formulations of the problem that are equivalent:

Problem 224 (Version 1)

Let $C \in \mathbb{R}^{n\times n}$ be a random matrix with iid entries each with distribution $n \cdot \text{Exp}$ (so in particular the entries have mean n). Define

$$A_n = \frac{1}{n} \min_{\sigma \in S_n} \left(\sum_{i=1}^{n} C_{i,\sigma(i)} \right).$$

In other words, choose a rook-placement with minimum total sum of entries, and look at the average of those entries. We wish to study $\mathbb{E}[A_n]$.

Problem 225 (Version 2)

Consider a complete bipartite graph $G_{n,n}$ (where we have n vertices on the left and right and we draw all n^2 edges between the left and the right). Think of the vertices on the left side as indexing rows and vertices on the right side as indexing columns, so C_{ij} is the weight of the edge connecting i on the left to j on the right. Letting each edge have weight $n \cdot \text{Exp}$ as before, we can define

$$A_n = \frac{1}{n} \min_{M} \left\{ \sum_{i=1}^{n} C_{i,M(i)} \right\}$$

where we take the minimum over **perfect matchings** M (which are a subset of edges such that every vertex is adjacent to exactly one of those edges), and again we want to study $\mathbb{E}[A_n]$.

Here's the result (from [2]) that we'll be covering today:

Theorem 226 (Aldous)

In the setting above, we have $\lim_{n\to\infty} \mathbb{E}[A_n] = \sum_{i=1}^{\infty} \frac{1}{i^2} = \zeta(2) = \frac{\pi^2}{6} \approx 1.6$.

Fact 227

This answer was first conjectured in 1987 by Mezard and Parisi, and previous work had showed that

$$1 + \frac{1}{e} < 1.51 \le \liminf \mathbb{E}[A_n] \le \limsup \mathbb{E}[A_n] \le 1.94 < 2.$$

Various "famous people" worked on these bounds, so many people cared about the answer to this problem. The fact that the lim inf and lim sup are equal came from earlier work by Aldous in [1], and further work (see [6] and [8]) showed $\mathbb{E}[A_n] = \sum_{i=1}^{n} \frac{1}{i^2}$. Today, we'll just show the original proof that the limit is $\frac{\pi^2}{6}$, but improvements (see [11]) have further simplified the proof since then.

First of all, we may notice that $\mathbb{E}[A_n]$ does not scale with n, so we'll start by explaining that. If we look at this problem as a bipartite graph $G_{n,n}$ with weights $C_{ij} \sim n \cdot \text{Exp}$, then for any matching M, we can define its **cost**

$$\text{cost}(M) = \frac{1}{n} \sum_{i=1}^{n} C_{i,M(i)}.$$

For any deterministic matching M, the expected value of the cost is $\frac{1}{n}\sum_{i=1}^{n} n = n$, but we're saying that the expected cost of the best matching still stays constant. To understand this, one heuristic is to think about the edges incident to a single vertex i on the left side, whose weights are n independent random variables $C_1, \cdots, C_n$. By definition, for all $t \geq 0$, we have

$$\mathbb{P}(C_i \geq t) = e^{-t/n} \implies \mathbb{P}\left(\min_{1\leq i\leq n} C_i \geq t\right) = \left(e^{-t/n}\right)^n = e^{-t},$$

which is the cdf of a **standard** exponential variable. Let $C_{(1)}$ be the smallest of the C_is, $C_{(2)}$ be the second smallest, and so on (these are all still connected to a fixed vertex i). Since the exponential distribution is **memoryless** – that is, $\mathbb{P}(C \geq s + t | C \geq s) = \mathbb{P}(C \geq t)$ – we can condition on the value of $C_{(1)} = \min_{1\leq i\leq n} C_i$. Then we have $n-1$ weights left, and we know they're all exponential and larger than $C_{\min}$, so the memoryless property tells us that

$$C_{(2)} \stackrel{d}{=} C_{(1)} + \frac{n}{n-1} \cdot \text{Exp}$$

(here we use the fact that the minimum of k iid $n \cdot \text{Exp}$ random variables is distributed as $\frac{n}{k} \cdot \text{Exp}$). Repeating this process, if $(C_{(1)}, \cdots, C_{(n)})$ is the sorted list of weights attached to a vertex on the left, we have

$$(C_{(1)}, \cdots, C_{(n)}) \stackrel{d}{=} \left(E_1, E_1 + \frac{n}{n-1}E_2, \cdots, E_1 + \frac{n}{n-1}E_2 + \cdots + nE_n\right),$$

where the E_is are iid exponential random variables. As $n \to \infty$, this (heuristically) converges to $(E_1, E_1 + E_2, E_1 + E_2 + E_3, \cdots)$ which is a collection of points on the real line that can be described by a **Poisson point process of rate 1** (in which the distance between points is given by an exponential random variable, and the number of points within any interval $[a, b]$ is given by a Poisson distribution). So in summary, if we look at the edges incident to a specific vertex i, the first 100 smallest edges will be of constant order as n gets large, so it's pretty reasonable to assume constant-order weight for each $C_{i,M(i)}$ in our matching.

This is just a heuristic, though, and what we've said doesn't really imply that the limit of $\mathbb{E}[A_n]$ should go to a constant. So we have to be more specific:

Definition 228

The **Poisson-weighted infinite tree** is constructed as follows: let Π be the random collection of points $(E_1, E_1 + E_2, E_1 + E_2 + E_3, \cdots)$, where the E_i are iid exponential. In our tree, each vertex has infinitely many children indexed by $\mathbb{N}$, and the weights of the connecting edges are randomly sampled iid as Π for each vertex. (So the first edge from each vertex is the lightest and given by E_1, the next edge is given by $E_1 + E_2$, and so on.)

We define a matching $\mathcal{M}$ on this infinite tree in the same way as on our graph – it's a subset of the edges such that every vertex is covered exactly once. It turns out the graph $G_{n,n}$ in our problem converges locally weakly to the Poisson-weighted infinite tree T – we did the calculation for depth 1 above, and there's a lot of independence after that. So the main idea of Aldous' proof is to consider the random bipartite graph $(G_{n,n}, M^*)$, where M^* is the best (lowest-cost) matching for $G_{n,n}$. Since $G_{n,n}$ converges locally weakly to T, and M^* is just a collection of 1s and 0s (telling us whether each of the edges are included), it seems reasonable to expect that we have the local weak convergence

$$(G_{n,n}, M^*) \stackrel{\text{lwc}}{\to} (T, \mathcal{M}^*).$$

Trees are easier to analyze than complete bipartite graphs, so that motivates using the infinite tree to attack this problem! One way to get a matching on the tree T is to always pick the lightest edge from the root, and then for all children that aren't connected, connect them to their lightest child, and so on. Because the lightest edge has expectation 1, the average cost of this matching must be 1, which is less than $\frac{\pi^2}{6}$. So something has gone wrong –

the point is that $\mathcal{M}^*$ has to arise from a local weak limit process, which is not true here! In particular, any local weak limit should be **spatially invariant**, because the choice of root of our tree should not play any special role in whether an edge belongs in the matching by the definition of local weak convergence. We'll use the next result without proof:

Theorem 229 (Aldous)
If $W_{\phi,M(\phi)}$ denotes the weight of the edge from the root ϕ to $M(\phi)$, then

$$\lim_{n\to\infty} \mathbb{E}[A_n] = c^* = \inf_{\text{spatially invariant } M} \mathbb{E}[W_{\phi,M(\phi)}].$$

In words, consider any spatially invariant matching M of our infinite tree. Because all edges are equivalent, the average cost can be defined as the expected cost of the edge next to the root! One direction of this theorem, showing that $\liminf \mathbb{E}[A_n] \geq c^*$, is more straightforward, because we can use a local weak limit argument (which is essentially compactness). Specifically, we can show that a subsequence of $(G_{n,n}, M^*)$ converges locally weakly to (T, M) for some spatially invariant matching M, and then the average cost for M must be at least c^* because c^* is optimal. But the other direction is more difficult because we have to produce a good approximation for the ideal c^* on a finite graph, and we won't go through it here.

Instead, in the rest of this lecture, we'll show that $c^* = \frac{\pi^2}{6}$. Here's a heuristic that the paper also starts off with (we're going to subtract infinity from itself, so there will be some issues with the rigor). Let the cost of the Poisson-weighted infinite tree T be the minimum cost of a matching on T (obviously every matching has infinite weight, but ignore that for now). Then we can write down a recursion based on what happens at the first level:

$$\text{cost}(T) = \min_{j\geq 1} \left(W_{\phi,j} + \sum_{i\neq j} \text{cost}(T^{(i)}) + \text{cost}(T^{(j)} \setminus j) \right).$$

Basically, we pay the weight of the edge ϕ, j connected to the root, plus the cost from subtrees $T^{(i)}$ at level 1 for all $i \neq j$, and then we can't use vertex j itself for $T^{(j)}$. If we compare this to the cost of the tree without the root vertex, we get a lot of cancellation because the cost of $(T \setminus \phi)$ is just the sum of the costs $T^{(i)}$:

$$\text{cost}(T) - \text{cost}(T \setminus \phi) = \min_{j\geq 1} \left(W_{\phi,j} - \left(\text{cost}(T^{(j)}) - \text{cost}(T^{(j)} \setminus j) \right) \right).$$

Defining $X_\phi = \text{cost}(T) - \text{cost}(T \setminus \phi)$ and $X_j = \text{cost}(T^{(j)}) - \text{cost}(T^{(j)} \setminus j)$, we get

$$\boxed{X_\phi = \min_{j\geq 1} (W_{\phi,j} - X_j)}.$$

But if we have a spatially invariant matching, X_ϕ and X_j should agree in distribution (because they are both defined to be the difference between the cost of a full tree and the tree without a particular vertex).

Lemma 230
Let $0 < \zeta_1 < \zeta_2 < \cdots$ be a Poisson point process (with $(\zeta_1, \zeta_2, \cdots) \overset{d}{=} (E_1, E_1 + E_2, \cdots)$), and let X, X_i be iid from some probability measure μ on $\mathbb{R}$. Then we have $X \overset{d}{=} \min_{i\geq 1} (\zeta_i - X_i)$ if and only if μ is the **logistic distribution** with density $f(x) = \frac{1}{(e^{x/2}+e^{-x/2})^2}$ and cdf $F(x) = \frac{e^x}{1+e^x}$.

This lemma is proved by calculus, and once we know the distribution function of μ, we can calculate useful properties directly:

Lemma 231

Let X, X_i be iid from the logistic distribution μ, and define $h(x) = \mathbb{P}(X_1 + X_2 \geq x)$ for all $x \geq 0$. Then

$$\int_0^\infty h(x) = 1, \quad \int_0^\infty xh(x)dx = \frac{\pi^2}{6}.$$

To understand how this relates to the matching problem, we can go back to the boxed equation above. We can guess that in the optimal matching for our tree, the weight of the edge adjacent to the root has distribution $W_{\phi,M^*(\phi)} \stackrel{d}{=} \zeta_{i^*}$, where $i^* = \text{argmin}(\zeta_i - X_i)_{i\geq 1}$. So we will calculate $\mathbb{E}[\zeta_{i^*}]$, and hopefully that's the number that we want – there's various ways to do this, but we'll follow what the paper does. Consider the process (ζ_i, X_i), which is a sequence of points scattered on the right half-plane – call the ζ-axis the z-axis and the X-axis the x-axis. Because the ξ_is form a Poisson point process, the number of points in any z-interval $[a, b]$ is a Poisson random variable with rate $b - a$. This means that within any **region** of the zx-plane, the number of points inside this region is also Poisson. Specifically, (ξ_i, X_i) is a Poisson point process on $(0, \infty) \times \mathbb{R}$ with intensity measure $\text{Leb} \otimes \mu$, where μ is **logistic** (that is, the X_is are iid logistic) and

$$\#\{\text{points in } R\} \sim \text{Pois}((\text{Leb} \otimes \mu)(R)).$$

for any region R. In particular, for any two disjoint regions R and S, behavior inside those regions is independent, so if we condition on the event $A_y = \{\text{some point } j \text{ of } (\zeta_i, X_i) \text{ lands on the vertical line } z = y\}$ (we will actually condition on finding a point where $z \in [y, y + dy]$), then the rest of the process is distributed as an independent Poisson process (and $z = y$ is disjoint from the other regions). This allows us to compute the expected value of $i^* = \text{argmin}(\zeta_i - X_i)$. We claim that

$$\mathbb{P}(i^* = j | A_y) = \mathbb{P}\left(y - X_j \leq \min_{i\geq 1}(\tilde{\zeta}_i - \tilde{X}_i)\right),$$

where the j is the same one as in the event A_y (which we know the value of because we're conditioning on A_y) and the $\tilde{\zeta}_i$s and $\tilde{X}_i$s are independently sampled from X_j. (In words, this is calculating the probability that if we find a point at horizontal coordinate y, it has the smallest value of $\xi_i - X_i$ among all points in the process.) To explain this equation, notice that $i^* = j$ means that $\zeta_j - X_j$ is the minimal value among all $\zeta_i - X_i$s, but we condition on $\zeta_j = y$ and then the remaining (ξ_i, X_i) pairs are all independent of X_j so we can reindex them as an independent Poisson point process. Write $\tilde{X} = \min_{i\geq 1}(\tilde{\zeta}_i - \tilde{X}_i)$; by Lemma 230, because the $\tilde{X}_i$s are logistic, so is $\tilde{X}$. Putting this all together and remembering that X and $\tilde{X}$ are independent, we find that

$$\mathbb{P}(i^* = j | A_y) = \mathbb{P}(y - X_j \leq \tilde{X}) = \mathbb{P}(X + \tilde{X} \geq y) = h(y)$$

by Lemma 231. So now we can calculate the unconditioned probability density function for ζ_{i^*} (which is what we're after) – to have $\zeta_{i^*} \in [y, y + dy]$, we must have the event $A_{[y,y+dy]}$ occur, so

$$\mathbb{P}(\zeta_{i^*} \in [y, y + dy]) = \mathbb{P}(A_{[y,y+dy]}) \cdot \mathbb{P}(i^* = j | A_{[y,y+dy]}) = h(y)dy$$

(because the probability of having a point in the Poisson point process in a horizontal strip of length dy is dy). In other words, $\zeta_{i^*} \stackrel{d}{=} W_{\phi,M^*(\phi)}$ has density h, so the expected value of $W_{\phi,M^*\phi}$ on our infinite tree is just the expected value $\int xh(x)dx = \frac{\pi^2}{6}$.

To conclude, we'll briefly discuss how to make the rest of the proof actually work. Our random variable i^* is related to the matching on our infinite tree T, but we still need to construct the actual matching $\mathcal{M}^*$. (Morally, we want to sample the infinite tree and then solve the boxed recursion above for the X_is, but there are infinitely many equations and variables). So instead, the idea is to only look at the edge-weights of T up to some finite level n and solve for X_is

within that, and we can avoid issues with dependence by using **directed edges** instead of undirected edges. Instead of the boxed recursion above, we now require the modified statement that

$$X_{v\to w} = \min_{u\neq v, w\to u}(W_{w\to u} - X_{w\to u}).$$

So we construct (T, X) (the infinite tree and the set of X_i random variables) as follows. First of all, we sample the edge-weights of T up to some finite-level n, and we let $W_{v,w}$ be the weight along the edge connecting v with w (in both directions). We can then determine the directed edge weights $X_{v\to w}$ and $X_{w\to v}$ within level n as follows. At all vertices of level n, sample iid logistic random variables pointing to their children. Then applying the recursion at each level n vertex allows us to find the directed edge weights from level $(n-1)$ to n, and repeatedly doing this allows us to determine $X_{v\to w}$ for all edges pointing away from the root. After that, we can also determine the edge weights pointing towards the root by continuing to apply the recursion relation – finally, we can apply the Kolmogorov extension theorem to get the infinite construction (T, X).

But having X in addition to T now allows us to construct the matching on T. Indeed, the edge between i and j is in the matching if

$$j = \text{argmin}_{k\sim i}(W_{i,k} - X_{k\to i}).$$

This gives us a unique neighbor for each vertex, and we can check that i's neighbor is chosen to be j if and only if j's neighbor is chosen to be i, because if j is chosen to be i's neighbor, then

$$W(i,j) - X_{i\to j} < \min_{k\sim i, k\neq j} W(i,k) - X_{i,k} = X_{j\to i}$$

by our modified recursion relation. In other words, an edge is included in the matching if the two X variables associated to it add to more than the weight W_{ij}, and we can check that this will not hold for any other vertex! This thus gives us a construction with the desired weights (and thus a spatially invariant matching with expected weight $\frac{\pi^2}{6}$ which we can plug into Theorem 229), and we can finish the proof by showing that any deviation from this matching cannot satisfy the recursion relation. (But we can read the paper for more details!)

References

[1] D. Aldous. Asymptotics in the random assignment problem. *Probab. Theory Relat. Fields*, 93(4):507–534, 1992.

[2] D. Aldous. The $\zeta(2)$ limit in the random assignment problem. *Random Struct. Alg.*, 18(4):381–418, 2001.

[3] R. Arratia and S. Tavaré. The cycle structure of random permutations. *Ann. Probab.*, 20(3):1567–1591, 1992.

[4] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, Cambridge, 2019.

[5] H. Kesten and B. Stigum. Limit theorems for decomposable multi-dimensional galton-watson processes. *J. Math. Anal. Appl.*, 17(2):309–338, 1967.

[6] S. Linusson and J. Wästlund. A proof of parisi's conjecture on the random assignment problem. *Probab. Theory Relat. Fields*, 128:419–440, 2004.

[7] R. Lyons, R. Pemantle, and Y. Peres. Conceptual proofs of l log l criteria. *Ann. Probab.*, 23(3):1125–1138, 1995.

[8] C. Nair, B. Prabhakar, and M. Sharma. Proofs of the parisi and coppersmith-sorkin random assignment conjectures. *Random Struct. Alg.*, 27(4):413–444, 2005.

[9] E. Shamir and J. Spencer. Sharp concentration of the chromatic number on random graphs $g_{n,p}$. *Combinatorica*, 7(1):121–129, 1987.

[10] N. Sun. 18.675. theory of probability. https://math.mit.edu/ nsun/f19-18675.html.

[11] J. Wästlund. An easy proof of the $\zeta(2)$ limit in the random assignment problem. *Electron. Commun. Probab.*, 14:261–269, 2009.