18.675 — Theory of Probability

Class by Konstantinos Kavvadias

Notes by Sanjana Das

Fall 2024

Lecture notes from the MIT class **18.675** (Theory of Probability), taught by Konstantinos Kavvadias. All errors are my own.

Contents

§1 September 5, 2024

[**TODO** — The first 20 minutes of this lecture are currently missing. They contain an introduction to measure theory and integration, the definition of a σ-algebra and measure, and an example where the ground set E is countable (where we take the σ-algebra to consist of all subsets of E, we have a weight function $m\colon E \to \mathbb{R}_{\geq 0}$, and we define the measure of a set $A \subseteq E$ as $\mu(A) = \sum_{x \in A} m(x)$.)]

§1.1 Generators for a σ-algebra

As you can see here, if E is countable, then we can define a notion of *size* on every single subset; so countable spaces are very nice, and appear in nature in a very natural way. But what if E is not countable anymore — for example, what if $E = \mathbb{R}$? This is much more complicated, and is not countable. In that case, σ-algebras can be more complicated. That's why we restrict our attention to simpler subsets of σ-algebras, which we call *generators* (which somehow encode all the properties of the σ-algebra).

Definition 1.1. Suppose we fix a set E, and $\mathcal{A}$ is a collection of subsets of E (i.e., $\mathcal{A} \subseteq \mathcal{P}(E)$). Then the σ-algebra generated by A, denoted $\sigma(\mathcal{A})$, is the set of all subsets of E which are contained in *every* σ-algebra of E that contains A.

In other words, we've considered *all* the σ-algebras we could define on E that contain $\mathcal{A}$, and we took their intersection; and we call that the σ-algebra generated by $\mathcal{A}$.

Why do we consider this kind of object? The point is this object might be useful because as mentioned earlier, if E is not countable anymore, the σ-algebra might be very complicated; but somehow all the information of the σ-algebra is encoded in the generator. So instead of studying the σ-algebra itself, we study the generator, which might be much simpler (e.g., the set of intervals in $\mathbb{R}$ generates a very nice σ-algebra).

Example 1.2

Consider $E = \mathbb{Z}$, and $\mathcal{A} = \{\{x\} \mid x \in E\}$ (so we consider all the singletons). Then $\sigma(\mathcal{A}) = \mathcal{P}(E)$.

The point is that every subset of E can be expressed as a countable union of elements of $\mathcal{A}$. So this is a simple example.

In the same spirit:

Example 1.3

Let $E = \mathbb{Z}$, and take $\mathcal{A} = \{\{x, x+1, x+2, x+3, \ldots\} \mid x \in \mathbb{Z}\}$. So in other words, we consider all subsets of the integers formed by letting x vary, and taking everything larger or equal to x. Then we have $\sigma(\mathcal{A}) = \mathcal{P}(E)$.

We can do this in any countable set; it turns out that all the information in countable spaces is encoded by the collection of singletons.

However, we might have much more complicated spaces, and then we need to be more careful when picking generators. Here the singletons were the simplest example; let's see a slightly more complicated example.

Example 1.4 (Borel σ-algebra)

Let $E = \mathbb{R}$, and consider $\mathcal{A} = \{U \subseteq \mathbb{R} \mid U \text{ open}\}$. Then $\sigma(\mathcal{A})$ is called the Borel σ-algebra.

The point is that here it's nontrivial this σ-algebra doesn't contain every single subset. It's known we can construct a subset of $\mathbb{R}$ which is not in the Borel σ-algebra.

An alternative way to describe this σ-algebra is that it's the one generated by intervals — i.e.,

$$\sigma(\{(a,b) \mid a < b \in \mathbb{R}\}).$$

Intervals are nice, and we can study them; based on intervals, we can say lots of things about the Borel σ-algebra. And we will also see a few more complicated sets, but our main focus will be $\mathbb{R}$ and the Borel σ-algebra. We are going to somehow define notions of measure, integration, and so on; and construct the Lebesgue measure, which is what assigns sizes to subsets of $\mathbb{R}$.

§1.2 π-systems and d-systems

We'll need a few more definitions, because they give us the language to say some interesting things about probability.

Another useful property of generators is we have some arbitrary σ-algebras, and we want to prove certain properties hold for them; and in many cases, it suffices to prove they hold for the generators, which is usually much simpler.

But sometimes it's not that easy to prove these properties hold for generators; so that's why we restrict ourselves to simpler collections of sets. One of those is the π system.

Definition 1.5. Let $\mathcal{A}$ be a collection of subsets of E. Then $\mathcal{A}$ is called a π-system if the following conditions hold:

(1) $\emptyset \in \mathcal{A}$.

(2) $\mathcal{A}$ is closed under finite intersections — i.e., if $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$.

So a π-system is a collection of sets containing $\emptyset$ closed under finite intersections. Most generating sets we'll see in this course are π-systems; so this is a nice property.

Another very nice collection of sets we'll consider is d-systems.

Definition 1.6. Let $\mathcal{A}$ be a collection of subsets of E. We say $\mathcal{A}$ is a d-system if the following conditions hold:

(1) $E \in \mathcal{A}$.

(2) If $A, B \in \mathcal{A}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{A}$. (So $\mathcal{A}$ is closed under differences of this form.)

(3) For any increasing sequence $(A_n) \subseteq \mathcal{A}$ (i.e., $A_1 \subseteq A_2 \subseteq \cdots$), we have $\bigcup_n A_n \in \mathcal{A}$.

So a d-system contains E, if you take any two sets with one contained int eh other, then their difference is there too; and it's closed underunions of increasing sequences.

What's the point for why we introduce these two notions of sets? The point is that a collection of sets is a σ-algebra if and only if it is both a d-system and a π-systsem — so we have broken into two pieces the definition of a σ-algebra. (We'll write this as a result; we're suggested to verify it on our own.)

Proposition 1.7

A collection $\mathcal{A}$ is a σ-algebra if and only if it is both a π-system and a d-system.

That's a nice way to somehow break into two pieces the definition of a σ-algebra.

We're being a bit bombarded with definitions right now, but these definitions are necessary because they give us the language we need.

§1.3 Dynkin's π-system lemma

Now that we have introduced certain notions, let's proceed to the proof of the first result of the course, which will be quite useful for various reasons.

Lemma 1.8 (Dynkin's π-system lemma)

Let $\mathcal{A}$ be a π-system. Then any d-system which contains $\mathcal{A}$ also contains the σ-algebra generated by $\mathcal{A}$ (i.e., $\sigma(\mathcal{A})$).

So here we start with any π-system $\mathcal{A}$, and then we consider a d-system which contains $\mathcal{A}$; then we're saying it also contains the σ-algebra generated by $\mathcal{A}$.

Before we prove this, why is it important? If we are not familiar with proofs at the moment, you don't have to actually remember the proof; the statement is actually how to learn to use the lemma that'll be important in the course.

How is this used? The point is we have some σ-algebra generated by $\mathcal{A}$, and we want to show this σ-algebra satisfies some nice properties. This lemma is a key input — we want to show that all elements of $\sigma(\mathcal{A})$ satisfy a certain property. For this, we first show the elements of $\mathcal{A}$ do, and that the things satisfying the property form a d-system; then this lemma says that that collection contains $\sigma(\mathcal{A})$.

Remark 1.9. When a set is uncountable, you can still consider the σ-algebra with all subsets, but then it's not easy to define a notion of size; you *can* take this σ-algebra, but it's not a good one. You have to restrict yourself. For example, in $\mathbb{R}$ there exist sets which are not Borel-measurable (we don't have time to go over it, but there's a construction due to Vitali).

The method of how we prove this is more or less standard to how people in measure theory prove things.

Proof of Dynkin's lemma. Let $\mathcal{D}$ be the intersection of all d-systems that contain $\mathcal{A}$. In other words, this is the smallest d-system which actually contains $\mathcal{A}$. Someone might ask whether there are actually d-systems containing $\mathcal{A}$; one such thing is the σ-algebra generated by $\mathcal{A}$.

The thing we want to show that $\mathcal{D}$ actually contains $\sigma(A)$; if we can show this, then we're done.

Now how do we do this? It suffices to show that this collection $\mathcal{D}$ (the intersection of all d-systems containing $\mathcal{A}$) is a π-system.

Why is this the case (i.e., why does this suffice)? It's easy to see intersections of d-systems are still d-systems. So if we manage to show that cD is a π-system, then it's a σ-algebra (by the above proposition). So it's a σ-algebra containing $\mathcal{A}$, which by the definition of $\sigma(\mathcal{A})$ means that $\mathcal{D}$ has to contain $\sigma(\mathcal{A})$; and then we're done. (Since $\mathcal{D}$ is the intersection of all d-systems containing $\mathcal{A}$, any such d-system has to contain $\sigma(\mathcal{A})$.)

The proof consists of two steps (in order to show this is a π-system).

Claim 1.10 — If $B \in \mathcal{D}$ and $A \in \mathcal{A}$, then $B \cap A \in \mathcal{D}$.

This is not exactly what we want to show — we want to show that for all $A, B \in \mathcal{D}$ we have $B \cap A \in \mathcal{D}$. But here we're first doing a simpler step where we only consider $A \in \mathcal{A}$. Then in the second step, we're going to prove that $\mathcal{D}$ is indeed a π-system.

Claim 1.11 — For any $B, A \in \mathcal{D}$, we have $B \cap A \in \mathcal{D}$.

If we manage to show this, then we're done, because that's the definition of a π-system. (Obviously $\emptyset$ is there because B is a d-system, and every d-system contains $\emptyset$ — if we take $A = B = E$, then $E \setminus E = \emptyset$. So d-systems contain $\emptyset$.)

First we'll prove the first claim. The standard way to prove this is — we have a property and want to show it's true for every element of $\mathcal{D}$. So we first consider

$$\mathcal{D}' = \{B \in \mathcal{D} \mid B \cap A \in \mathcal{D} \text{ for all } A \in \mathcal{A}\}$$

to be the set of all B satisfying the desired property. Clearly $\mathcal{D}' \subseteq \mathcal{D}$; the first claim is equivalent to showing that $\mathcal{D}' = \mathcal{D}$. So that's the main thing to show.

For this, let's make some observations. The first is that $\mathcal{A} \subseteq \mathcal{D}'$. Why? This is because $\mathcal{A}$ is contained in this family $\mathcal{D}$, and it's a π-system.

So then it suffices to show that $\mathcal{D}'$ is a d-system (becasue $\mathcal{D}$ is the smallest d-system containing $\mathcal{A}$, so if $\mathcal{D}'$ is a d-system containing $\mathcal{A}$, then it has to contain $\mathcal{D}$; and since of course $\mathcal{D}' \subseteq \mathcal{D}$ as well, this means they have to be equal).

Here what we want to do is prove a property is ture for every element in $\mathcal{D}$; so we consider the collection of all the elements of $\mathcal{D}$ for which this property is satisfied, and we want to show this is indeed the collection $\mathcal{D}$.

We know that $\mathcal{A} \subseteq \mathcal{D}'$. So if we show that $\mathcal{D}'$ is a d-system, then it's a d-system containing $\mathcal{A}$; and since $\mathcal{D}$ is the smallest, this means it has to contain $\mathcal{D}$. But $\mathcal{D}'$ is a subset of $\mathcal{D}$, so this means they have to be equal. So that's the reasoning in proving the first claim.

Now let's see why $\mathcal{D}'$ is really a d-system. The first thing is that clearly $E \in \mathcal{D}'$, because if you take any element of A, then $A \cap E = A \in \mathcal{A} \subseteq \mathcal{D}$. So the first condition in the definition of a d-system holds.

Now let's see why the second condition is true, for $\mathcal{D}'$. Here, suppose we take any two sets $B_1, B_2 \in \mathcal{D}'$ such that $B_1 \subseteq B_2$. Then for every set $A \in \mathcal{A}$, we have that $(B_2 \setminus B_1) \cap A = (B_2 \cap A) \setminus (B_1 \cap A)$ (this is easy to see, from set theory; you can also verify by yoursef that we have this equality). So by the definition of $\mathcal{D}'$, we have that that $B_1 \cap A$ and $B_2 \cap A \in \mathcal{D}$, so the fact that we have a d-system means that $(B_1 \cap A) \setminus (B_2 \setminus A) = \emptyset$. (by the definition of d-systems), so by the second property of d-systems, this is true for all shifts, so $B_1 \setminus B_2 \subset |B_1|$.

The last thing we need to shwo is that $\mathcal{D}'$ is closed under increasing sequences. Here, suppose that $(B_n)_{n \geq 0}$ is an increasing sequence in $\mathcal{D}'$. Then for every set $A \in \mathcal{A}$, we have

$$(\bigcup_i B_i) capA = \bigcup_{n \geq 1} (B_i \cap A)$$

(by set theory). And all these sets $B_i \cap D$ belong to d (by the way we defined $\mathcal{D}'$). S owe have an increasing sequence of sets belonging to $\mathcal{D}$. And since $\mathcal{D}$ is a d-system, this means this set has to be in the union as well. So this means $\mathcal{D}'$ is a d-system containing $\mathcal{A}$, as desired.

So we have shown the first claim is true for every element of $\mathcal{D}$; that's the first step.

(It's quite standard that we want to prove a property is true for every element in a collection fo sets, and we prove the collection satisfies some nice properties — like being a d-system or π-system.)

Now we need to show the second claim. Again, the reason is exactly the same — we consider the collection of all $B \in \mathcal{D}$ such that $B \cap A \in \mathcal{D}$ for all $A \subseteq \mathcal{D}$. We'll show again that this is a d-system containing $\mathcal{A}$, so it has to contain all the d-systems cnotaining $\mathcal{A}$.

We'll briefly sketch this (but not give the whole proof). For this, we let $\mathcal{D}''$ be the collection fo sets

$$\mathcal{D}'' = \{B \in \mathcal{D} \mid B \cap A \in \mathcal{D} \text{ for all } A \in \mathcal{D}\}.$$

So our goal is to prove that $\mathcal{D}'' = \mathcal{D}$. Since we have already shown that $\mathcal{D}' = \mathcal{D}$, we have that $\mathcal{A} \subseteq \mathcal{D}''$; similarlly $\mathcal{D}''$ is a d-system. (This is exactly the same arguemnt as before.) That completes the proof — this shows $\mathcal{D}''$ is a d-system that cihtaains a, so it has to also contain $\mathcal{D}$.

So condition (2) holds; this means $\mathcal{D}$ is a π-system, therefore a sigma-algebra; and we're done. $\square$

We chose to prove this in a ddetailed way so that we could get familiar with some of the notions and concepts of measure theory. But we're mostly going to use the statement of the lemma, because it will be useful to prove things, not the proof itself. But it would be nice to be familiar with he proof.

Now we've mentioned several definitions; at a first glance they're not easy to absorb, but in time we'll hopefully start getting familiar with them.

Let's also mention a few more definitions, which are again goig to be useful later; we'll end this lecture with those definitions, and also give time to discuss questions.

If you haven't taken many pure math courses, you've dealt with functions defined on $\mathbb{R}$ (e.g., we have an interval $[0,1]$, and then we consider functions defined on that interval). But w ecan define functions on sets as well — we have a collection fo sets, and we can define a function ont hat collection of sets. That's what set functions are, and we'll deal with many.

Definition 1.12 (Set function)**.** Suppose that $\mathcal{A}$ is a collection of subsets of E containing $\emptyset$. Then a set function is a function $\mu\colon \mathcal{A} \to [0,\infty]$ (we allow values of ∞) such that $\mu(\emptyset) = \emptyset$.

We can imagine μ as a way of assigning sizes to the elements of this family. But there are other types of functions as well.

Recall that in $\mathbb{R}$, we say a function is *increasing* if for any two numbers $x < y$, we have $f(x) \le f(y)$. This is possible because in $\mathbb{R}$ there is a natural order of numbers; we can say whether one is greater than the other. The point now is we can do something smilar with sets; we think of $A \le B$ if $A \subseteq B$.

Definition 1.13. We say a set function is increasing if it has the property that $\mu(A) \le \mu(B)$ for all $A, B \in \mathcal{A}$ such that $A \subseteq B$.

Definition 1.14. A set function is called additive if whenever $A, B \in \mathcal{A}$ are such that $A \cup B \in \mathcal{A}$ and $A \cap B = \emptyset$, we have $\mu(A \cup B) = \mu(A) + \mu(B)$.

This is very natural; it's realted to what we said earlier where if you want to assign sizes to two different collection sof sets, you should jus ttake the su m of their sizes.

But we can also do the same with countably many sets instead of just finitely many, and in that case the function is called *countably* additive.

Definition 1.15. We say a set function μ is countably additive if whenever we have a sequence $(A_n) \subseteq \mathcal{A}$ such that $\bigcup_n A_n \in \mathcal{A}$ and the A_n's are pairwise disjoint (i.e., $A_n \cap A_m = \emptyset$ for all $n \neq m$), then

$$\mu\left(\bigcup A_i\right) = \sum \bigcup A_i.$$

So now instead of having a finite collection of objects, we have an infinite one; adn we stil say tehri sizes is just the sum of hte sies.

One last thing, before we complete today's lecture, is what we mea by saying *countably subadditive functions*. We're going to define these things now, and refer to them from now on; it's sometimes nice to define objects snad then based on these definitions, start to rigorously construct theories.

Definition 1.16. A set function is countably subadditive if whenever we have a countable collection fo sets $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}$ with $\bigcup_n A_n \in \mathcal{A}$, we have

$$\mu\left(\bigcup A_n\right) \le \sum \mu(A_n).$$

In this definition, we do not assume the A_n's are pairwise disjoint, and we have an inequality; this makes sense, because if we took two sets which are not disjoint, we cannot have an equality because we cannot count an element twice; so that's why we need na inequality (when we consider the size of the union, we don't consider elements twice). If the sets are pairwise disjoint, then we could have equality. (Here we're giving some intuition if you have a measure; a measure is both countably subadditive and countably additive, and when these sets are disjoint you have equality.)

Next lecture we'll mention Caratheodory's extension theorem; we'll see how we can construct more complicated measures than the simple ones we've mentioned so far.

After each lecture, we will upload a summary of what we did so you can have a guide if you miss a lecture. We've also uploaded some lecture notes online. These are quite compact but contain the main theorems; on the board we'll try to clarify some things and expand the proofs.

§2 September 10, 2024

The first pset will be uploaded by the end of this week. There will be 6 problem sets; most will be problems of understanding, to help us understand the material. There'll be one every two weeks; we'll have around 4–5 problems. There will be office hours every Friday, 5:30–7:30.

§2.1 Overview

Last time we saw an introduction to the course, with some motivation for measure theory and probabiltiy. Konstantinos also gave us several definitions, which might at first glance seem abstract nonsense, but are important because they give us the vocabulary we need.

We also mentioned what a measure is — a countably additive set function — but we haven't actually constructed a measure. We gave some elementary examples of measures, but we didn't give a more general way to construct measures. This is what we'll do today; we'll state Caratheodory's extension theorem, which lets us construct measures, and then hopefully we can construct the Lebesgue measure on $\mathbb{R}$.

§2.2 Two more definitions

Before we do this, we'll need two more definitions. As usual, we fix a (very large) set E, and we consider a collection of subsets $\mathcal{A}$ of E.

Definition 2.1 (Ring)**.** A collection of subsets $\mathcal{A}$ is called a ring on E if:

- $\emptyset \in A$.
- For all $A, B \in \mathcal{A}$, we have $B \setminus A \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.

So a ring is a collection of subsets which contains $\emptyset$ and is closed under finite unions and under differences (in the sense that if we take any two elements in the family and take their difference, the result is still in the family). We'll need this to state Caratheodory's theorem in its full generality; this is all we need it for.

Definition 2.2 (Algebra)**.** A collection of subsets $\mathcal{A}$ is called a algebra on E if:

- $\emptyset \in \mathcal{A}$.
- For every $A, B \in \mathcal{A}$, we have $A^c \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$ (where $A^c = E \setminus A$).

This is more or less the same thing as a σ-algebra, but it's not necessarily closed under *countable* unions, just finite ones. We now have the basic definitions we're going to use; from now on we're going to use these definitions extensively to do more interesting things.

§2.3 Caratheodory's extension theorem

Caratheodory's extension theorem is a general way of constructing measures; we'll only briefly sketch the proof, but we'll use it. Caratheodory was a mathematician who made major contributions in measure theory and analysis at the beginning of last century.

Theorem 2.3 (Caratheodory's extension theorem)

Let $\mathcal{A}$ be a ring on E, and suppose μ is a countably additive set function on $\mathcal{A}$. Then the set function μ extends to a measure on $\sigma(\mathcal{A})$ (the σ-algebra generated by $\mathcal{A}$).

We will see there's an algorithm for how to extend the function to the σ-algebra.

This is a very powerful theorem. First, it's very general — we have a set E and a ring on that set (a collection of sets satisfying the earlier properties), and we know we can define a countably additive function on that set. And this theorem says this extends to a measure. So to construct a measure, we just have to define it on a simpler family of sets, where it's much easier to define (it's usually much easier to define it on a ring than on the entire σ-algebra). The construction of the Lebesgue measure is based on this.

Proof. We'll talk about how to construct the measure, but we won't prove it satisfies the desired properties (due to the lack of time).

First, for *any* $B \subseteq E$, we set

$$\mu^*(B) = \inf\left\{\sum_n \mu(A_n) \mid (A_n) \subseteq \mathcal{A} \text{ and } B \subseteq \bigcup_n A_n\right\}.$$

We call μ^* an outer measure (it's some sort of generalized notion of a measure). What this means is we fix B, and consider any sequence (A_n) in our ring that covers B, and take the sum of the values in that sequence; and μ^* is intuitively just the smallest of those sums. (Of course there doesn't necessarily need to be a sequence (A_n) attaining equality, but that's the intuition.) If no such sequence (A_n) exists, then we set $\mu^*(B) = \infty$. (It might not be possible to cover B by sets in $\mathcal{A}$; in that case we set its outer measure to ∞. In general, we define $\inf \emptyset = \infty$.)

Remark 2.4. If such a sequence does exist but for all of them the sum diverges, you still set $\mu^*(B) = \infty$.

Student Question. *Do we take A_n to be disjoint?*

Answer. No; (A_n) is any sequence in our ring which covers B, and we sum the measures of those sets.

We've defined a function on the entire space, but this is not enough, because we need to define a notion of nice sets — unfortunately not every set is nice.

First, it's clear that $\mu^*(\emptyset) = \emptyset$, because we can take $A_n = \emptyset$ for all n. And μ^* is an increasing function; this is not hard to see (it's just from the definition).

Now, what are the nice sets? We say that a set $A \subseteq E$ is μ^*-*measurable* if

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

for all $B \subseteq E$. In other words, this means A nicely divides any subset of our space. Why does this make sense? If μ^* were a measure, then this condition would hold by the countable additivity of a measure; so it makes sense to define nice sets in this way.

Then we define $\mathcal{U}$ as the collection of μ^*-measurable sets. And that's our collection of nice sets.

There are two observations, which we're not going to prove (we don't have the time to go through all the details):

(1) The collection $\mathcal{U}$ is a σ-algebra containing $\mathcal{A}$ (in particular, this means it contains $\sigma(\mathcal{A})$).

(2) The function μ^* is a measure on the σ-algebra $\mathcal{U}$, and its restriction to $\mathcal{A}$ is μ.

So that's exactly what our extension is — μ^* is defined on a larger σ-algebra, but the fact that $\mathcal{A} \subseteq \mathcal{U}$ means that $\sigma(\mathcal{A}) \subseteq \mathcal{U}$. And Caratheodory proved that μ^* is a measure on this bigger σ-algebra $\mathcal{U}$. (And we've got an explicit form of the extension.) □

We don't have time, but this gives an idea of how we construct a measure starting with an arbitrary set function on a ring. This is used extensively — it's a very nice way to construct measures.

Student Question. *If μ^* is a measure on $\mathcal{U}$, it follows that it's also a measure on $\sigma(\mathcal{A})$?*

Answer. Yes — if it's countably additive on $\mathcal{U}$, then it's also countably additive on $\sigma(\mathcal{A})$, for example. In general, if you restrict a measure to a smaller σ-algebra, then it's still a measure there.

Remark 2.5. If the ring $\mathcal{A}$ is a σ-algebra, then μ^* takes a simpler form — we have

$$\mu^*(B) = \inf\{\mu(A) \mid A \in \mathcal{A},\, B \subseteq A\}.$$

This is because in our definition $\mu^*(B) = \sum \mu(A_n)$, the union $\bigcup A_n$ will also be in $\mathcal{A}$, and μ is countably additive; so in that case we get this simpler form.

§2.4 Uniqueness of extensions

A very natural question someone could ask: Caratheodory's theorem says you can construct a measure, but what about *uniqueness*? This is an extension of μ, but can there be another? In general, uniqueness is not true. But if E has finite measure, then in that case, the extension *is* unique; we're going to prove that in the next theorem. This uses the machinery we've developed so far — in particular, Dynkin's lemma from last class.

Theorem 2.6

Suppose that we have two measures μ_1 and μ_2 om $(E, \mathcal{E})$ (where $\mathcal{E}$ is some σ-algebra on E) such that $\mu_1(E) = \mu_2(E)$ and both are finite. Then if $\mathcal{A}$ is a π-system which generates $\mathcal{E}$ (i.e., $\sigma(\mathcal{A}) = \mathcal{E}$), and μ_1 and μ_2 agree on $\mathcal{A}$, then $\mu_1 = \mu_2$.

What this says is that if we start with any two measures which give the same mass of the space and agree on a π-system generating the σ-algebra of the space, then they have to be the same; this gives a nice criterion for Caratheodory's extension theorem where the existence is unique (if we take $\mathcal{A}$ to be a π-system where the function you start with has finite mass, then the extension is unique).

Proof. The proof is very similar to what we did last time. Here's what we need to show — we consider

$$\mathcal{D} = \{A \in \mathcal{E} \mid \mu_1(A) = \mu_2(A)\}$$

(i.e., the collection of sets on which the two measures agree). Our goal is to show that $\mathcal{D}$ is the entire σ-algebra. To do this, last time we showed that if you have a d-system containing $\mathcal{A}$, then this will also contain $\sigma(\mathcal{A})$. So what we actually have to prove, by Dynkin's lemma, is that $\mathcal{D}$ is a d-system; if we can prove this, then we're done by Dynkin's lemma (and this is why Dynkin's lemma is useful).

First, we know that $\mathcal{D}$ indeed contains $\mathcal{A}$ (this is by our assumption that μ_1 and μ_2 agree on $\mathcal{A}$). So by Dynkin's lemma, it suffices to show that $\mathcal{D}$ is indeed a d-system.

To show this, recall that there are three conditions that we need to actually verify. The first condition is that $E \in \mathcal{D}$. But this is obvious, by our assumption that the two measures give the same mass on the entire space.

The second thing we have to show is that if we take any $A, B \in \mathcal{D}$ with $A \subseteq B$, then $B \setminus A \in \mathcal{D}$ (this is the second condition for being a d-system). That's quite easy in our case, for the following reason. By our assumptions, we have that $\mu_1(B) = \mu_1(A) + \mu_1(B \setminus A)$ (this is because μ_1 is countably additive, because it's a measure); the right-hand side is something finite. Similarly we have

$$\mu_2(B) = \mu_2(B) + \mu_2(B \setminus A) < \infty.$$

And these two quantities are equal, because by definition they belong to $\mathcal{E}$. Meanwhile, $\mu_1(A) = \mu_2(A)$. So when we subtract the first equation from the second, we get that

$$\mu_1(B \setminus A) = \mu_2(B \setminus A),$$

which means that $B \setminus A \in \mathcal{D}$.

However, here we have hidden something — here you're subtracting two numbers, so we need to know those numbesr are finite. That's hwere we're using that $\mu_i(E)$ is finite; we'll see counterexamples otherwise.

So the second condition of a d-system is satisfied. The last condition is as follows: suppose we take an increasing sequence $(A_n) \subseteq \mathcal{D}$ such that $A = \bigcup A_n \in \mathcal{D}$. We then need tos how that A itself is also in $\mathcal{D}$.

But this follows from a limit argument — we have

$$\mu_1(A) = \lim_{n \to \infty} \mu_1(A_n)$$

(we'll prove this on the first problem set — that when we have a limit of increasing sets, the measure of the limit is the limit of their measures). But this is $\lim_{n\to\infty} \mu_2(A_n)$, which is $\mu_2(A)$. So we indeed have $A \in \mathcal{D}$; this implies $\mathcal{D}$ is a d-system containing $\mathcal{A}$.

And $\mathcal{A}$ is a π-system, so the fact that $\mathcal{D}$ contains $\mathcal{A}$ measn it contains a π-ystem generated by $\mathcal{A}$. □

So under special conditions, we do have uniqueness of extension; we are going to use this for the Lebesgue measure.

Student Question. *Is there motivation for the definition of a d-system?*

Answer. The notion of d-systems and π-systems somehow break into two pieces the notion of a σ-algebra, which make them easier to understand. In measure theory we study complicated objects by restricting to something easier to understand, and then building from there. The notion of a d-system is simpler than a σ-algebra; so we design a d-system and π-system, and combine them to get a standard σ-algebra. At a first glance this may seem like abstract nonsense, but it turns out to be very important.

Student Question. *Are we assuming the two measures are increasing?*

Answer. Any measure is necessarily increasing — if you have $A, B \in \mathcal{E}$ where $A \in B$, then $B \setminus A \in \mathcal{E}$ (because $B \setminus A = B \cap A^c$), So this means $\mu(B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A)$. And measures are always nonnegative, so this is always at least $\mu(A)$. So measures are automatically increasing.

Student Question. *Are the definitions of rings and algebras related to the ones from algebra?*

Answer. Yes, if you consider the set operations as operations — in algebra, you have sets closed under standard operations (e.g., addition). And here we have something similar, where the algebra is closed under unions. You can understand addition as union (very vaguely speaking); if you make this abstract connection, then you have something analogous. (This is probably the reason for the name.)

§2.5 An example of non-uniqueness

Now that we've seen uniqueness, we'll construct an example where uniqueness breaks down (where we'll take a set of infinite measure). Let $E = \mathbb{Z}$, and consider the σ-algebra $\mathcal{E} = \mathcal{P}(E)$ (the set of all subsets of E).

Suppose we have a π-system mentioned last time, specifically

$$\mathcal{A} = \{\{x, x+1, x+2, \ldots\} \mid x \in E\} \cup \{\emptyset\}.$$

As mentioned last time, this is a π-system which generates the entire σ-algebra (this is not hard to see — every subset of $\mathbb{R}$ ca be written using such sets).

Then we need to define our measures; we define $\mu_1(\mathcal{A}) = |\mathcal{A}|$ (so this tracks the number of elements; note that $\mathcal{A}$ might be infinite, in which case $\mu_1(\mathcal{A}) = \infty$).

And we simply take $\mu_2(\mathcal{A}) = 2\mu_1(\mathcal{A})$, for every $A \in \mathcal{P}(E)$.

So we have two very elementary measures; uyou can see that these two measures are off by a factor of 2. But they conicide on $\mathcal{A}$, because every element of $\mathcal{A}$ has infintie measure, and $2\infty = \infty$. So μ_1 and μ_2 coincide on our π-system.

So in that case, the previous theorem breaks down because these two theorems give *infinite* mass in our space, and our argument crucially uses finite measures.

§2.6 Borel measures

We'll now give a few definitions that will be useful, and hopefully we'll have time to go through the construction of Lebesgue's measure.

Definition 2.7 (Borel σ-algebra)**.** Let E be a topological space (a set with some topology). Then we define the Borel σ-algebra in E t obe the σ-algebra generated by the open subsets of E. ()

(The notion of topology is not important if you're not familiar with it; you can imagine E is just $\mathbb{R}$¿)

So we consider all the open sets, and the σ-algebra generated by those sets. In our case, we'll mostly b interestedd in the case $E = \mathbb{R}$ (with the usual topology).

We define the Borel σ-algebra as $\mathcal{B}(E)$.

Definition 2.8. A measure μ on the topological space $(E, \mathcal{B}(E))$ is called a Borel measure. Furthermore (in addition), if the measure of any finite set $E \in \mathcal{E}$ is finite, then we say μ is a Radon measure.

If we don't know what a compact set is in an arbitrary space, that's fine; what we actually have to know is we can imagine replacing K by intervals in $\mathbb{R}$. So if our measure gives finite mass to every closed interval, then it's a Radon measure.

§2.7 The Lebesgue measure

Now the main theorem, which we'll try to finish the proof of today, is the following.

Theorem 2.9

There exists a unique Borel measure μ on $\mathbb{R}$ with the property that $\mu([a,b]) = b - a$ for all $b \geq a$.

Intuitively, we can understand this — we take any interval and set its measure to its length. But the point is that this measure can be extended to a much larger σ-algebra, and this extension is actually unique.

We'll now go through the proof.

§2.7.1 Proof of uniqueness

As we might expect, for uniqueness we're going to use the previous thoerem we stated. But there's one problem — our measure μ is not finite (it's defined on the entire rela line, and it's not necessarily finite, because if you take the interval with length $[0,n]$, then its length is n). So we need a trick to apply our previous uniqueness theorem.

The trick is as follows: for every $n \in \mathbb{Z}$, we *restrict* our measures to

$$\mu_n(A) = \mu(A \cap (n, n+1]),$$

and we do the same for a second measure $\widetilde{\mu}_n(A)$ (we assume for contradiction that we have two such measures, and we want to eventually prove they're the same) — so we define

$$\widetilde{\mu}_n(A) = \widetilde{\mu}(A \cup (n, n+1]).$$

Then we need a π-system — we consider all half-open intervals $(a,b]$, where $a < b \leq \mathbb{R}$. By the definitions of our measures and the property we have, μ_n and $\widetilde{\mu}$ agree on our π-system — so $\mathcal{A}$ is a π-system with the properties that it generates the entire algebra, meaning $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$. And then by the previous lemma, we have $\mu_n = \widetilde{\mu}_n$.

Now uniquness is easy — for every $A \in \mathcal{B}(\mathbb{R})$, we have $A = \bigcup_m (A \cap (n, n+1])$. And these sets are pairwise disjoint, so by countable additivity we have $\mu(A) = \{\sum n(A \cap (n, n+1)]\}$. But then we have

$$\mu(A) = \sum_n \mu(A \cap (n, n+1]) = \sum_n \widetilde{\mu}(A \cap (n, n+1]) = \widetilde{\mu}(A).$$

So the idea was to partition into finite intervals that cover the space, and then apply the previous theorem (we were dealing with an infinite measure we couldn't use the previous theorem directly, but dealing with this wasn't too hard).

§2.7.2 Proof of existence

The harder part is the proof of existence; for that we'll use Caratheodory's theorem, but for this we need a ring that generates the entire Borel σ-algebra, as well as a function restricted to that ring. So we need to be a bit careful with this; and that's the main goal for the rest of the lecture (we want to apply Caratheodory, but we have to do something first).

We need to consider a ring which generates the entire Borel σ-algebra. For that, we consider $\mathcal{A}$ to be the set of finite unions of disjoint intervals $(a_i, b_i]$ — i.e.,

$$\mathcal{A} = \{A = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n]\}.$$

(The reason for the half-closed intervals is so that it's not hard to show this is a ring.)

Then $\mathcal{A}$ is a ring with the property that $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R})$; this is nontrivial, but not super hard to show. (It may be one of the problems on the problem set, so that we can play around with definitions.)

Then we should define $\mu(\mathcal{A})$ as the sum of lengths of those intervals — i.e.,

$$\mu(\mathcal{A}) = \sum_{i=1}^{n} (b_i - a_i).$$

Of course, here someone might say, if we express A in a different way, do we still get the same length? The answer is yes, because we require the intervals be disjoint; so μ is well-defined, and additive. (Additivity is again not very hard to see, meaning that the union of two disjoint sets is also the sum of their measures; it's a nice exercise.)

But we need to actually show that μ is *countably* additive. For this, let $(A_n) \subseteq \mathcal{A}$ be a sequence of disjoint sets, and let $A = \bigcup A_n$; we assume that $A \in \mathcal{A}$. (This is not true for every such sequence, but we only restrict to such unions, because we need to just check that μ is countably additive *on* our ring; so we only care about countable unions that don't leave the ring.)

We need to show countable additivity, meaning that

$$\mu(A) = \sum_{n} \mu(A_n).$$

(That's the definition of countable additivity, and it's one of the prerequisites of Caratheodory's extension theorem.) First, by *finite* additivity (as we discussed earlier), we can break $\mu(A)$ into two pieces as

$$\mu(A) = \mu(A_1) + \mu(A \setminus A_1).$$

Then we can further break the second part and write

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

And we can iterate this; then we get

$$\mu(A) = \sum_{i=1}^{n} \mu(A_i) + \mu(A \setminus (A_1 \cup \cdots \cup A_n)).$$

To complete the proof of countable additivity, we need to show that the last term tends to 0, i.e., that

$$\mu(A \setminus (A_1 \cup \cdots \cup A_n)) \to 0$$

as $n \to \infty$ (then we will have shown μ is countably additive on our ring, and then we are done by Caratheodory's extension theorem).

Now we're going to use some notation to make our lives easier — set $B_n = A \setminus (A_1 \cup \cdots \cup A_n)$. What we want to show is that $\mu(B_n) \to 0$ as $n \to \infty$ (this is exactly the same thing as before, we've just changed notation). We are going to argue by contradiction — we are going to assume that $\mu(B_n)$ does *not* converge to 0, and then we're going to reach a conclusion that contradicts a conclusion we had before.

So suppose for contradiction that $\mu(B_n)$ does not tend to 0 as $n \to \infty$. What does it mean not to converge to 0? Here, since the B_n's are also decreasing, this implies there exists $\varepsilon > 0$ such that $\mu(B_n) \geq 2\varepsilon$ for every n. To explain this a bit more, a sequence converges to 0 if for every ε, eventually the sequence goes below ε. So for a sequence to *not* converge to 0 means there exists ε such that the sequence stays above ε for infinitely many values of n. (And here we can go from infinitely many to all using the fact that the sequence is decreasing.)

Now we need to somehow approximate elements of B_n's by elements of our ring $\mathcal{A}$ — B_n is not necessarily an element of $\mathcal{A}$ because it can't be written as a finite union of disjoint intervals. But it *can* be written as an *infinite* union of intervals, which means elements of B_n can be approximated by elements of our ring. More specifically, for every n we can find $C_n \in \mathcal{A}$ such that $\mathrm{Cl}(C_n) \subseteq B_n$ and $\mu(B_n \setminus C_n)$ is small — specifically, $\mu(B_n \setminus C_n) \leq \varepsilon/2^n$. This is possible because each B_n is a union of finite intervals, and any such thing can be approximated by unions of finitely many disjoint intervals.

So then we have

$$\mu(B_n) - \mu(C_1 \cap \cdots \cap C_n) = \mu(B_n \setminus (C_1 \cap \cdots \cap C_n))$$

(because this intersection is contained in B_n by construction). But we can write

$$\mu(B_n \setminus (C_1 \cap \cdots \cap C_n)) \leq \bigcup_{m=1}^{n} (B_m \setminus C_m).$$

And now by countable additivity, this is at most

$$\sum_{m=1}^{n} \mu(B_m \setminus C_m).$$

And we chose C_m such that this is very small — specifically, at most $\sum_{m=1}^{n} \varepsilon/2^m \leq \varepsilon$ (because $\sum 1/2^m = 1$ — that's why we chose this sequence of error terms, because it's summable).

Here what we've done is approximated our sets B_n by nice sets C_n, and then approximated $\mu(B_n)$ in terms of the measures of these nice sets C_n.

But we also have $\mu(B_n) \geq 2\varepsilon$, so this means $\mu(\bigcap C_m) \geq \varepsilon$ (by the above inequality). And this is true for all n.

Now consider $K_n = \bigcap_{m=1}^{n} \mathrm{Cl}(C_m)$. The point is here that this sequence K_n is a sequence of decreasing compact sets, which have the property that $\mu(K_n) \geq \varepsilon$. In particular, the K_n are nonempty for all n. And we can use the 'finite intersection property' of $\mathbb{R}$, which says that if we have a decreasing sequence of compact sets in $\mathbb{R}$ which are nonempty, then their intersection is also nonempty. (We may not be familiar with this and that's fine, but we need it for this proof.)

Then the finite intersection property implies that $\bigcap_n K_n$ is nonempty (because if we had a set with positive measure, then it cannot be the empty set; and we've constructed a decreasing sequence of compact sets which have positive measure, so must be nonempty; and then this property ensures their intersection is nonempty).

But by construction, $\bigcap_n K_n \subseteq \bigcap_n B_n$. And this intersection is the empty set, since $A = \bigcup_n A_n$ and so this is $A \setminus A = \emptyset$.

And so that's how we obtain the contradiction — by constructing this decreasing sequence of compact sets. The whole idea of the proof was to get a contradiction in this way. Some of us might not be familiar with these techniques, which is fine (we don't have to come up with these proofs, some very smart people did); but this is an idea of how these things sometimes work, and it's very nice and elegant.

And now we use Caratheodory's extension theorem — the function we defined on the ring satisfies the assumptions of this theorem, so it extends (in the way described earlier).

Since we've also proved uniqueness, this tells us there is a *unique* measure where the measure of an interval is the size of that interval. We define this measure as the Lebesgue measure.

> **Definition 2.10.** The Lebesgue measure is the unique Borel measure μ on $\mathbb{R}$ with the property that $\mu([a, b]) = b - a$ for all intervals $[a, b]$.

And that's how we construct the Lebesgue measure. This procedure generalizes to higher dimensions too, though we're not going to do that (you can use the same method, but instead of intervals you use squares or cubes or so on).

Student Question. *Are the Lebesgue-measurable sets precisely the Borel ones?*

Answer. No — there are actually sets which are Lebesgue-measurable but not Borel. In our setup, what we called Lebesgue-measurable sets correspond to the σ-algebra $\mathcal{U}$ from Caratheodory's extension theorem, while what we call Borel-measurable sets are $\sigma(\mathcal{A})$. These are not the same — you can construct sets which are Lebesgue-measurable but not Borel-measurable. We won't do this because there's not enough time.

§3 September 12, 2024

The first problem set is uploaded, and the deadline is September 27. It has 7 problems, but you only have to submit the first 6 for credit; Prof. K added a bonus problem for those of us who are interested in solving more problems. He designed the problems to be problems of understanding; they shouldn't require deep mathematical background for the moment. If you feel you have mathematical gaps or want to ask anything, you can come to the office hours; these will be Friday 5:30–7:30 at the math department (this is also announced on Canvas). Now we also know our TA and grader; you can find their information on the syllabus.

§3.1 The Lebesgue measure

Last time, we rigorously constructed the Lebesgue measure on the real line (or the Borel σ-algebra); we also proved it's the unique Borel measure with the property that the measure of any interval was its length. The proof was a bit technical and required some nontrivial steps.

§3.2 σ-finiteness

One observation about the Lebesgue measure is it's not finite — the mass it gives to $\mathbb{R}$ is infinite. However, it is *σ-finite*.

Definition 3.1 (σ-finite)**.** Let $(E, \mathcal{E})$ be a measure space (so that E is a set together with a σ-algebra $\mathcal{E}$), and μ a measure on this space. Then we say μ is σ-finite if there exists a sequence $(E_n) \subseteq \mathcal{E}$ of measurable sets with the property that $E = \bigcup_{n\in\mathbb{N}} E_n$ and $\mu(E_n) < \infty$ for all n.

Here the space might have infinite measure, but in the case of σ-finite measures, we can partition the space into smaller pieces such that each piece has finite mass under our measure. Of course the Lebesgue measure is σ-finite — you can take $E_n = [-n, n]$, for example (for $n \in \mathbb{N}$). Then $\bigcup_n E_n = \mathbb{R}$, but $\mu(E_n) = 2n$, which is a finite number (for each n).

Why do we use this definition? It's important because most properties that are true for *finite* measures are also true for σ-finite ones, because of this nice partition. So it'll often be useful to study σ-finite measures.

§3.3 Translational invariance

Now that we have this definition, let's prove another property of the Lebesgue measure. Intuitively, if we have an interval with endpoints a and b, and then we translate that interval by x — so we take the interval with endpoints $a + x$ and $b + x$ — then these intervals have the same length. And the point is that for the

Lebesgue measure, this is actually true for *any* Borel subset A — i.e., $\mu(A+x) = \mu(A)$. We'll now prove this; it's intuitively clear, but proving it rigorously is nontrivial.

Proposition 3.2

The Lebesgue measure is *translation invariant*, meaning that $\mu(A) = \mu(A+x)$ for every set $A \in \mathcal{B}(\mathbb{R})$ and for every $x \in \mathbb{R}$.

Recall that $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra (defined last lecture); and $A + x = \{a + x \mid a \in A\}$.

By now we have all the tools needed to prove this proposition. We'll prove this by invoking uniqueness.

Proof. Fix any $x \in \mathbb{R}$, and consider the measure μ_x defined by

$$\mu_x(A) = \mu(A + x)$$

for all Borel sets A (where μ denotes the Lebesgue measure). This is well-defined, since if you take a Borel set and add a fixed number, you still get a Borel set.

Then we have that μ_x is a Borel measure with the property that for every interval, we have

$$\mu_x([a,b]) = \mu([a+x, b+x]) = (b+x) - (a+x) = b - a.$$

So the measure μ_x is a Borel measure such that the measure of any interval is equal to its length. And now we can invoke uniqueness — by uniqueness, μ_x *has* to be the Lebesgue measure, because the Lebesgue measure is the *unique* Borel measure satisfying this property.

So by the uniqueness of the Lebesgue measure, we have that $\mu_x = \mu$. But that's exactly what we wanted to prove (because x was arbitrary). □

As we can see, the proof was almost straightforward, but it required as input the uniqueness of the Lebesgue measure, which is nontrivial (we had to do some preparation in order to prove that, and of course the essence behind that was Dynkin's π-system lemma from the first lecture).

Question 3.3. Are there other Borel translation-invariant measures?

The answer is no, in a certain sense; this gives another characterization of the Lebesgue measure in some sense.

Proposition 3.4

Suppose that $\widetilde{\mu}$ is a Borel measure on $\mathbb{R}$ which is translation-invariant and such that $\widetilde{\mu}([0,1]) = 1$. Then $\widetilde{\mu}$ is the Lebesgue measure.

Here we had to require $\widetilde{\mu}([0,1]) = 1$ to fix the scaling; otherwise we'd get that $\widetilde{\mu}$ is a multiple of the Lebesgue measure.

Proof. Based on the characterization of the Lebesgue measure, it suffices to show that

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

for all $a < b$ (because this actually characterizes the Lebesgue measure).

The idea behind this is that if we fix some $a < b$, then because $\mathbb{Q}$ is dense in $\mathbb{R}$, we can find sequences $(p_n), (q_n) \subseteq \mathbb{Q}$ such that $p_n \downarrow a$ and $q_n \uparrow b$ as $n \to \infty$ (this means (p_n) is decreasing and converges to a, and (q_n) is increasing and converges to b). Then we have

$$\widetilde{\mu}([a,b]) = \lim_{n\to\infty} \widetilde{\mu}([p_n, q_n]).$$

So this means it suffices to show the equality $\widetilde{\mu}([a,b]) = b-a$ when $a,b \in \mathbb{Q}$ — because then we get that $\widetilde{\mu}([p_n,q_n]) = q_n - p_n$, and $q_n - p_n \to b-a$.

Now we're going to use translational invariance to prove that this equality is true for rational numbers. For rational $p<q$, we first have $\mu(\{p\}) = \lim_{n\to\infty} \widetilde{\mu}([p,p+\frac{1}{n}])$.

Claim 3.5 — We have $\widetilde{\mu}([p,p+\frac{1}{n}]) = \frac{1}{n}$.

Proof. First, by translational invariance, we have $\widetilde{\mu}([p,p+\frac{1}{n}]) = \widetilde{\mu}([0,\frac{1}{n}])$. But we can also partition $[0,1]$ into the smaller intervals $[0,\frac{1}{n}]$, $[\frac{1}{n},\frac{2}{n}]$, …; and all these intervals have the same measure, by translational invariance. So we actually have $\widetilde{\mu}([0,1]) = n\cdot\widetilde{\mu}([0,\frac{1}{n}])$. And this is 1, giving $\widetilde{\mu}([0,\frac{1}{n}]) = \frac{1}{n}$. □

This immediately gives us $\mu(\{p\}) = \frac{1}{n}$. Furthermore, if we take any $p = \frac{m}{n}$, then the same reasoning gives that $\widetilde{\mu}([0,\frac{m}{n}]) = m\cdot\widetilde{\mu}([0,\frac{1}{n}]) = \frac{m}{n}$. (The point is that we divide $[0,\frac{m}{n}]$ into the smaller intervals $[0,\frac{1}{n}]$, $[\frac{1}{n},\frac{2}{n}]$, …, and we know that each of these has measure 1.)

More generally, if we take any two rational numbers $p<0<q$, then we get $\mu([p,q]) = \widetilde{\mu}([p,0]) + \widetilde{\mu}([0,q]) = -p+q$.

Using this, we can get that this is true for *any* $p,q\in\mathbb{Q}$.

And then we use our above argument (approximating reals with rationals) to conclude $\widetilde{\mu}([a,b]) = b-a$. And then using the uniqueness of the Lebesgue measure, we get that $\widetilde{\mu}$ has to be the Lebesgue measure. □

Student Question. *Where did we use $\widetilde{\mu}(\{p\}) = 0$?*

Answer. This is because when you take the limit, you actually need to say that $\mu(\{a\}) = \mu(\{b\}) = 0$ — our equality with the limit actually should say that

$$\widetilde{\mu}((a,b)) = \lim_{n\to\infty} \widetilde{\mu}([p_n,q_n]).$$

And this argument shows that $\widetilde{\mu}$ has no atoms, which means we can replace the open interval with a closed one.

Student Question. *How did we get to the limit $\widetilde{\mu}(\{p\}) = \lim_{n\to\infty}\widetilde{\mu}([p,p+\frac{1}{n}])$?*

Answer. The point is that you can write $\{p\} = \bigcap_n [p,p+\frac{1}{n}]$; and when you have a decreasing sequence of sets, the measure of the limit (i.e., intersection) is the limit of the measures. (We'll prove this on the problem set.)

Student Question. *How do we conclude after having shown that $\widetilde{\mu}([0,\frac{m}{n}]) = \frac{m}{n}$?*

Answer. If you fix any rational numbers $p<0<q$, then you can split $[p,q] = [p,0)\cup[0,q]$, which means $\mu([p,q]) = \mu([p,0)) + \mu([0,q])$. And our arguments above show that the measure of the first interval is $-p$ and the second is q.

Remark 3.6. We mentioned last time that the σ-algebra formed by Lebesgue measurable sets is actually bigger than the Borel σ-algebra. It's highly nontrivial to construct a set that's Lebesgue measurable but not Borel; we won't do it here, but they have been constructed.

There's also the question of whether every subset of $\mathbb{R}$ is Lebesgue measurable; the answer is no, and there's a brief and clever construction (but we also don't have time to discuss it in class; we can ask about it in office hours). In other words, $\mathcal{U}$ (the σ-algebra constructed in Caratheodory's extension theorem) is not $\mathcal{P}(\mathbb{R})$ — there's a subset of $\mathbb{R}$ which is not here.

§3.4 Probability — some definitions

Now we'll talk a bit about probability and introduce some basic concepts we'll use in this course; after all, this course is called probability and not measure theory.

Definition 3.7 (Probability measure and space)**.** Let $(E, \mathcal{E})$ be a measure space with $\mu(E) = 1$. Then we say μ is a probability measure, and we say $(E, \mathcal{E}, \mu)$ is a probability space.

Probability theory focuses on this kind of space, where the mass of the entire space is 1. From now on, we're going to denote our probability space by Ω, our σ-algebra by $\mathcal{F}$, and our probability measure by $\mathbb{P}$.

Definition 3.8 (Sample space)**.** In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we call Ω the sample space.

Intuitively, you think of Ω as the set of 'all possible results.'

Definition 3.9 (Event)**.** In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we call the elements of $\mathcal{F}$ events.

Finally, the *probability* of an event will simply be its measure.

Definition 3.10 (Probability)**.** For each event $A \in \mathcal{F}$, we call $\mathbb{P}[A]$ the probability of A.

These are standard definitions, but we should write them down once, and we'll refer to them from now on. From now on, we'll only focus on probability spaces.

§3.5 Independence

One fundamental notion in probability is that of independence; let's see what we mean by that.

Definition 3.11 (Independence of events)**.** A sequence of events (A_n) is called independent if for every finite $J \subseteq \mathbb{N}$, we have $\mathbb{P}[\bigcap_{n\in J} A_n] = \prod_{n\in J} \mathbb{P}[A_n]$.

In words, this means if we take any finite collection of our sets A_n, then the probability of the intersection of those events is the product of the individual probabilities.

Intuitively, two events are independent if the occurrence of the first event doesn't affect the other, and vice versa — if we have any two events A and B, the fact that A occurs doesn't depend on whether or not B occurs. That's the intuitive interpretation of independence. For example, if we take a fair coin and flip it twice, the two results are completely independent — the second time we toss the coin, it really doesn't matter what we had before.

But of course, we'll see in this course much more complicated independent events.

And we can do the same with σ-algebras — we can also talk about independent σ-algebras.

Definition 3.12 (Independence of σ-algebras)**.** A sequence of σ-algebras $(\mathcal{A}_n)$ such that $\mathcal{A}_n \subseteq \mathcal{F}$ for all n is said to be independent if for any sequence (A_n) such that $A_n \in \mathcal{A}_n$ for all n, the sequence (A_n) is independent.

So a collection of σ-algebras is independent if when we take *any* collection of events, each from the corresponding σ-algebra, then this collection of events is independent in the way described earlier. So we can also talk about the independence of σ-algebras.

The connection between these two definitions is the following (which is an easy exercise, which we're recommended to do to get familiar with these notions).

Proposition 3.13

The events (A_n) are independent if and only if the σ-algebras $(\sigma(\{A_n\}))$ are independent.

It's not too difficult to see why this is true; you just apply the definitions of independence of σ-algebras and collections of events.

In general, it's nice to have some methods to prove whether two σ-algebras are independent — because the concept of independence is of fundamental importance in probability. As earlier, we often want to restrict ourselves to simpler sets, e.g., π-systems — if you have two σ-algebras generated by two π-systems, and the π-systems are independent, we'd like to say the σ-algebras are independent. And indeed, we have the following theorem, which allows us to check independence between σ-algebras.

Theorem 3.14

Suppose that $\mathcal{A}_1$ and $\mathcal{A}_2$ are both π-systems in $\mathcal{F}$. If these are independent, in the sense that

$$\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_1]\mathbb{P}[A_2]$$

for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, then we have that $\sigma(\mathcal{A}_1)$ and $\sigma(\mathcal{A}_2)$ are independent.

(From now on, we always assume we're living in our big σ-algebra $\mathcal{F}$.)

Based on what we've done so far, this is not surprising; and you can start to see why π-systems are important in general. So we have two σ-algebras generated by two π-systems. These may be extremely weird and complicated to work with, but if we know the π-systems that generate them are independent in this sense, then we actually get to conclude that these σ-algebras are independent.

The main idea of the proof is that if we have a finite measure, and if we know the measures on the π-system, then we actually know it on the entire σ-algebra generated by that π-system (we saw this last class).

Proof. Suppose we fix some $A_1 \in \mathcal{A}_1$, and we consider the measures μ and ν defined by

$$\mu(A) = \mathbb{P}[A \cap A_1],$$
$$\nu(A) = \mathbb{P}[A]\mathbb{P}[A_1]$$

(for all $A \in \mathcal{F}$). It's easy to see these both satisfy the properties of measures, so they're measures. And they're finite measures (though not necessarily probability measures) — in particular, we have $\mu(\Omega) = \mathbb{P}[A_1] \leq 1 < \infty$, and $\nu(\Omega) = \mathbb{P}[A_1] < \infty$ as well; this in particular means $\mu(\Omega) = \nu(\Omega)$.

Furthermore, we have $\mu|_{\mathcal{A}_2} = \nu|_{\mathcal{A}_2}$, by our hypothesis. So we have two measures that agree on a π-system, and this means they have to agree on the σ-algebra generated by this π-system $\mathcal{A}_2$ — in particular, what this means is that

$$\mathbb{P}[A \cap A_1] = \mathbb{P}[A]\mathbb{P}[A_1]$$

for all $A \in \sigma(\mathcal{A}_2)$.

We're not quite done, because we need this to hold for every $A_1 \in \sigma(\mathcal{A}_1)$, and right now we only have it for every $A_1 \in \mathcal{A}_1$. But completing the proof is easy — to do so, fix any $A \in \sigma(\mathcal{A}_1)$, and now consider the measures

$$\widetilde{\mu}(B) = \mathbb{P}[A \cap B]$$
$$\widetilde{\nu}(B) = \mathbb{P}[A]\mathbb{P}[B]$$

for every $B \in \mathcal{F}$. Then the same story happens — $\widetilde{\mu}$ and $\widetilde{\nu}$ coincide on our π-system $\mathcal{A}_1$, meaning that $\widetilde{\mu}|_{\mathcal{A}_1} = \widetilde{\nu}|_{\mathcal{A}_1}$, and this means they actually coincide on the σ-algebra generated by $\mathcal{A}_1$. And by definition, this means

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

for every $B \in \sigma(\mathcal{A}_1)$ (and this is true for every $A \in \sigma(\mathcal{A}_2)$, so we're done). □

What this means is it suffices to check independence only on the π-system, which might be much simpler (e.g., for the Borel σ-algebra on $\mathbb{R}$, we can instead take the π-system consisting of unions of disjoint intervals, which is much easier to deal with).

Student Question. *Why is $\mu|_{\mathcal{A}_2} = \nu|_{\mathcal{A}_2}$?*

Answer. This is by our hypothesis that $\mathbb{P}[A_1 \cap A_2] = \mathbb{P}[A_1]\mathbb{P}[A_2]$ for all A_2 in our π-system $\mathcal{A}_2$.

Student Question. *Does it make sense to ask about independence for an uncountable family of events?*

Answer. Yes, but you have to be careful. For example, you can take any finite collection of that collection and say it's independent — you use the exact same definition, just without the assumption that the index set is countable.

Student Question. *How did we get to $\widetilde{\mu}|_{\mathcal{A}_1} = \widetilde{\nu}|_{\mathcal{A}_1}$?*

Answer. This comes from the previous part — we proved that $\mathbb{P}[A \cap A_1] = \mathbb{P}[A] \cdot \mathbb{P}[A_1]$ for every $A \in \sigma(\mathcal{A}_2)$, but this is true for *every* $A_1 \in \mathcal{A}_1$. So now we're fixing $A \in \sigma(\mathcal{A}_2)$, and letting A_1 vary over $\mathcal{A}_1$.

§3.6 The Borel–Cantelli lemmas

This is a very nice way to check whether two σ-algebras are independent. Another thing that we're interested in in probability is to check other types of events. For example, you might have a collection of events, and wonder what's the probability this collection happens infinitely many times. For example, suppose you toss a coin and ask the probability of the event you get heads infinitely many times. Intuitively, this should be 1, but how do we formalize that?

For that, we need a few notions — namely, the lim sup and lim inf of a collection of events (we'll use them extensively in our course, and they're used extensively in statistics as well).

Definition 3.15. Suppose we have a sequence of events (A_n). Then we define

$$\limsup_{n\to\infty} A_n = \bigcap_n \bigcup_{m \geq n} A_m \quad \text{and} \quad \liminf_{n\to\infty} A_n = \bigcup_n \bigcap_{m \geq n} A_m.$$

What $\limsup A_n$ means is that for every n, there exists m larger than n such that A_m occurs. In other words, $\limsup A_n$ is the event that infinitely many A_n's occur. Similarly, $\liminf A_n$ means there exists n such that for every $m \geq n$, we have that A_m occurs — so this is the event that eventually all these events occur.

We'll be interested in computing the probabilities of events of this form as well, and for that we'll need the Borel–Cantelli lemmas.

Example 3.16

For the coin example, we can define A_n as the event that the nth toss is heads; then $\limsup A_n$ is the event that we get infinitely many heads.

Now we'll state some results on these definitions; the first such result is the first Borel–Cantelli lemma, which is used extensively in mathematics.

Lemma 3.17 (Borel–Cantelli I)

Let (A_n) be any collection of events. If $\sum_n \mathbb{P}[A_n] < \infty$, then $\mathbb{P}[\limsup A_n] = 0$.

So if we take any collection of events and the sum of their probabilities is finite, then the probability that infinitely many of them happen is 0.

Proof. Let $A = \limsup A_n$. Then by definition, we have

$$\mathbb{P}[A] = \mathbb{P}\left[\bigcap_n \bigcup_{m \geq n} A_m\right].$$

But then we have

$$\mathbb{P}[A] \leq \mathbb{P}\left[\bigcup_{m \geq n} A_m\right]$$

for all n. And by countable subadditivity, this is at most $\sum_{m \geq n} \mathbb{P}[A_m]$. But this tends to 0 as $n \to \infty$, because $\sum_n \mathbb{P}[A_n]$ is finite (which is true if and only if its tail sums go to 0).

And this means $\mathbb{P}[A] = 0$, so we're done. □

This proof is only a few lines, but the lemma is indeed very important. We first applied the definition, then the fact that measures are increasing functions ($\bigcap_n \bigcup_{m \geq n} A_m$ is contained in $\bigcup_{m \geq n} A_m$ for any specific m); and then we used countable subadditivity to get a sum that goes to 0.

Student Question. *What does the* lim sup *mean intuitively?*

Answer. It's the event that infinitely many of the A_n's happen; and lim inf represents that eventually all of them happen.

Student Question. *We didn't use the fact that $\mathbb{P}$ is finite here, did we?*

Answer. Yes, we didn't — in fact, this holds for every measure. We don't use the fact that $\mathbb{P}$ is finite. So if we replace $\mathbb{P}$ by an arbitrary measure μ, we still have the same statement — that if $\sum \mu(A_n) < \infty$, then $\mu(\limsup A_n) = 0$.

Now we'll proceed with the second Borel–Cantelli lemma, which is also extremely important. Here we need to make some further assumptions — specifically, that our events are independent.

Lemma 3.18 (Borel–Cantelli II)

Let (A_n) be a sequence of *independent* events. Then if $\sum_n \mathbb{P}[A_n] = \infty$, then we have $\mathbb{P}[\limsup A_n] = 1$.

This is a way to rigorously prove that if we toss a fair coin infinitely many times, then with probability 1 we get heads infinitely many times (we can define A_n to be the event that the nth toss is heads; clearly the A_n's are independent, and this lemma says that then with probability 1 we get heads infinitely many times). So this is something we know intuitively, and we can prove it rigorously using this lemma.

Proof. The first thing that we need is the following:

> **Claim 3.19 —** The sequence (A_n^c) is also independent.

We'll leave this as an exercise (it may be included in some problem set; but otherwise we can just accept that it's true, but we're suggested to try to prove it).

Then we have that if we take the probability of the first N terms, we have

$$\mathbb{P}\left[\bigcap_{m=n}^{N} A_m^c\right] = \prod_{m=n}^{N} \mathbb{P}[A_m^c]$$

(this follows by the definition of independence). But we can rewrite this as

$$\prod_{m=n}^{N} (1 - \mathbb{P}[A_m]).$$

And now we somehow want to bound this term on the right. To do so, there's a simple inequality $1+x \le e^x$ for all $x \in \mathbb{R}$ (this is something from analysis, and you can easily see it by taking derivatives, to prove the functino $e^x - x - 1$ is decreasing up to $x = 0$ and then increasing). So we get that our probability is at most

$$\prod_{m=n}^{N} e^{-\mathbb{P}[A_n]} = e^{-\sum_{m=n}^{N} \mathbb{P}[A_n]}.$$

And this sum tends to ∞ (for any fixed n, as $N \to \infty$), which means the exponential tends to 0; and what this means is that

$$\mathbb{P}\left[\bigcap_{m=n}^{\infty} A_m^c\right] = 0$$

for all n. And by countable subadditivity, this implies

$$\mathbb{P}\left[\bigcup_{n\ge 1}\bigcap_{m\ge n} A_n^c\right] \le \sum_{n\ge 1} \mathbb{P}\left[\bigcap_{m\ge n} A_m^c\right] = 0.$$

And if the probability of this event is 0, then the probability of its *complement* is 1; and the complement of $\bigcup_{n\ge1}\bigcap_{m\ge n} A_m^c$ is just $\limsup A_n$, giving that $\mathbb{P}[\limsup A_n] = 1$. And that's exactly what we wanted to prove. □

Here we did use the fact that the measure is finite — we used the fact that $\mathbb{P}[A] = 1 - \mathbb{P}[A^c]$ several times, and we need finiteness for that.

> **Student Question.** *Are there similar results for the* $\liminf$*?*
>
> **Answer.** Usually there's duality — you can write $\liminf A_n = (\limsup A_n^c)^c$. We state results for $\limsup$ because it's easier to work with, but you can use this duality to get the corresponding results for $\liminf$.

§3.7 Measurable functions

We'll now introduce measurable functions, which are a fundamental object in measure theory (and in probability theory).

Recall that when we have a function $f: E \to F$, where E and F are some topological spaces, we say f is *continuous* if $f^{-1}(U)$ is open in E whenever U is open in F. With measurable functions, we do the same; we just replace the notion of an open set with the notion of a measurable set. So a function will be called measurable if it inverts measurable sets to measurable sets (where we have σ-algebras instead of topological spaces). We'll now write this down properly, but this is the main idea (coming from topology).

Definition 3.20 (Measurable functions)**.** Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be two measure spaces. Then we say a map $f: E \to G$ is measurable if for every $A \in \mathcal{G}$, we have that $f^{-1}(A) \in \mathcal{E}$.

Note that by $f^{-1}(A)$, we mean $\{x \in E \mid f(x) \in A\}$.

So this is the same definition as with continuous functions, but now we're dealing with measurable sets instead of open sets. From now on, we're going to deal with such functions.

Definition 3.21. If E is a topological space with the Borel σ-algebra (meaning that $\mathcal{E} = \mathcal{B}(E)$), then we say f is a Borel function.

So a Borel function is a function that's measurable with respect to the Borel σ-algebra (on the first set).

Remark 3.22. Recall that the Borel σ-algebra is the σ-algebra generated by the open subsets of E (in our topology).

Student Question. *How do you prove that if A and B are independent then $\sigma(A)$ and $\sigma(B)$ are independent?*

Answer. The σ-algebra generated by A is simply $\sigma(A) = \{\emptyset, A, A^c, E\}$, and likewise with $\sigma(B)$. Then it's just a matter of checking that if we take any elements from each of these, the probability of their intersection is the product of probabilities.

Student Question. *In topology, you have the notion of a basis (which you can generate the open sets from). Is there something similar here?*

Answer. Yes. The point is that in order to check that a function is measurable, it suffices to restrict to π-systems. The role of open sets in a topology is played by π-systsems in a σ-algebra — if you know that f inverts into measurable sets for elements of a π-system, then it inverts *any* measurable set to a measurable set.

§4 September 17, 2024

§4.1 Measurable functions

Last lecture, we proved several results in measure theory, such as the Borel–Cantelli lemmas. We also defined the notion of a *measurable function* — this is analogous to the definition of a continuous function.

Question 4.1. In practice, how do we check whether a function is measurable?

We said a function is measurable if it inverts measurable sets to measurable sets. It turns out we're lucky — if we have a π-system generating our σ-algebra and the function inverts elements of that system to measurable sets, then the function is measurable. In particular, we have the following lemma.

Lemma 4.2

Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measurable spaces such that $\mathcal{G} = \sigma(\mathcal{A})$ for some $\mathcal{A}$. If $f^{-1}(A) \in \mathcal{E}$ for every $A \in \mathcal{A}$, then f is measurable.

So we don't actually have to check that the function inverts *every* element of $\mathcal{G}$; we just have to check this on a set of generators. This makes it much easier to check whether a function is measurable.

The proof has a similar flavor to things we've done so far.

Proof. Since $\mathcal{G}$ is generated by $\mathcal{A}$, it suffices to show that

$$\mathcal{F} = \{A \subseteq G \mid f^{-1}(A) \in \mathcal{E}\}$$

is a σ-algebra. (This is because we know this is a collection of sets containing $\mathcal{A}$, by our hypothesis; so then it's a σ-algebra containing $\mathcal{A}$, which means it has to contain $\sigma(\mathcal{A}) = \mathcal{G}$; and then $f^{-1}(A) \in \mathcal{E}$ for every $A \in \mathcal{G}$, which is what we want.)

This is fairly easy to see. First, we have $f^{-1}(\emptyset) = \emptyset$, so $\emptyset$ indeed belongs to this family $\mathcal{F}$.

Also, if we take any $A \in \mathcal{F}$, then $f^{-1}(G \setminus A) = E \setminus f^{-1}(A)$ (this is easy to check set-theoretically). And since $f^{-1}(A) \in \mathcal{E}$ and σ-algebras are closed under complements, this set also belongs to $\mathcal{E}$, as desired.

Finally, we have $f^{-1}(\bigcup_n A_n) = \bigcup_n f^{-1}(A_n)$. And if $(A_n) \subseteq \mathcal{F}$, then all the sests on the right-hand side are in $\mathcal{E}$, so their union is too; and then $\bigcup_n A_n \in \mathcal{F}$.

So we've shown $\mathcal{F}$ is a σ-algebra, and it contains our generating set $\mathcal{A}$; so it must contain $\sigma(\mathcal{A})$. □

Note that $\mathcal{A}$ doesn't have to be a π-system — it can be *any* collection of sets generating the σ-algebra. Most of the time when we use it, we'll take $\mathcal{A}$ to be a π-system; but this actually holds for any collection of sets.

§4.1.1 Some examples

Now let's see a few applications of this lemma to prove that a given function is indeed measurable.

Example 4.3

Suppose $f: E \to \mathbb{R}$, and we want to check whether f is Borel measurable (i.e., measurable with respect to the Borel σ-algebra on $\mathbb{R}$). The Borel σ-algebra on $\mathbb{R}$ is generated by sets of the form $(-\infty, x]$ — i.e.,

$$\mathcal{B}(\mathbb{R}) = \sigma(\{(-\infty, x] \mid x \in \mathbb{R}\}).$$

(This is because intervals generate $\mathcal{B}(\mathbb{R})$, and they can be written as intersections of such sets.) So to check that f is $\mathcal{B}(\mathbb{R})$-measurable, it suffices to check that

$$\{x \in E \mid f(x) \leq y\} = f^{-1}((-\infty, y])$$

is in $\mathcal{E}$ for all y.

So that's how we check that a real-valued function is Borel measurable; this is usually much easier than taking every single Borel set and checking that its inverse image is also measurable.

Example 4.4

Suppose E and F are topological spaces, and $f: E \to F$ is continuous. Then f is measurable (under the Borel σ-algebras on E and F).

Proof. First, for any *open* set $U \subseteq F$, we have that $f^{-1}(U) \subseteq E$ is open (by the definition of continuous functions — a continuous function inverts open sets into open sets). But we have that the Borel σ-algebra on F is generated by all open subsets of F — i.e.,

$$\mathcal{B}(F) = \sigma(\{G \subseteq F \mid G \text{ open}\}).$$

This means f is $\mathcal{B}(F)$-measurable (because f inverts open sets to open and therefore measurable sets, and the open sets generate the σ-algebra on F). □

Example 4.5

For $A \subseteq E$, the *indicator function* of A is the function $1_A: E \to \mathbb{R}$ defined as

$$x \mapsto \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

Then 1_A is measurable if and only if $A \in \mathcal{E}$.

Example 4.6

Constant functions are measurable; the identity function is also measurable.

Example 4.7

The composition of measurable functions is also measurable. More precisely, if $(E, \mathcal{E})$, $(F, \mathcal{F})$, and $(G, \mathcal{G})$ are measurable spaces and $f: E \to F$ and $g: F \to G$ are measurable, then the composition $g \circ f: E \to G$ is also measurable.

Proof. The proof is just set theory — if we take any $A \in \mathcal{G}$, then since g is measurable, we have $g^{-1}(A) \in \mathcal{F}$; this then means $f^{-1}(g^{-1}(A)) \in \mathcal{E}$. But $(g \circ f)^{-1}(A) = f^{-1}(g^{-1}(A))$, so we are done. □

§4.2 σ-algebras generated by functions

Before constructing other measurable spaces, let's mention a few words about the σ-algebra generated by a *function*. So far we've seen σ-algebras generated by collection of sets, but you can also generate a σ-algebra by a function. For that, we need the following definition.

Definition 4.8 (σ-algebra generated by functions)**.** Suppose we have a set E and a family of real-valued functions $\{f_i\}_{i \in I}$ on E (for some index set I). Then we define the σ-algebra generated by $\{f_i\}_{i \in I}$ as

$$\sigma(\{f_i\}_{i \in I}) = \sigma(f_i^{-1}(A) \mid A \in \mathcal{B}(\mathbb{R}),\ i \in I).$$

In other words, this is the smallest σ-algebra on E that makes all the functions measurable (that's the intuition behind this).

§4.3 Product measure spaces

We'll soon see a way to take a collection of measurable functions and construct new measurable functions. For this, we'll need the notion of a *product* of measurable spaces.

Definition 4.9 (Product measurable spaces)**.** Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measure spaces. Then we define the product measure space, as $E \times G$ with the σ-algebra $\mathcal{E} \otimes \mathcal{G}$ generated by the following two functions:

- $\pi_1 : E \times G \to E$ is the projection onto E — i.e., the map $(x, y) \mapsto x$.
- $\pi_2 : E \times G \to G$ is the projection onto G — i.e., the map $(x, y) \mapsto y$.

In other words, the product σ-algebra $\mathcal{E} \otimes \mathcal{G}$ is the σ-algebra generated by rectangles — i.e.,

$$\mathcal{E} \otimes \mathcal{G} = \sigma(\{A \times B \mid A \in \mathcal{E}, B \in \mathcal{G}\})$$

(these are two ways of writing the same thing, in terms of functions vs. sets).

We can generalize this notion to allow products of an arbitrary number of sets as well.

Definition 4.10. If we have a collection of measure spaces $\{(E_i, \mathcal{E}_i)\}_{i \in I}$, then the product measure space has underlying set $\prod_{i \in I} E_i$, with the σ-algebra generated by the projection maps $\pi_i : \prod_{j \in I} E_j \to E_i$.

So here we're doing the same thing as before, but now we might have infinitely many terms we're taking the product of (the index set here is arbitrary; in the previous case the index set was $I = \{1, 2\}$).

Student Question. *How is the product* set *defined if I is uncountable?*

Answer. The definition is a bit subtle — it's actually in terms of sequences. When I is countable, this is simple; you can just take sequences indexed by the integers. When I is $\mathbb{R}$ for example, this is more complicated, but it can still be done.

This gives us another criterion for measurability.

Proposition 4.11

Let $f_i : E \to F_i$ be functions (indexed by a set I). Then the functions f_i are all measurable if and only if the function $(f_i) : E \to \prod_{j \in I} f_j$ defined by $x \mapsto (f_i(x))_{i \in I}$ is measurable.

So we fix x and then consider the elements of the product space with coordinates $f_i(x)$; and *every* one of the functions is measurable if and only if this one function defined on the product space is measurable.

Remark 4.12. Note that (f_i) is a *single* function.

Proof. First, if the map (f_i) is measurable, then for each i, we can write $f_i = \pi_i \circ (f_i)$, which means f_i is indeed measurable (we proved that a composition of measurable functions is measurable). So one direction is clear — if our function defined on the product space is measurable, then each of these functions is measurable.

Conversely, suppose that all the f_i are measurable. Then we know that the σ-algebra of $\prod_{i \in I} F_i$ is generated by inverse images of the projections — i.e., sets of the form $\pi_j^{-1}(A)$ over $A \in \mathcal{F}_j$ and $j \in I$. And we also know that $(f_i)^{-1}(\pi_j^{-1}(A)) = f_j^{1}(A) \in \mathcal{E}$ for any j (because f_j is measurable). So this means $(f_i)^{-1}$ inverts elements of our generating set to measurable sets; and then it has to be measurable by our previous result. □

§4.4 Operations with measurable functions

Now that we have this, we'll see a nice way to prove that a whole lot more functions are measurable.

Proposition 4.13

Let $(E, \mathcal{E})$ be a measurable space, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative measurable functions on E. Then the following functions are measurable:

- $f_1 + f_2$ and $f_1 \cdot f_2$;
- $\max\{f_1, f_2\}$ and $\min\{f_1, f_2\}$;
- $\inf_{n\in\mathbb{N}} f_n$ and $\sup_{n\in\mathbb{N}} f_n$;
- $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$.

So if we start with a collection of measurable functions (f_n), then all these functions are measurable; this is a nice way to prove lots of measurable functions.

We're going to prove the first two cases (the sum and product) using the previous propositions; Prof. Kavvadias suggests we try to prove the remainder as a nice exercise.

Proof. To prove that $f_1 + f_2$ is measurable, we're going to apply the previous proposition to an appropriate function. Define $(+)\colon [0,\infty) \times [0,\infty) \to [0,\infty)$ as $(x, y) \mapsto x + y$; this is a nice well-defined function. Then we can write

$$f_1 + f_2 = (+) \circ (f_1, f_2),$$

where (f_1, f_2) is defined in the same way as the above proposition — explicitly, $(f_1, f_2)\colon F \to \mathbb{R} \times \mathbb{R}$ is the function $x \mapsto (f_1(x), f_2(x))$. By the previous proposition, we know (f_1, f_2) is measurable; and $(+)$ is measurable because it's continuous. So $f_1 + f_2$ is measurable, since it's a composition of two measurable functions.

The proof for $f_1 \cdot f_2$ is exactly the same thing, except we replace the addition with multiplication (and write $f_1 f_2$ as the composition of the multiplication function with the product function). The rest of these statements can be proven in a similar way. □

Student Question. *Where are we using nonnegativity?*

Answer. We're actually not using it here; we can replace $[0,\infty)$ with $\mathbb{R}$ everywhere (so we shouldn't need this condition).

This gives us machinery for producing lots of measurable functions.

§4.5 The monotone class theorem

We'll now state and prove a very important theorem, the analog of Dynkin's lemma. It sounds awkward, but it's very useful (as we'll see when rigorously constructing the Lebesgue integral and so on).

Theorem 4.14 (Monotone class theorem)

Let $(E, \mathcal{E})$ be a measurable space, and let $\mathcal{A}$ be a π-system which generates $\mathcal{E}$, i.e., with $\sigma(\mathcal{A}) = \mathcal{E}$. Let $\mathcal{V}$ be a vector space of functions with the following properties:

(i) We have $1_E \in \mathcal{V}$ (where 1_E is the all-1's function).

(ii) For any $A \in \mathcal{A}$, we have $1_A \in \mathcal{V}$.

(iii) $\mathcal{V}$ is closed under *bounded monotone limits*. More precisely, if we have a collection of functions $(f_n) \subseteq \mathcal{V}$ which are bounded and nonnegative and such that $f_n \uparrow f$ pointwise, then $f \in \mathcal{V}$.

Then $\mathcal{V}$ contains *all* bounded measurable functions on E.

The third assumption means that if we take any function f which is the pointwise limit of an increasing, bounded, nonnegative collection of functions on $\mathcal{V}$, then this function f also has to belong to $\mathcal{V}$. (We use $\uparrow$ to denote an increasing limit.)

To recap a bit, we have a measurable space $(E, \mathcal{E})$, and we have a π-system generating the σ-algebra. Then we consider a vector space of functions, meaning that if we take two functions and sum them or scale one of them, the resulting function is still in our space. And we suppose that this vector space contains the constant-1 function, and the indicator function of every $A \in \mathcal{A}$; and that it has the property that if we take any collection of bounded nonnegative functions which converge increasingly (pointwise) to some f, then $f \in \mathcal{V}$. And then we can conclude that $\mathcal{V}$ contains *all* bounded monotone sets.

Student Question. *Doesn't the first assumption follow from the second?*

Answer. No — $\mathcal{A}$ is a π-system, not necessarily a σ-algebra.

By 'measurable' we mean *Borel* measurable — in particular, all these functions map $E \to \mathbb{R}$.

Student Question. *Are Borel and Lebesgue measurable interchangeable?*

Answer. Lebesgue measurable is with respect to the Lebesgue σ-algebra, and Borel measurable is with respect to the Borel σ-algebra. There are functions which are Lebesgue-measurable but not Borel-measurable — we've mentioned there are such *sets*, and you can take the indicator vector of one.

Student Question. *Does* bounded *require the functions to be* uniformly *bounded?*

Answer. No, we just need each function to be bounded separately.

Now we'll prove this. This is pretty similar to things we've seen before — the idea is that the conditions on $\mathcal{V}$ are similar to that of a d-system, and taking a monotone limit is similar to taking an increasing union.

Proof. First we're going to see that $1_A \in \mathcal{V}$ for *all* $A \in \mathcal{E}$ (i.e., the indicator function of any measurable set is in $\mathcal{V}$). To do this, we consider

$$\mathcal{D} = \{A \in \mathcal{E} \mid 1_A \in \mathcal{V}\}.$$

We know by assumption that $\mathcal{D} \supseteq \mathcal{A}$. So by Dynkin's π-system lemma, it suffices to show $\mathcal{D}$ is a d-system — if it contains our generating set, then it has to contain the σ-algebra generated by that generating set by Dynkin's π-system lemma (since our generating set $\mathcal{A}$ is a π-system).

So we want to show $\mathcal{D}$ is a d-system. For this, we first know that $1_E \in \mathcal{V}$, which means $E \in \mathcal{D}$.

The next thing we need to prove (as in the definition of a d-system) is that $\mathcal{D}$ is closed under complements. If we take some A such that $1_A \in \mathcal{V}$, then we have

$$1_{E \setminus A} = 1_E - 1_A.$$

And this is where we use the assumption that we have a *vector space* — we know $1_A \in \mathcal{V}$ by the definition of $\mathcal{D}$, and $1_E \in \mathcal{V}$ as well; so their difference $1_E - 1_A$ is also in $\mathcal{V}$ (because $\mathcal{V}$ is a vector space). This implies $E \setminus A \in \mathcal{D}$; so $\mathcal{D}$ is indeed closed under taking complements.

The third condition is that we need $\mathcal{D}$ to be closed under increasing unions of sets. In order to show this, take some increasing sequence (A_n) in $\mathcal{D}$. Then the indicator functions 1_{A_n} increasingly converge pointwise to $1_{\bigcup A_n}$ as $n \to \infty$; this means we must have $1_{\bigcup A_n} \in \mathcal{V}$ by the third property of $\mathcal{V}$. And this means $\bigcup A_n \in \mathcal{D}$.

So $\mathcal{D}$ is a d-system (because it satisfies the three properties of a d-system), and therefore $\mathcal{D}$ is the entire σ-algebra $\sigma(\mathcal{A}) = \mathcal{E}$ (by Dynkin's π-system lemma).

(These ideas are pretty standard, and we've seen them before — how to use the π-system lemma in this way.)

Now we've shown $\mathcal{V}$ contains the indicator functions of measurable sets. We're not done, though, because we need to show that $\mathcal{V}$ contains *arbitrary* bounded measurable functions. The idea for how we'll do this is by approximating it by linear combinations of such indicator functions.

Suppose we have a function $f: E \to [0, \infty)$ which is bounded and measurable. (We're assuming nonnegativity for now; we'll relax this later.) We want to show that then $f \in \mathcal{V}$. And the idea is that we're going to approximate them.

To do so, we define $f_n = 2^{-n} \lfloor 2^n f \rfloor$. This is maybe a weird definition, but we can express it as a power series as

$$f_n = \sum_{k=0}^{\infty} k \cdot 2^{-n} \cdot 1_{\{k \cdot 2^{-n} \le f < (k+1) \cdot 2^{-n}\}}.$$

This is well-defined because f is bounded, so this summation will stop at some point (for each fixed n). Explicitly, since f is bounded, there exists N such that

$$f_n = \sum_{k=0}^{N \cdot 2^n} k \cdot 2^{-n} \cdot 1_{\{k \cdot 2^{-n} \le f < (k+1) \cdot 2^{-n}\}}$$

(here we take N to be the bound on f; then this is because if $k \cdot 2^n$ goes above N, this set is empty).

These sets $\{k \cdot 2^{-n} \le f < (k+1) \cdot 2^{-n}\}$ are measurable — this is because f is measurable, and this set can be written in terms of inverse images of intervals under f, explicitly as

$$f^{-1}([k \cdot 2^{-n}, \infty)) \setminus f^{-1}([(k+1) \cdot 2^{-n}, \infty)).$$

And by what we've proven, this means their indicator functions are in $\mathcal{V}$. Then f_n is a finite linear combination of such sets, so since $\mathcal{V}$ is a vector space, this means each f_n is in $\mathcal{V}$. And $f_n \uparrow f$ pointwise, so by the third property it follows that $f \in \mathcal{V}$.

Now we've shown that the claim is true for bounded *nonnegative* linear functions, by approximating them by finite linear combinations of indicator functions of measurable sets.

Finally, if f is bounded and measurable (but not necessarily nonnegative), then we can write $f = \max\{0, f\} - \max\{-f, 0\}$. So f is the difference of two nonnegative bounded measurable functions. And both are in $\mathcal{V}$ by the previous step, so their difference is also in $\mathcal{V}$.

So we've finished the proof — we showed that an arbitrary bounded measurable function is in $\mathcal{V}$. □

Unfortunately, we need to wait a few lectures to see applications of this theorem; we'll see it when discussing Lebesgue integration.

§4.6 Constructing new measures

Now we'll see ways to construct new measures — starting with a measurable function and some measure. The typical measure we construct is called the *image measure*, and we're going to discus it extensively when discussing integration.

Definition 4.15 (Image measure)**.** Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measure spaces, and suppose that μ is a measure on $\mathcal{E}$ and f is a measurable function $E \to G$. Then we define the image measure ν on G by

$$\nu(A) = \mu(f^{-1}(A))$$

for every $A \in \mathcal{G}$.

This is a well-defined measure — f is a measurable function, so the inverse image of any measurable set is measurable, which means its measure with respect to μ is well-defined. This is a quite standard way to define measures, but it'll be important in integration (specifically, when we integrate the composition of two functions).

This is a simple way to construct a measure, but there are more complicated ones as well. Before we do this, we'll talk about functions with certain properties.

§4.6.1 Generalized inverses

From real analysis, we know that if we have a strictly increasing and continuous function f (meaning that $f(x) < f(y)$ for all $x < y$), then we can invert f. But you can define other notions of inverses for other functions.

You can relax this condition by saying that f is *nondecreasing*, meaning that $f(x) \le f(y)$ for all $x < y$. We also assume that f is *right-continuous* — meaning that for any sequence $x_n \downarrow x$, we have $f(x_n) \downarrow f(x)$ as $n \to \infty$. (So whenever we approximate x *from the right*, $f(x_n) \to f(x)$. This doesn't imply continuity — for that we'd also need f to be *left-continuous*. So this is a relaxation of continuity.)

The point is that for functions with these two (more relaxed) properties, we still have a notion of inverses.

Lemma 4.16

Let $g\colon \mathbb{R} \to \mathbb{R}$ be nonconstant, nondecreasing, and right-continuous. Then there is a nondecreasing *left-continuous* function $f\colon I \to \mathbb{R}$ such that for all $x \in I$ and $y \in \mathbb{R}$, we have

$$x \le g(y) \text{ if and only if } f(x) \le y.$$

We will call f the *inverse* of g. This is a generalization of the classical inverse — if you take a function which is strictly increasing and continuous, then g is just the inverse of f. Here we relax these conditions; then you still have a notion of inverse, but not quite the same one; instead we have a notion with this inequality.

In fact, you can construct this inverse *explicitly* — we define

$$f(x) = \inf\{y \in \mathbb{R} \mid x \le g(y)\}.$$

We need to prove that this function has the required properties.

Proof. We define our interval I as the open interval with endpoints given by $\lim_{x\to-\infty} g(x)$ and $\lim_{x\to\infty} g(x)$. These two limits exist because g is nondecreasing (this follows from real analysis).

For $x \in I$, we consider the set

$$J_x = \{y \in I \mid x \le g(y)\}.$$

Then since g is nondecreasing, if we have some $y \in J_x$, then for *every* $y' \geq y$, we also have $y' \in J_x$ (because $g(y') \geq g(y) \geq x$). Since g is right-continuous, we have that if we take any sequence $(y_n) \subseteq J_x$ such that $y_n \downarrow y$, then we also have $y \in J_x$ — because $g(y_n) \geq x$ and $g(y_n) \to g(y)$ (by right-continuity), so $g(y) \geq x$. This means

$$J_x = [f(x), \infty)$$

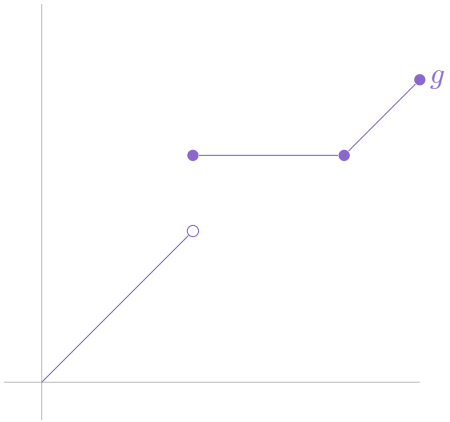
(by the definition of $f(x)$). This means for $y \in \mathbb{R}$, we have $x \leq g(y)$ if and only if $f(x) \leq y$.

Now we need to prove the remaining properties of f (that it's nondecreasing and left-continuous). First, if we take any $x \leq x'$, then we have $J_x \subseteq J_{x'}$. Then since $J_x = [f(x), \infty)$, this means $f(x) \leq f(x')$; so f is nondecreasing.

Finally, if we take any sequence $x_n \uparrow x$, then we have $J_x = \bigcap_n J_{x_n}$; this implies $f(x_n) \to f(x)$ (by the above formula $J_x = [f(x), \infty)$), which means $f$is left-continuous. □

From now on, we're going to interpret f as the generalized inverse of g.

We'll finish by representing f and g graphically.



§5 September 19, 2024

Last lecture, we stated and proved the monotone class theorem. It's very important, but unfortunately we can't see its importance yet; it'll be important later when we construct the Lebesgue integral.

§5.1 From right-continuous functions to Radon measures

We said one of our goals in measure theory is to construct new measures. One obvious measure we can construct is the image measure (as defined last class). Then we said we're looking for more elaborate ways of constructing measures. That's why we needed the generalized notion of the inverse of a function — we started with a non-decreasing right-continuous function, and we said that for such a function, we can always define a notion of inverse. We did this to construct new measures on $\mathbb{R}$. We've earlier mentioned *Radon measures*, and the Lebesgue measure is one of them; it'll turn out that we can recover *all* of them using nondecreasing right-continuous functions.

Theorem 5.1

Let $g: \mathbb{R} \to \mathbb{R}$ be nonconstant, nondecreasing, and right-continuous. Then there exists a unique Radon measure dg on $\mathcal{B}(\mathbb{R})$ such that for every half-closed interval $(a, b]$, we have

$$dg((a, b]) = g(b) - g(a).$$

Moreover, we can obtain all nonzero Radon measures on $\mathbb{R}$ in this way.

So if we start with such a function, we can always produce a Radon measure (a Borel measure on $\mathbb{R}$ with certain properties). But the point is we can do something stronger. In particular, we can recover *every* Radon measure in this way. So this gives a mechanism for producing all the nonzero Radon measures on $\mathbb{R}$ — they're in corresopndence with nondecreasing right-continuous functions.

We've already seen an example of such a construction.

Example 5.2

If we take g to be the identity function, then we get the Lebesgue measure.

For the proof, we've already done most of the work in the construction of the Lebesgue measure.

Proof. First, recall that we can define an inverse $f: I \to \mathbb{R}$ of g (in the generalized form discussed in the previous lecture), where $I = (g(-\infty), g(\infty))$ (where by $g(-\infty)$ we mean $\lim_{x \to -\infty} g(x)$, and likewise $g(\infty) = \lim_{x \to \infty} g(x)$ — both limits exist but may be infinite, because g is nondecreasing). Recall that f is left-continuous and nondecreasing (as proved last lecture).

Now we can just define the measure we're looking for to be the pushforward of the Lebesgue measure with respect to f — i.e., $dg = \mu \circ f^{-1}$. Now we'll verify that this pushforward measure indeed satisfies the desired property. In particular, we have

$$dg((a, b]) = \mu(\{x \in I \mid a < f(x) \leq b\})$$

by the definition of the pushforward measure. But by the definition of f, this is also

$$\mu(\{x \in I \mid g(a) < x \leq g(b)\})$$

(this was the definition of the function f). But this is just $\mu((g(a), g(b)])$, which by the definition of the Lebesgue measure is just $g(b) - g(a)$. And this is true for all a and b.

(It's obvious that this measure is a Radon measure because it's a pushforward of the Lebesgue measure with respect to a left-continuous measure.)

Note that there are no other Radon measures ν satisfying $\nu((a, b]) = g(b) - g(a)$, by the same argument used to prove the uniqueness of the Lebesgue measure (we just replace the identity function with the function g). So the function g *uniquely* defines the Radon measure.

Now we need to show the other direction — that given a Radon measure, we can find a function g which produces it in this way. We're going to explicitly construct g. Let ν be a Radon measure on $\mathcal{B}(\mathbb{R})$. Now we want to construct a function g such that $dg = \nu$. For this, we set

$$g(y) = \begin{cases} -\nu((y, 0]) & y \leq 0 \\ \nu((0, y]) & y > 0. \end{cases}$$

Then by the definition of g, we have $\nu((a, b]) = g(b) - g(a)$ (that's why we defined g in this way — so that this condition holds). Also, ν is nonzero, so the function g is not constant. Then we have $\nu = dg$, because they agree on a π-system generating the Borel σ-algebra (finite unions of such intervals). □

So we have characterized all Radon measures on $\mathbb{R}$ — they're of the form dg for some nonconstant nondecreasing right-continuous g .This is a generalization of the construction of the Lebesgue measure, where we took g to be the identity.

Student Question. *How does the uniqueness of the Lebesgue measure generalize to this?*

Answer. We obtained two characterizations of the Lebesgue measure. The first characterization was the unique Borel measure with $\mu((a,b]) = b - a$, and the second was the unique Borel measure which is translation-invariant. Here we're referring to the proof of uniqueness from the first characterization.

Student Question. *Why do we need ν to be Radon?*

Answer. This is because we're defining g using ν, so we need ν to have finite values — one of the conditions on Radon measures is that the measure of any compact set is finite. Here we need $\nu((0,y])$ to be finite — otherwise g wouldn't be well-defined. (We also need it for the uniqueness argument.)

Student Question. *Why is g right-continuous?*

Answer. For example, we can write $(a,b] = \bigcap_n (a, b + \frac{1}{n}]$, which means $\nu((a,b]) = \lim_{n\to\infty} \nu((a, b+\frac{1}{n}])$. And this gives

$$g(b) - g(a) = \lim_{n\to\infty} g\left(b + \frac{1}{n}\right) - g(a).$$

§5.2 Random variables and distribution functions

We'll now look at these ideas in the context of probability. First we'll rigorously define random variables.

Definition 5.3 (Random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(E, \mathcal{E})$ be a measurable space. Then an E-valued random variable is a measurable function $X : \Omega \to E$.

So a random variable is a measurable function defined on a probability space. Usually we'll be interested in the case where $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

When we have a random variable X, we might ask questions like, what's the probability X is in a set A? In other words, we're asking for the size of the things sent to A. This is also called the *distribution* or *low* of our random variable. And it's just given by the image (or pushforward) measure.

Definition 5.4 (Distribution)**.** Given a random variable $X : \Omega \to E$, the distribution (or low) of X is defined as the image measure

$$\mu_X = \mathbb{P} \circ X^{-1}.$$

We usually write

$$\mathbb{P}[X \in A] = \mu_X(A) = \mathbb{P}[X^{-1}(A)].$$

In the special case where $E = \mathbb{R}$ (so X is a real-valued random variable), μ_X is determined by its values on the π-system

$$\{(-\infty, y] \mid y \in \mathbb{R}\}.$$

So the distribution of a random variable is determined by the values of its low on this π-system (since π-systems determine measures, as we've seen).

Definition 5.5. For a $\mathbb{R}$-valued random variable, we define the distribution function of X as

$$F_x(x) = \mu_X((-\infty, x]) = \mathbb{P}[X \le x].$$

So the distribution function of X determines the low of X (because μ_X is determined by its values on this π-system).

We'll now state some properties of such distribution functions. (We'll state this as a proposition, but Prof. Kavvadias encourages us to do it as an exercise.)

Proposition 5.6

The distribution function F_X of a random variable X satisfies the following properties:

- We have $F_X(x) \to 0$ as $x \to -\infty$, and $F_X(x) \to 1$ as $x \to \infty$.
- F_X is nondecreasing and right-continuous.

So a distribution of a random variable on $\mathbb{R}$ is a nondecreasing right-continuous function with limits 0 and 1 as $x \to -\infty$ and $x \to \infty$. And we call *any* function with these properties a *distribution function*.

Definition 5.7 (Distribution function)**.** A distribution function is a nondecreasing right-continuous function $f : \mathbb{R} \to [0, 1]$ with the properties that $f(x) \to 0$ as $x \to -\infty$ and $f(x) \to 1$ as $x \to \infty$.

Then a natural question arises. We proved the distribution function of a random variable is a distribution function (in the sense it satisfies these properties).

Question 5.8. Is *every* distribution function the distribution function of some random variable — i.e., given f, can we always find a random variable X with $f = F_X$?

The answer is yes, and the proof is elementary, though not necessarily easy to come up with.

Proposition 5.9

Let F be any distribution function. Then there exists some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X such that $F_X = F$.

So we can recover any distribution function as the distribution function of some random variable. The proof is constructive — we start with any distribution function F, and we need to construct our probability space and a random variable defined on it. This might seem abstract, but the proof is quite simple.

Proof. We need a set, a σ-algebra on that set, and a probability measure. What's the simplest set to think of? The unit interval. And the most natural σ-algebra on the unit interval is the Borel measure, and the most natural measure is the Lebesgue measure (restricted to that interval). So we simply take $(\Omega, \mathcal{F}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]), \mu)$ (note that this is a probability space because $\mu([0, 1]) = 1$).

Then we take $X : \Omega \to \mathbb{R}$ to be given by

$$X(\omega) = \inf\{x \in \mathbb{R} \mid \omega \leq F(x)\}$$

for all $\omega \in [0, 1]$. This is a well-defined function, and it's easy to verify that it's a measurable function. The important thing is that we have $X(\omega) \leq x$ if and only if $\omega \leq F(x)$. This means

$$F_X(x) = \mathbb{P}[X \leq x] = \mathbb{P}[(0, f(x)]].$$

But the Lebesgue measure of an interval is just its length; and here the length is $f(x)$, so we are done. (This is true for all x.) □

So any distribution function is the distribution function of some random variable, and we can explicitly construct that random variable. The proof is just a few lines, but this construction is very useful in practice — this procedure is used to sample random variables using computer programs, for example.

Student Question. *Where in the proof are we using the fact that we're using the Borel σ-algebra?*

Answer. We need this to justify that X is a measurable function. We could also take a σ-algebra larger than the Borel one; but the Borel one is just a natural choice.

§5.3 Independence of random variables

Now we'll talk about independence of random variables. We've already talked implicitly about this when talking about independence between σ-algebras, but now we'll write down explicitly what we mean.

Definition 5.10 (Independent random variables)**.** A family (X_n) of random variables is said to be independent if the family of σ-algebras generated by these random variables — i.e., $(\sigma(X_n))$ — is independent.

(We've already discussed what we mean by a sequence of independent σ-algebras; so this is a fairly expected definition.)

Now let's see some characterizations of independence. In general, we'll deal with independence a lot in this course, so this will be very useful.

Proposition 5.11

Two $\mathbb{R}$-valued random variables X and Y are independent if and only if

$$\mathbb{P}[X \le x,\, Y \le y] = \mathbb{P}[X \le x]\mathbb{P}[Y \le y]$$

for all $x, y \in \mathbb{R}$. More generally, if (X_n) is a sequence of $\mathbb{R}$-valued random variables, then they are independent if and only if

$$\mathbb{P}[X_1 \le x_1,\, X_2 \le x_2,\, \ldots, X_n \le x_n] = \prod_{j=1}^{n} \mathbb{P}[X_j \le x_j]$$

for every $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathbb{R}$.

We'll only prove the first statement, since the essence of the proof is there (we can use induction to prove the second statement). The proof is not hard, but the result is very important when checking independence.

Proof. The backwards direction (where we assume independence) is obvious from the definition of independence. Now we'll show the other direction. For this, consider the family of sets

$$\{(-\infty, x] \mid x \in \mathbb{R}\}.$$

As mentioned earlier, this is a π-system generating the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. By our assumption, the σ-algebras generated by X and Y are independent when restricted to the special sets

$$\{\omega \in \Omega \mid X(\omega) \le x\}$$

and likewise with Y. But these are π-systems generating our σ-algebras. And we proved that if two σ-algebras are independent on a π-system, then they're independent everywhere (we proved this last lecture). □

Remark 5.12. Here we have two functions X and Y, and we want to say that the σ-algebra generated by X is independent from the one generated by Y. By definition, $\sigma(X)$ is independent of $\sigma(Y)$ if and only if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

for every $A \in \sigma(X)$ and $B \in \sigma(Y)$. (This is by definition; and that's what we want to show.) But we've shown we don't have to show this equality for *every* A and B in the two σ-algebras — if we can find two π-systems and prove this for A and B from those π-systems, then this is enough. And the π-system for $\sigma(X)$ given by

$$\{\omega \in \Omega \mid X(\omega) \leq x\}$$

(this generates $\sigma(X)$ because $(-\infty, x]$ is a π-system generating $\mathcal{B}(\mathbb{R})$). We can do the same for Y, and the given assumption gives independence on these sets; so independence follows.

§5.4 Independent and identically distributed sequences

Lots of probability involves independent and identically distributed random variables. But it's not actually clear how to construct such sequences of random variables. We're now going to do this rigorously. We'll first consider the Bernoulli random variable with parameter $\frac{1}{2}$ (which is in some sense the 'simplest' random variable — taking 0 with probability $\frac{1}{2}$ and 1 with probability $\frac{1}{2}$).

Proposition 5.13

Let $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}(0,1), \mu)$ (where μ is the Lebesgue measure). Then there exists a sequence of independent random variables $\text{Ber}(\frac{1}{2})$.

We're going to construct this explicitly, and for that, we're going to use the binary expansion of a number in $(0,1)$.

Proof. Suppose we fix $\omega \in (0,1)$. Then we can expand ω as

$$\omega = \sum_{n \geq 1} \omega_n \cdot 2^{-n},$$

where $\omega_n \in \{0,1\}$. It's a theorem (from analysis or number theory) that every $\omega \in (0,1)$ can be written in this way, which we'll use. And this expansion is unique — if we take another ω' and consider the sequence (ω'_n), it'll be different.

Now that we have this, we define

$$R_n(\omega) = \omega_n.$$

So for every ω, we just look at the nth coefficient of its binary expansion.

Claim 5.14 — The function R_n is measurable.

Proof. First, we can write $R_1(\omega) = 1_{(1/2,1]}(\omega)$ by definition. And we've seen that indicator functions of measurable sets are measurable, so R_1 is measurable.

Similarly, we can write R_2 as

$$R_2(\omega) = 1_{(1/4,1/2]}(\omega) + 1_{(3/4,1]}(\omega).$$

This is still a measurable function, because it's a sum of two measurable functions, and we saw such a sum is also measurable.

In general, to prove that R_n is measurable, we can express

$$R_n(\omega) = \sum_{j=1}^{2^{n-1}} 1_{(2^{-n}\cdot(2j-1),2^{-n}\cdot 2j]}(\omega).$$

(You can prove this using induction on n.) So R_n is a sum of measurable functions (indicators of measurable sets), and is therefore measurable. (The hard part is coming up with this expansion.) □

Claim 5.15 — Each R_n is a Bernoulli variable with parameter $\frac{1}{2}$.

Proof. Using our formula for R_n above, we have

$$\mathbb{P}[R_n = 1] = \sum_{j=1}^{2^{n-1}} 2^{-n}((2j) - (2j-1)) = \sum_{j=1}^{2^{n-1}} 2^{-n} = \frac{1}{2}.$$

This also means $\mathbb{P}[R_n = 0] = 1 - \mathbb{P}[R_n = 1] = \frac{1}{2}$. So R_n is a Bernoulli random variable with parameter $\frac{1}{2}$, as desired. □

So these functions are measurable and have the correct low. The last part is to check independence. For this, if $n \neq m$, it's easy to check (using the formula for R_n) that

$$\mathbb{P}[R_n = R_m = 0] = \frac{1}{4} = \mathbb{P}[R_n = 0]\mathbb{P}[R_m = 0],$$

and likewise with some of the 0's replaced by 1's. So we've constructed a probability space and a sequence of random variables which are independent, and which all have the low of $\text{BER}(\frac{1}{2})$.

(Here we only considered two of the random variables, but you actually have to do this for an arbitrary finite collection; but you can do that in the same way; then $\frac{1}{4}$ gets replaced by $\frac{1}{2^n}$.) □

So we can construct a space of independent random variables which have the distribution function of a Bernoulli random variable.

Question 5.16. Given an *arbitrary collection* of fixed distribution functions, can we find a collection of independent variables with those distribution functions?

The answer is yes. (This is a generalization of the above result, and surprisingly, the proof is easier once we have the above construction.)

Proposition 5.17

Let $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}((0,1)), \text{Lebesgue})$. Given any sequence (F_n) of distribution functions, there exists a sequence (X_n) of independent random variables with the property that $F_{X_n} = F_n$ for all n.

In the previous result, all the F_n's were the same, and were the distribution of a Bernoulli random variable with parameter $\frac{1}{2}$. This is a generalization — given an arbitrary sequence of distribution functions, we can produce a sequence of independent random variables with those distributions.

Proof. Somehow we need to generate the appropriate random variables, and we're going to somehow use the Bernoullis from earlier. So we first let (R_j) be as above — so (R_j) is a sequence of independent random variables, each with the low of a Bernoulli random variable. The goal is to relate these somehow. For this,

we consider some bijection $m: \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ (it's known from set theory that there is such a bijection, because they have the same cardinality; a *bijection* is a function which is injective and surjective). Then we set

$$Y_{(k,n)} = R_{m(k,n)}.$$

So we take the sequence of Bernoullis constructed earlier, and we parametrize it with respect to $\mathbb{N} \times \mathbb{N}$ rather than $\mathbb{N}$.

Now we define the random variable $Y_n = \sum_{k \geq 1} Y_{(k,n)} 2^{-k}$. This is a well-defined random variable. And we know that the Y_n's are independent (because (R_j) is independent, and if we take any two indices n, then all the indices of the $Y_{(k,n)}$'s in the sums differ). So (Y_n) is an independent sequence of random variables. And each is *uniform* — i.e., it has the uniform low on $(0, 1)$ (which means its distribution is just the Lebesgue measure on $(0, 1)$).

Then we do a similar trick as from a couple of propositions before — we consider the function

$$G_n(y) = \inf\{x \mid y \leq F_n(x)\},$$

and we set $X_n = G_n(Y_n)$. Then by the proof that every distribution function is the distribution of a random variable, we have that (X_n) is a sequence of random variables with the required property $F_{X_n} = F_n$. □

So this proof combines the last two results mentioned in class.

Student Question. *Why is Y_n measurable?*

Answer. This is because it's the pointwise limit of measurable functions — if we fix n, then $Y_n = \lim_{m \to \infty} \sum_{k=1}^{m} Y_{(k,m)} \cdot 2^{-k}$. And each of these finite sums is measurable, and the limit of measurable functions is a measurable function (this follows from the proposition from the previous lecture, with lim sup).

Student Question. *Why are the Y_n's independent?*

Answer. This is because the $Y_{(k,n)}$ are independent, and the indices in Y_m and Y_n are disjoint — we always have $(k_1, m) \neq (k_2, n)$, because the second coordinates are always different. So then $Y_{(k_1,m)}$ and $Y_{(k_2,n)}$ will correspond to different R_j's, and we know the R_j's are independent.

Student Question. *Could we also say the Y_n's are measurable by the monotone class sequence?*

Answer. Yes (since we have an increasing sequence).

Student Question. *This construction only lets us construct a* countable *family; can we construct an uncountable one?*

Answer. Yes, but then you have to use some more complicated results from set theory, and you have to be careful (first you have to change the definition of independence; and the actual construction is much more complicated, since here we're summing and you have to be careful what you mean if you sum over uncountable sets).

We'll now highlight a result here which is connected to measure theory using probability theory. Suppose we have the same setup, with the R_j's being Bernoullis with parameter $\frac{1}{2}$. If we fix $\varepsilon > 0$, we'll learn later that the *weak law of large numbers* (which we are going to talk about later — these limiting theorems are one of the most important theorems in science, and appear everywhere) says that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{j=1}^{n} R_j - \frac{1}{2}\right| \geq \varepsilon\right] \to 0$$

as $n \to \infty$ (i.e., the probability that the average of the R_j's is ε away from $\frac{1}{2}$ tends to 0 as $n \to \infty$). We also have another result, the *strong law of large numbers* (which we'll also prove later in the course), which is stronger than this one — it implies that

$$\mathbb{P}\left[\omega \in (0,1) \mid \frac{1}{n}\sum_{j=1}^{n} R_j \to \frac{1}{2}\right] = 1.$$

In words, almost every number in the unit interval has an equal number of 0's and 1's in its binary expansion. This is called the *normal number theorem*; it's a theorem in number theory, and it can be shown using the strong law of large numbers.

§6 September 24, 2024

In the previous lecture, we showed how to rigorously construct a sequence of random variables which are independent and all have some fixed distribution. We constructed both a probability space and random variables defined there.

The next thing we'll consider is the convergence of measurable functions. In analysis, you don't just have pointwise convergence, but also uniform convergence and various other notions. Here we'll have analogous notions, but this time involving some sort of measure.

§6.1 Convergence almost everywhere

The first notion is that of convergence almost everywhere; we've already mentioned the notion of convergence *Lebesgue*-almost everywhere, and now we'll generalize this.

Definition 6.1 (Convergence almost everywhere). Let $(E, \mathcal{E}, \mu)$ be a measure space. Suppose we have a sequence of measurable functions (f_n) and a fixed function f. Then we say that f_n converges to f almost everywhere as $n \to \infty$, written $f_n \xrightarrow{ae} f$, if

$$\mu(\{x \in E \mid f_n(x) \not\to f(x) \text{ as } n \to \infty\}) = 0.$$

If $(E, \mathcal{E}, \mu)$ is actually a probability space, then this is also called almost sure convergence.

In other words, f_n converges to f pointwise outside of a set which is negligible (where 'negligible' means that its measure is 0 under μ).

It's not obvious *a priori* that this set $\{x \in E \mid f_n(x) \not\to f(x) \text{ as } n \to \infty\}$ is even measurable, but this is not hard to see. To see this, we can write this as

$$\{x \in E \mid \limsup_{n\to\infty} |f_n(x) - f(x)| > 0\}.$$

And we know the function $\limsup_{n\to\infty} |f_n(x) - f(x)|$ is a measurable function. In other words, we can write this as $\{x \in E \mid g(x) > 0\}$, where

$$g(x) = \limsup_{n\to\infty} |f_n(x) - f(x)|,$$

and we have shown that g is a measurable function. And this is by definition $g^{-1}((0,\infty))$, which is measurable because g is measurable. So this definition makes sense, since the set we're trying to take a measure of is indeed measurable.

§6.2 Convergence in measure

The next notion of convergence we have is convergence in measure, which is slightly weaker.

Definition 6.2 (Convergence in measure). Let $(E, \mathcal{E}, \mu)$ be a measure space, and let (f_n) be a sequence of measurable functions and f be a measurable function. Then we say that f_n converges to f in measure as $n \to \infty$ if for all $\varepsilon > 0$, we have

$$\mu(\{x \in E \mid |f_n(x) - f(x)| \geq \varepsilon\}) \to 0$$

as $n \to \infty$. If $(E, \mathcal{E}, \mu)$ is a probability space, this is also called converges in probability.

So if we fix ε and consider the set of points x in our space for which $|f_n(x) - f(x)| \geq \varepsilon$, this set becomes smaller and smaller as $n \to \infty$.

§6.3 Relationships between types of convergence

Someone might wonder how these notions are related.

Question 6.3. Does convergence in measure imply almost everywhere convergence or vice versa?

We'll show that these two results *are* related, at least under certain assumptions on our measure space.

Theorem 6.4

(i) If $\mu(E) < \infty$ and $f_n \to f$ almost everywhere, then $f_n \to f$ in measure.

(ii) For any E, if $f_n \to f$ in measure, then there exists a subsequence (f_{n_k}) such that $f_{n_k} \to f$ almost everywhere (as $k \to \infty$).

So if the measure space is finite, then almost everywhere convergence is stronger than convergence in measure. But if the space is infinite, this is not the case anymore; we're going to see a counterexample.

On the other hand, if $f_n \to f$ in measure, we don't necessarily get that $f_n \to f$ almost everywhere; but we *do* get a *subsequence* for which this is true. So we don't actually have convergence almost everywhere of the *entire* sequence, but we do have it along a *subsequence*.

The proof is not too complicated, but it's a nice result.

Proof of (i). Fix $\varepsilon > 0$, and set

$$A_n = \{x \in E \mid |f_n(x) - f(x)| \leq \varepsilon\}.$$

It's a fact (which is on the first problem set) that we have

$$\liminf_{n\to\infty} \mu(A_n) \geq \mu(\liminf_{n\to\infty} A_n)$$

(where the right-hand side is the lim inf in the sense of sets). This means

$$\liminf_{n\to\infty} \mu(\{x \in E \mid |f_n(x) - f(x)| \leq \varepsilon\}) \geq \mu(\{x \in E \mid |f_m(x) - f(x)| \leq \varepsilon \text{ for all sufficiently large } m\})$$

(by the definition of the lim inf of sets). But this is at least

$$\mu(\{x \in E \mid f_m(x) \to f(x)\})$$

(by the definition of convergence). And we assumed almost everywhere convergence, so this is simply $\mu(E)$. This implies (reversing the order) that

$$\begin{aligned}\limsup_{n\to\infty}\mu(\{x\in E\mid |f_n(x)-f(x)|>\varepsilon\}) &= \limsup_{n\to\infty}(\mu(E)-\mu(\{x\in E\mid |f_n(x)-f(x)|\le\varepsilon\}))\\ &= \mu(E)-\liminf_{n\to\infty}\mu(\{x\in E\mid |f_n(x)-f(x)|\le\varepsilon\})\\ &\le 0\end{aligned}$$

(by what we've shown). And the lim sup of a nonnegative seequence is of course nonnegative, so it has to be 0. This implies

$$0\le\liminf_{n\to\infty}\mu(\{x\in E\mid |f_n(x)-f(x)|>\varepsilon\})\le\limsup_{n\to\infty}\mu(\{x\in E\mid |f_n(x)-f(x)|>\varepsilon\})\le 0,$$

which means both must be 0. And this means the sequence has a limit, and the limit is 0; so $\mu(\{x\in E\mid |f_n(x)-f(x)|>\varepsilon\})\to 0$, as desired. □

Note that we assumed the measure of the space is finite for the finiteness of

$$\mu(E)-\liminf_{n\to\infty}\mu(\{x\in E\mid |f_n(x)-f(x)|\le\varepsilon\}).$$

Proof. Here we need to choose a subsequence. We choose our subsequence (f_{n_k}) such that

$$\mu\left(\left\{x\in E\mid |f_{n_k}(x)-f(x)|>\frac{1}{k}\right\}\right)\le 2^{-k}.$$

(We can do this because for every fixed k, we know the left-hand side tends to 0 as $n\to\infty$; so we can pick n_k sufficiently large to make the right-hand side arbitrarily small.)

And $\sum 2^{-k}$ is summable, so we get

$$\sum_{k\ge 1}\mu\left(\left\{x\in E\mid |f_{n_k}(x)-f(x)|>\frac{1}{k}\right\}\right)\le\sum_{k\ge 1}2^{-k}=1.$$

In particular, this is finite, so we can apply the first Borel–Cantelli lemma (which is true in any space, not just a probability space); this tells us that

$$\mu\left(\left\{x\in E\mid |f_{n_k}(x)-f(x)|>\frac{1}{k}\text{ infinitely often}\right\}\right)=0.$$

But this implies that $f_{n_k}\to f$ almost everywhere as $k\to\infty$. □

Now we'll see a counterexample where (i) doesn't hold for a measure space with $\mu(E)=\infty$.

Example 6.5

Consider $(E,\mathcal{E},\mu)=(\mathbb{R},\mathcal{B}(\mathbb{R}),\text{Lebesgue})$, and take $f_n(x)=1_{[n,\infty)}$ (for each $n\in\mathbb{N}$). Then clearly $f_n\to 0$ almost everywhere (for every fixed x, we can find n for which $n>x$, so $f_n(x)$ becomes 0). But for every n, we have

$$\mu\left(\left\{x\in\mathbb{R}\mid |f_n(x)|>\frac{1}{2}\right\}\right)=\mu([n,\infty))=\infty.$$

This means f_n does *not* converge to 0 in measure.

(Of course a counterexample has to have infinite measure, because if the space has finite measure, then our theorem implies we have convergence in measure.)

Remark 6.6. This is one of those easy counterexamples in measure theory that's actually quite frequent.

§6.4 Convergence in distribution

So far we've mentioned two notions of convergence that hold in arbitrary measure spaces; we'll now talk about notions of convergence that hold in just probability spaces. Specifically, we'll talk about convergence in distribution, which is very important and commonly used.

Definition 6.7 (Convergence in distribution). Let X_n and X be random variables with distribution functions F_{X_n} and F_X. Then we say that $X_n \to X$ in distribution as $n \to \infty$ if $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathbb{R}$ at which F_X is continuous.

So we require pointwise convergence of the distribution functions, but not for all $x \in \mathbb{R}$ — we only require it for those points x at which F_X is continuous.

Why do we impose this notion of convergence? The idea is that if the resulting distribution has a jump at X, it shouldn't matter which side of the jump X_n is at.

Another comment is that we're just dealing with distribution functions; our random variables are not necessarily defined on the same probability space.

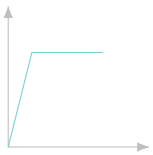
Now we'll see an example that illustrates convergence in distribution without having pointwise convergence everywhere.

Example 6.8

Let X_n is the uniform random variable on $[0, \frac{1}{n}]$. Then $X_n \to 0$ in distribution.

This intuitively makes sense, but to prove it rigorously, we can explicitly compute the distribution function of X_n — we have

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x \le 0 \\ nx & \text{if } 0 \le x \le \dfrac{1}{n} \\ 1 & \text{if } x \ge \dfrac{1}{n}. \end{cases}$$



Meahwhile, the distribution function of 0 is of course

$$F_0(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \ge 0. \end{cases}$$

So $F_{X_n}(x) \to F_0(x)$ at all $x \neq 0$. However, we do *not* have pointwise convergence at $x = 0$, because $F_{X_n}(0) = 0$ and $F_0(0) = 1$. Still, this is fine according to our definition of convergence in distribution; and we can still make sense of X_n being close to the 0 random variable.

These are more or less the main notions of convergence that we're going to deal with in this course — convergence almost everywhere, convergence in measure, and convergence in distribution.

§6.5 Tail events

Before rigorously constructing the Lebesgue integral, we'll talk about tail events and tail σ-algebras.

Very intuitively, tail events are events that depend only on the asymptotic behavior of a sequence of random variables. We're going to see examples of such events soon. But before we do this, we'll explain what we mean by the tail σ-algebra, the σ-algebra generated by those events. This is very important because it lets us study the asymptotic behavior of a sequence of random variables (which will be useful in proving things like the central limit theorem and the strong law of large numbers and so on).

Definition 6.9 (Tail σ-algebra). Let (X_n) be a sequence of random variables. Then we let $\mathcal{T}_n$ be the σ-algebra

$$\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, X_{n+3}, \ldots),$$

and we define $\mathcal{T} = \bigcap_n \mathcal{T}_n$. We call $\mathcal{T}$ the tail σ-algebra.

As we can see, $\mathcal{T}$-measurable events are exactly those which only depend on the asymptotic behavior of the X_n's — i.e., they don't depend on any finite collection.

Example 6.10

Consider the random variables $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i$ and $\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i$ (the lim sup and lim inf of the average of the first n terms). These random variables are $\mathcal{T}$-measurable — this is because if we fix any finite collection of X_a's, the lim sup and lim inf are independent of them.

Then knowing this, you can construct many events that belong to the tail σ-algebras.

Example 6.11

The event $\{\lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^n X_i) \text{ exists}\}$ can be written as

$$\left\{\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i\right\},$$

so it is in $\mathcal{T}$ (because both random variables here are $\mathcal{T}$-measurable).

We will be interested in this event, and we'll later show that its probability is 1 (under some conditions).

§6.5.1 Kolgomorov's 0–1 law

We'll now see our second main theorem for this lecture.

Theorem 6.12 (Kolgomorov's 0–1 law)

Let (X_n) be a sequence of independent real-valued random variables, and let $\mathcal{T}$ be their tail σ-algebra.

(i) Then any event $A \in \mathcal{T}$ has probability either 0 or 1 — i.e., $\mathbb{P}[A] = 0$ or $\mathbb{P}[A] = 1$.

(ii) If X is a $\mathcal{T}$-measuable random variable, then it is constant almost everywhere — there exists $c \in \mathbb{R}$ such that $\mathbb{P}[X = c] = 1$.

So events in the tail σ-algebra have probability either 0 or 1; we're going to use this many times in the course to prove several limiting theorems.

Proof. We're going to show that if we take any $A \in \mathcal{T}$, then A is independent of itself, i.e.,

$$\mathbb{P}[A \cap A] = \mathbb{P}[A] \cdot \mathbb{P}[A].$$

Of course $A \cap A = A$, so this immediately means $\mathbb{P}[A] \in \{0, 1\}$. (The first time when you see the statement 'A is independent of itself' it seems weird, but it does make sense in this way.)

In fact, we are going to show that the σ-algebra $\mathcal{T}$ is independent of itself (this immediately implies the first claim).

Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ be the σ-algebra generated by the first n random variables. Then we've seen earlier that $\mathcal{F}_n$ is generated by the π-system of events of the form

$$A = \{X_1 \leq x_1,\, X_2 \leq x_2,\, \ldots,\, X_n \leq x_n\}$$

for $x_1, \ldots, x_n \in \mathbb{R}$ — the collection of sets A of this form is a π-system which generates $\mathcal{F}_n$ (we showed this in a previous lecture).

Similarly, we consider the σ-algebra $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, X_{n+3}, \ldots)$; this is generated by sets of the form

$$B = \{X_{n+1} \leq x_{n+1},\, X_{n+2} \leq x_{n+2},\, \ldots,\, X_{n+k} \leq x_{n+k}\}$$

for every $k \in \mathbb{N}$ and $x_{n+1}, \ldots, x_{n+k} \in \mathbb{R}$. This is again another π-system which generates $\mathcal{T}_n$.

The first step is to show that $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent. This is intuitively clear because if we take any sets A and B of this form, we have

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B].$$

Why? Here we're taking the first n indices and then the next k indices of the X_j's; and since the X_j's are independent by assumption, that means these events are independent.

So far, we have considered the π-systems generating our σ-algebras, and we have shown that they are independent; this means $\mathcal{F}_n$ is independent of $\mathcal{T}_n$ (since we've shown that if you have independence between two π-systems generating your σ-algebras, then you have independence between the entire σ-algebras).

This also implies that $\mathcal{F}_n$ is independent of $\mathcal{T}$ for all n, because $\mathcal{T} = \bigcap_n \mathcal{T}_n$.

But now if we consider $\bigcup_n \mathcal{F}_n$, this is a π-system which generates $\mathcal{F}_\infty = \sigma(X_1, X_2, \ldots)$. And we have shown $\mathcal{T}$ is independent of our π-system generating $\mathcal{F}_\infty$, which implies that $\mathcal{T}$ is independent of $\mathcal{F}_\infty$.

And by definition $\mathcal{T} \subseteq \mathcal{F}_\infty$; so since $\mathcal{T}$ is independent of a bigger σ-algebra than itself, this in particular means it's independent of itself.

So we've shown $\mathcal{T}$ is independent of itself, which by the definition of independence between σ-algebras means

$$\mathbb{P}[A \cap A] = \mathbb{P}[A] = \mathbb{P}[A]^2,$$

proving the first part of Kolgomorov's theorem.

Now we'll prove the second part — that any $\mathcal{T}$-measurable X is constant almost everywhere (which means we need to pick the constant c). To do so, first we know that for every x, we have $\{X \leq x\} \in \mathcal{T}$ (because X is $\mathcal{T}$-measurable), so $\mathbb{P}[X \leq x] \in \{0, 1\}$. For some x this probability has to be 1, so we can just take c to be the smallest such x — i.e., we define

$$c = \inf\{x \in \mathbb{R} \mid \mathbb{P}[X \leq x] = 1\}.$$

Then for this choice of c, we have $\mathbb{P}[X = c] = 1$ — this is because $\mathbb{P}[X < c] = 0$ by our choice of c.

So every random variable which is measurable with respect to $\mathcal{T}$ is constant almost everywhere, as desired. □

Student Question. *Why is* $\limsup_{n\to\infty} \frac{1}{n}\sum X_n$ *is measurable — doesn't it depend on* X_1*?*

Answer. It doesn't actually depend on X_1, because we're taking $n \to \infty$; we have

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \limsup_{n\to\infty} \frac{1}{n}\sum_{k=2}^{n} X_k,$$

because $\frac{1}{n}X_1 \to 0$. In general, if you take two sequences (a_n) and (b_n) with $a_n \to 0$ as $n \to \infty$, then $\limsup_{n\to\infty} a_n = \limsup_{n\to\infty} a_n + b_n$ (this is a result from real analysis), and that's exactly what we did here.

More generally, if you exclude from this sum any finite collection of X_k's, then the lim sup is still the same; that's the main point (that the lim sup doesn't depend on any finite collection of random variables).

§6.6 The Lebesgue integral

We've seen several main notions of convergence and stated and proven Kolgomorov's 0–1 law; now it's time to talk about integration. We're going to rigorously construct the integral of a measurable function.

How will we construct the integral in this abstract form? The idea is to first construct the integral for functions which are simple to study (so-called 'simple functions'). And then you can take any arbitrary function and *approximate* it by simple functions, and define the integral of your arbitrary function as the limit of the integrals of these approximating simple functions.

§6.6.1 Simple functions

We'll start by defining the notion of a simple function that we'll consider.

Definition 6.13 (Simple function)**.** Let $(E, \mathcal{E}, \mu)$ be a measure space. A simple function is a measurable function which can be written as a *finite* linear combination of nonnegative multiples of indicator functions — i.e., a function of the form $f = \sum_{k=1}^{n} a_k 1_{A_k}$, where $a_k \geq 0$ and $A_k \in \mathcal{E}$ for each k.

So simple functions are just finite (nonnegative) linear combinations of indicator functions of measurable events. The idea to construct the integral is to construct it for such functions; and then to take arbitrary measurable functions and approximate them by functions of this form. We'll see this makes sense in its abstract form (here we have an abstract measure space, not necessarily something like $\mathbb{R}$ with the Borel σ-algebra).

Proposition 6.14

A function is simple if and only if it is measurable, nonnegative, and takes only finitely many values.

So this is another way to define a simple function (as any function which is measurable and nonnegative and takes values in a finite set); this is left as an exercise.

Simple functions are much easier to study than arbitrary functions; we'll start by defining the integral of such functions, in the way you'd expect.

Definition 6.15 (Integral of simple functions)**.** The integral of a simple function $f = \sum_{k=1}^{n} a_k 1_{A_k}$ is

$$\mu(f) = \sum_{k=1}^{n} a_k \mu(A_k).$$

This makes sense — it's the natural definition. We use the convention that $0 \cdot \infty = 0$ — because in this expression you might have an event A_k with $\mu(A_k) = \infty$, and then you multiply it by a coefficient $a_k = 0$; in that case, you get 0.

Student Question. *Do we require the A_k's to be disjoint?*

Answer. We don't need to — if you take any two linear combinations which give you the same function, then the integral (defined in this way) will be the same. This is not obvious, but it's not difficult; mostly it is a tedious exercise of just combining things.

§6.6.2 Integrals of general measurable functions

Now we have the integral of a simple function; the next step is to extend this to nonnegative measurable functions (and then we can get to arbitrary measurable functions via a method described earlier). And this definition will again be the one that it should be.

Definition 6.16 (Integral)**.** Let f be a nonnegative measurable function. Then we define the integral of f, denoted $\mu(f)$, as

$$\mu(f) = \sup\{\mu(g) \mid g \le f,\ g \text{ is simple}\}.$$

So we take all the simple functions that remain below f, and look at all their integrals; and we define the integral of f as the maximum (or rather, supremum) of all these integrals. (We do something similar in a Riemann integral, where we approximate the graph of a function from above and below; here we're just approximating from below.)

Definition 6.17. For an arbitrary measurable function f, we write $f = f^+ - f^-$ where $f^+ = \max\{f, 0\}$ and $f^- = -\min\{f, 0\}$, and we define $|f| = f^+ + f^-$.

For an arbitrary function f (which may take positive and negative values), the integral is not necessarily defined. The integral of a nonnegative function is always well-defined in the above way (though it may be infinite), but the integral of an arbitrary function might not be defined — we'd like to define it by $\mu(f^+) - \mu(f^-)$, but this doesn't make sense if one of these is infinite. However, it turns out that it *does* make sense if $|f|$ has finite integral, leading to the following definition.

Definition 6.18. We say f is integrable if $\mu(|f|) < \infty$. In this case, we write $\mu(f) = \mu(f^+) - \mu(f^-)$.

If only one of f^+ and f^- has infinite integral, we can still make sense of the integral, but it will be infinite. (If both integrals were infinite, then we could not make sense of this.)

§6.6.3 Lebesgue integral vs. Riemann integral

In the case where we're integrating over a subset of $\mathbb{R}$ with respect to $\mathcal{B}(\mathbb{R})$ and the Lebesgue measure, we call this the *Lebesgue integral*.

But we've also learned the Riemann integral; so it's natural to ask for the relationship between the two. It turns out that any function which is Riemann integrable is also Lebesgue integrable; we will state but not prove this.

Proposition 6.19

Let $f: [0,1] \to \mathbb{R}$ be Riemann integrable. Then f is also Lebesgue integrable, and the integrals agree.

So if you have a function which is Riemann integrable, then it'll also be Lebesgue integrable with the same integral. But the point is that there are lots of other functions which are Lebesgue integrable but not Riemann integrable. So the collection of Lebesgue integrable functions is much wider. There's a very typical example that you usually see in real analysis.

Example 6.20

Let $f = 1_{[0,1]\setminus\mathbb{Q}}$ be the indicator function of irrational numbers in $[0,1]$. Then f is not Riemann integrable (if you approximate its graph from above and below, you'll get different limits). But it is Lebesgue integrable (since f is a simple function); in particular, by the definition of the integral of a simple function, we have

$$\mu(f) = \mu([0,1] \setminus \mathbb{Q}) = 1 < \infty.$$

§6.6.4 Some properties of the Lebesgue integral

We'll now quickly mention a few important properties of the integral.

Proposition 6.21

If f and g are simple and $a, b \geq 0$, then $\mu(af + bg) = a\mu(f) + b\mu(g)$.

In other words, we have linearity — when we restrict to simple functions and nonnegative coefficients. We're going to see that this is true even without these restrictions (probably next lecture).

Proposition 6.22

If $f \leq g$, then $\mu(f) \leq \mu(g)$.

This holds for any integrable f and g, and follows by the definition of the integral.

Proposition 6.23

If f is a nonnegative measurable function, then f is 0 almost everywhere if and only if $\int f = 0$.

So if we have a nonnegative measurable function whose integral is 0, then this function has to be 0 almost everywhere. This just follows from the definition of the integral. This is not true without the nonnegativity restriction — you could have a function for which the set of points where it's positive and negative cancel out.

Next lecture we're going to prove these properties, for arbitrary integrable functions.

§7 September 26, 2024

In the previous lecture, we proved Kolgomorov's $\{0,1\}$-law, which is a very important result (as we'll see later). We also started rigorously constructing the integral in an abstract measure space. The first step was to construct the integral for simple functions; then we constructed it for nonnegative measurable functions

by approximating them by simple functions and taking limits; and then we wrote arbitrary functions as differences of nonnegative functions, and under certain conditions, the integral of the original is the difference of their integrals.

§7.1 Some properties of the integral

Our next goal is to prove properties of the integral. One such property is linearity. This holds for free for simple functions, but we want to show it's true for arbitrary nonnegative functions.

To prove such properties, if we approximate a nonnegative measurable function by simple functions, we want to somehow be able to interchange limits and integrals. For this, we will use some convergence results from measure theory.

The first is the monotone convergence theorem; we won't prove it (it's standard and very useful).

Theorem 7.1 (Monotone convergence theorem)

Suppose that (f_n) and f are nonnegative measurable functions with the property that $f_n(x) \uparrow f(x)$ as $n \to \infty$ μ-almost everywhere. Then $\mu(f_n) \uparrow \mu(f)$.

In other words, if we have any sequence of nonnegative measurable functions which converge increasingly pointwise to some measurable function f, then their integrals converge as well. Here we assume everything is nonnegative; this is not necessarily the case for arbitrary functions (we'll see another theorem, the dominated convergence theorem, to deal with that).

We'll now see how this is applied to prove basic properties of the integral.

Theorem 7.2

Let f and g be nonnegative measurable functions, and fix $a, b \geq 0$. Then:

(i) Linearity: $\mu(af + bg) = a\mu(f) + b\mu(g)$.

(ii) Monotonicity: If $f \leq g$, then $\mu(f) \leq \mu(g)$.

(iii) We have $f = 0$ μ-almost everywhere if and only if $\mu(f) = 0$.

Last lecture, we commented the last property is not true if f is an arbitrary measurable function (i.e., without the nonnegativity condition), if the positive and negative parts of f cancel each other.

The idea is to fix f and g and approximate them by simple functions; we know these properties are true for simple functions, so combined with the monotone convergence theorem, we will get the properties for arbitrary nonnegative measurable functions.

Proof of (i). As before, let $f_n = \min\{2^{-n} \lfloor 2^n f \rfloor, n\}$, and similarly define $g_n = \min\{2^{-n} \lfloor 2^n g \rfloor, n\}$. Then both f and g are simple, and they converge pointwise to f and g, respectively (in an increasing way) — this is immediate by the definitions of the functions f_n and g_n.

Now we can apply the monotone convergence theorem — this gives that $\mu(f_n) \uparrow \mu(f)$ and $\mu(g_n) \uparrow \mu(g)$ as $n \to \infty$. And we also have $\mu(af_n + bg_n) \uparrow \mu(af + bg)$, for the same reason.

But (i) is true for the simple functions f_n and g_n — this is actually clear by the definition of the integral of simple functions. So we have

$$\mu(af_n + bg_n) = a\mu(f_n) + b\mu(g_n).$$

This is true for all n. And now we can just take limits — as $n \to \infty$, the left-hand side converges to $\mu(af+bg)$ by the monotone convergence theorem, while the right-hand side converges to $a\mu(f) + b\mu(g)$. And since our two terms are equal, their limits are also equal; this shows the linearity property (i) is true. □

Proof of (ii). There are many ways to prove (ii); it can be done using the monotone convergence theorem, but it can also be done directly from the definition. Recall that

$$\mu(g) = \sup\{\mu(h) \mid h \text{ is simple}, h \leq g\}.$$

But this is at least

$$\sup\{\mu(h) \mid h \text{ is simple}, h \leq f\},$$

because any $h \leq f$ also satisfies $h \leq g$ (which means the second set is a subset of the first, and if we have $A \subseteq B$, then $\sup A \leq \sup B$).

But the last term is just $\mu(f)$ by definition. So we get $\mu(g) \geq \mu(f)$. □

Proof of (iii). We'll first prove the backwards direction — we assume $\mu(f) = 0$, and we want to show $f = 0$ μ-almost everywhere.

We'll argue by contradiction — suppose that f is *not* 0 μ-almost everywhere, meaning $\{x \in E \mid f(x) \neq 0\}$ has positive measure under μ. Then we can write

$$\{x \in E \mid f(x) \neq 0\} = \bigcup_n A_n,$$

where for each $n \in \mathbb{N}$, we define $A_n = \{x \in E \mid f(x) \geq \frac{1}{n}\}$. Since we assumed $\{x \in E \mid f(x) \neq 0\}$ has nonzero measure, we obtain that for some n, we have $\mu(A_n) > 0$ (because if all the A_n's had measure 0, then their union would also have to have measure 0 by the subadditivity property of the measure).

Now we will use the fact that (ii) holds and get a contradiction. For this, we consider the function $h = \frac{1}{n}1_{A_n}$. Then we have $f \geq h$, because if A_n holds, then by the definition of A_n we have $f \geq \frac{1}{n} = h$, while if A_n doesn't hold then the right-hand side is 0, so this is always true.

And so $f \geq h$, which means $\mu(f) \geq \mu(h)$ by (ii). But $\mu(h) = \frac{1}{n}\mu(A_n) > 0$ by assumption. And that's a contradiction, since we assumed $\mu(f) > 0$.

So this implies $\mu(\{x \in E \mid f(x) \neq 0\}) = 0$, so f is 0 μ-almost everywhere.

Now we'll prove the other direction — this means we assume f is 0 μ-almost everywhere, and we need to prove $\mu(f) = 0$. Again, we're going to approximate f for this — let $f_n = \max\{2^{-n}\lfloor 2^n f \rfloor, n\}$. Then we have $f_n \uparrow f$ as $n \to \infty$, and $f_n = 0$ μ-almost everywhere (because $f_n \leq f$, and this is true of f).

This implies $\mu(f_n) = 0$, because (iii) is true for simple functions. Combined with the monotone convergence theorem, this implies

$$\mu(f) = \lim_{n\to\infty} \mu(f_n) = 0.$$ □

So as we can see, this was an immediate application of the monotone convergence theorem combined with the fact that these properties are true for simple functions. So the integral satisfies the expected properties for nonnegative measurable functions.

We will see that the monotone convergence theorem is applied many times in measure theory, and also in this course.

The next step is to prove the analogous properties for arbitrary measurable functions; we'll see (i) and (ii) are identical, but (iii) doesn't hold (as stated earlier).

Theorem 7.3

Let f and g be integrable, and let $a, b \geq 0$. Then:

(i) $\mu(af + bg) = a\mu(f) + b\mu(g)$.

(ii) If $f \leq g$, then $\mu(f) \leq \mu(g)$.

(iii) If $f = 0$ μ-almost everywhere, then $\mu(f) = 0$.

So one direction of (iii) is true, but the other one is not — if $\mu(f) = 0$, this does not imply $f = 0$ almost everywhere.

We'll prove this by applying the previous theorem.

Proof of (i). We're first going to show that this is true when $b = 0$, so that the integral satisfies a scaling property.

As before, we write $f = f^+ - f^-$. Then for any $a \geq 0$, we have $\mu(af) = \mu(af^+) - \mu(af^-)$. But f^+ and f^- are both nonnegative measurable functions, and for such functions we know the integral satisfies the scaling property. So this is just $a\mu(f^+) - a\mu(f^-) = a\mu(f)$. So indeed the integral satisfies the scaling property.

We're not quite done; we've proved the integral satisfies the scaling condition, but now we need to show that it's additive, meaning that $\mu(f + g) = \mu(f) + \mu(g)$. We know this is true for nonnegative measurable functions; now we need a little trick to make the same statement in general.

If we let $h = f + g$, then we can write

$$h^+ - h^- = (f^+ - f^-) + (g^+ - g^-).$$

And by rearranging, we have

$$h^+ + f^- + g^- = h^- + f^+ + g^+.$$

And now the advantage is that all these functions are nonnegative, and we know the integral satisfies the linearity property. So the integral of the left-hand side and right-hand side are equal and now we have linearity, meaning that

$$\mu(h^+) + \mu(f^-) + \mu(g^-) = \mu(h^-) + \mu(f^+) + \mu(g^+).$$

Now we're in a position to rearrange — everything here is finite (we assumed f and g are integrable), so this means

$$\mu(h^+) - \mu(h^-) = \mu(f^+) - \mu(f^-) + \mu(g^+) - \mu(g^-).$$

but now we are done, because the first term is just $\mu(h) = \mu(f + g)$ by definition; and the right-and side is $\mu(f) + \mu(g)$, again by the definition of the integral.

So we have shown

$$\mu(f + g) = \mu(f) + \mu(g).$$

So we have additivity and scaling, and combining these gives (i).

(There was one trick here, to express $f + g$ in this way and rearrange to somehow use nonnegativity.) □

Now that we've shown (i), the remaining properties are quite easy.

Proof of (ii). If $f \leq g$, then $g - f \geq 0$. And then by the previous theorem, we have $\mu(g - f) \geq 0$. But by (i) this is $\mu(g) - \mu(f)$, which implies $\mu(g) \geq \mu(f)$. □

Proof of (iii). Suppose that $f = 0$ μ-almost everywhere; this means both f^+ and f^- have to be 0 μ-almost everywhere. This implies $\mu(f^+) = \mu(f^-) = 0$, so $\mu(f) = \mu(f^+) - \mu(f^-) = 0$. □

As a counterexample for the reverse direction of (iii) from before:

Example 7.4

Consider the coordinate axis, and take the function

$$f(x) = \begin{cases} -1 & \text{if } -1 \le x < 0 \\ 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(i.e., the sign function restricted to $[-1, 1]$). This has integral 0, but is not zero almost everywhere.

These are the basic properties of the integral; they are what you'd expect.

Student Question. *After defining h^+ and h^-, why did we have to rearrange, take the measure, and rearrange back — why couldn't we just take the measure immediately?*

Answer. Because we don't yet know additivity — we know $\mu(h^+ - h^-) = \mu(f^+ - f^- + g^+ - g^-)$. But we don't yet know that $\mu(f^+ - f^- + g^+ - g^-)$ is the appropriate sum of individual integrals, because we don't yet know additivity — that's what we want to show. (The problem is that these functions are not nonnegative.)

The rearrangement solves this problem because now everything is nonnegative, so we *do* have additivity.

Student Question. *In the proof of (iii) using (i), aren't we using a negative scalar — we're saying $\mu(f^+ - f^-) = \mu(f^+) - \mu(f^-)$, but here this corresponds to $a = -1$?*

Answer. We only showed (i) for $a, b \ge 0$, but it's actually true for arbitrary a and b — because you can write $af = (-a)(-f)$, for example.

As mentioned, the converse to (iii) in the general case is not true. But we do have a partial converse — under certain directions, we do have a converse.

Proposition 7.5

If $\mathcal{A}$ is a π-system with the property that $E \in \mathcal{A}$ and $\sigma(\mathcal{A}) = \mathcal{E}$, and f is an integrable function with the property that $\mu(f \cdot 1_A) = 0$ for all $A \in \mathcal{A}$, then $f = 0$ μ-almost everywhere.

So to show $f = 0$ almost everywhere, we need to assume some further things; and what we assume is that the integral of f is 0 even restricted to any element of a π-system (which generates the σ-algebra).

Proof. This will essentially follow by Dynkin's π-system lemma. The first observation is that it suffices to show that $\mu(f \cdot 1_B) = 0$ for all $B \in \mathcal{E}$ — if we can show this is true, then we're done. This is because we can write

$$f = f \cdot 1_{\{x | f(x) \ge 0\}} + f \cdot 1_{\{x | f(x) < 0\}}.$$

Both of these sets are measurable, so we get that the integral of the first function is zero; but it's nonnegative, so that function must be zero almost everywhere (and the same is true for the second term).

As we usually do in such cases, we use Dynkin's π-system lemma. So here we're going to let

$$\mathcal{D} = \{A \in \mathcal{E} \mid \mu(f \cdot 1_A) = 0\}.$$

Our goal is to show that $\mathcal{D}$ is the entire σ-algebra. And the way we do this is using Dynkin's π-system lemma, as usual; so what we want to do is show $\mathcal{D}$ is a d-system. (It contains the π-system $\mathcal{A}$, so by Dynkin's, if we can show that it's a d-system then it contains $\sigma(\mathcal{A}) = \mathcal{E}$.)

The fact that $\mathcal{D}$ is a d-system follows by the properties of the integral from the previous theorem (this is immediate to check); and $\mathcal{A} \subseteq \mathcal{D}$. So this implies $\mathcal{D} = \sigma(\mathcal{A}) = \mathcal{E}$ by Dynkin's lemma, and we're done. □

This means if we make the further assumption that the integral of f is 0 on every element of a generating π-system, then we do get that f is 0 μ-almost everywhere.

§7.2 Exchanging summation and integration

In analysis, there's a fundamental question:

> **Question 7.6.** If you have a sum of functions, under which conditions is the integral of the sum the sum of integrals?

If the functions are nonnegative and measurable, you can always interchange the sum and integral — in particular, we have the following proposition.

> **Proposition 7.7**
> Suppose that (g_n) is a sequence of nonnegative measurable functions. Then we have $\mu(\sum_n g_n) = \sum_n \mu(g_n)$.

So if we have nonnegative measurable functions, then we can always interchange summation and integration. This is not true if you have arbitrary measurable functions — the sum might not even converge.

The proof follows directly from the monotone convergence theorem and the linearity property of the integral.

Proof. If we consider just the sum of the first N terms, we have $\sum_{n=1}^N g_n \uparrow \sum_{n\geq 1} g_n$ μ-almost everywhere as $N \to \infty$. Now linearity tells us that $\mu(\sum_{n=1}^N g_n) = \sum_{n=1}^N \mu(g_n)$, and the monotone convergence theorem tells us that this converges to $\mu(\sum_{n\geq 1} g_n)$.

But we also have that $\sum_{n=1}^N \mu(g_n) \uparrow \sum_{n\geq 1} \mu(g_n)$ as $n \to \infty$ (by definition of an infinite sum).

This implies $\mu(\sum_{n\geq 1} g_n) = \sum_{n\geq 1} \mu(g_n)$, since they're limits of the same sequence. And that's exactly what we wanted to prove. □

So for nonnegative measurable functions, we can always switch the order of integration and summation.

> **Student Question.** *Why did we need 'almost everywhere' when saying $\sum_{n=1}^N g_n \to \sum_n g_n$?*
>
> **Answer.** We don't; they actually converge for every point. (We only need almost everywhere convergence to apply the monotone convergence theorem; but in this setting we actually have convergence for every x.)

§7.3 Fatou's lemma

It's really nice to be able to exchange imits and integrals — this will be extremely useful in the future. And the monotone convergence theorem is one very useful tool that helps us do so.

We'll now see another important theorem that also helps with this.

Theorem 7.8 (Fatou's lemma)

Let (f_n) be a sequence of nonnegative measurable functions. Then

$$\mu(\liminf_{n\to\infty} f_n) \le \liminf_{n\to\infty} \mu(f_n).$$

In one of the problem sets, we proved the same statement, but when f_n is the indicator function of some event (i.e., set); so this is a more generalized version.

Proof. Here we start with the following observation: if we take any $k \ge n$, then we always have $\inf_{m\ge n}(f_m) \le f_k$ (by the definition of the infimum).

Now by the monotonicity of the integral (as shown in the previous theorems), this means

$$\mu(\inf_{m\ge n}(f_m)) \le \mu(f_k).$$

And this is true for every $k \ge n$.

This is true for *all* $k \ge n$, so this means $\mu(\inf_{m\ge n}(f_m)) \le \inf_{k\ge n} \mu(f_k)$. And the infimum on the right is at most $\liminf_{N\to\infty} \mu(f_N)$, by the definition of the lim inf.

So to complete the proof of Fatou's lemma, it suffices to show that $\mu(\inf_{m\ge n} f_m) \to \mu(\liminf_{n\to\infty} f_n)$. To see this, note that $\inf_{m\ge n} f_m \uparrow \liminf_{m\to\infty} f_m$ as $n \to \infty$ (this is the definition of a lim inf). By the monotone convergence theorem, this implies that $\mu(\inf_{m\ge n}(f_m)) \uparrow \mu(\liminf_{n\to\infty} f_n)$ as $n \to \infty$. And that's the end of the proof; so $\mu(\liminf f_n) \le \liminf \mu(f_n)$ (becaues the right-hand side bounds every term of our sequence). □

That's another very useful result. Something to note here is that the inequality here might be strict — we don't always have equality. We'll write down a counterexample.

Example 7.9

Let $(E, \mathcal{E}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \text{Lebesgue})$. Let $f_n = 1_{[n,n+1]}$. Then we have $\liminf_{n\to\infty} f_n = 0$, because these intervals are disjoint; this means $\mu(\liminf_{n\to\infty} f_n) = 0$. But we have $\mu(f_n) = 1$ for all n. This implies $\liminf_{n\to\infty} \mu(f_n) = 1$ as well, which is not 0.

So in Fatou's lemma, it's possible to have a strict inequality; it's not necessarily an equality.

Fatou's lemma helps us prove quite a lot of results about convergence.

§7.4 Dominated convergence theorem

We've already seen two main theorems so far which help us change the order of integration and taking limits. One was the monotone convergence theorem (which we didn't prove, but will take for granted); the other was Fatou's lemma.

> **Question 7.10.** What happens to e.g. the monotone convergence theorem if the functions aren't nonnegative?

In that case, the theorem is not true. But with slightly different assumptions, it turns out we *can* still interchange the order of limits and integrals.

Theorem 7.11 (Dominated convergence theorem)

Let (f_n) and f be measurable functions such that $f_n(x) \to f(x)$ as $n \to \infty$ for μ-almost every x (i.e., we have pointwise convergence μ-almost everywhere). Suppose that there is an integrable function g which dominates all the f_n's, in the sense that $|f_n| \leq g$ for all n. Then $\mu(f_n) \to \mu(f)$ as $n \to \infty$.

So here we have a slightly different assumption — we use a dominant function g which dominates all the f_n's in this sense. And under this assumption, the limit of the integral is the integral of the limit.

Proof. First, we'll check that f is integrable. By definition, we have $|f| = \lim_{n\to\infty} |f_n| \leq g$; so by the monotonicity of the integral, we have $\mu(|f|) \leq \mu(g)$. And the latter is finite, because we know g is integrable. So all the f_n's and f are integrable, so it does make sense to talk about their integrals.

Also note that $g + f_n$ is nonnegative, because $-f_n \leq g$; and similarly $g - f_n$ is also nonnegative.

Now we're going to apply Fatou's lemma for these two sequences of functions, and somehow extract the statement that we want.

First, by the linearity of the integral we have

$$\mu(g) + \mu(f) = \mu(g + f) = \mu\left(\liminf_{n\to\infty} g + f_n\right).$$

And now we can apply Fatou's lemma, because we have a sequence of nonnegative measurable functions; so this is at most $\liminf_{n\to\infty} \mu(g + f_n)$. But now we can apply linearity again, and say that this is $\liminf_{n\to\infty} \mu(g) + \mu(f_n)$. And $\mu(g)$ is a constant, so we can take it out of the lim inf; this gives us

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

And $\mu(g)$ cancels from both sides, so we obtain

$$\mu(f) \leq \liminf_{n\to\infty} \mu(f_n).$$

And now we can do the same for the function $g - f$ (which is still nonnegative); we have

$$\mu(g)-\mu(f) = \mu(g-f) = \mu(\liminf_{n\to\infty}(g-f_n)) \leq \liminf_{n\to\infty} \mu(g-f_n) = \liminf_{n\to\infty}(\mu(g)-\mu(f_n)) = \mu(g)+\liminf_{n\to\infty} -\mu(f_n).$$

But the lim inf of the negative of a sequence is the negative of the lim sup, so we get

$$\mu(g) - \mu(f) \leq \mu(g) - \limsup_{n\to\infty} \mu(f_n),$$

which tells us

$$\limsup_{n\to\infty} \mu(f_n) \leq \mu(f).$$

But now putthing these together, we have

$$\mu(f) \leq \liminf_{n\to\infty} \mu(f_n) \leq \limsup_{n\to\infty} \mu(f_n) \leq \mu(f).$$

This means all these terms have to be equal; in particular, since the lim sup and lim inf coincide, this means $\lim_{n\to\infty} \mu(f_n)$ exists and equals $\mu(f)$. □

So if we have a dominant function g, then we can interchange between integrals and limits.

Student Question. *Did we assume f was integrable?*

Answer. We don't need to assume it, we can prove it from the hypotheses of the theorem — we have $|f| \leq g$, and we know $\mu(g)$ is finite, so $\mu(|f|)$ is also finite; this means f is integrable.

§7.5 Constructing measures

The next thing we'll mention is how to construct new measures from old measures. This is something we've already done in previous lectures, but now we'll see a few more ways to do so — not in this lecture, but in the next.

One obvious way is the *restriction measure*, where we just take the original measure and restrict it to a smaller σ-algebra.

Definition 7.12 (Restriction of a measure space)**.** Let $(E, \mathcal{E}, \mu)$ be a measure space, and fix a measurable set $A \in \mathcal{E}$. Then the restriction of the measure space to A is $(A, \mathcal{E}_A, \mu_A)$, where

$$\mathcal{E}_A = \{B \in \mathcal{E} \mid B \subseteq A\},$$

and $\mu_A(B) = \mu(B)$ for all $B \in \mathcal{E}_A$.

So we just take the original measure space, take any measurable set, and restrict everything to that set. This is a very naive way to construct a measure; we will soon see more elaborate ways.

It is an easy exercise to check that $(A, \mathcal{E}_A, \mu_A)$ is indeed a measure space. And we also have that the restriction of any measurable function on the original space is still measurable; we'll write this down as a proposition.

Proposition 7.13

Let $(E, \mathcal{E}, \mu)$ and $(F, \mathcal{F}, \mu')$ be measure spaces, and fix any $A \in \mathcal{E}$. Suppose we have a function $f: E \to F$ which is measurable. Then $f|_A$ is $\mathcal{E}_A$-measurable.

Proof. If we fix any $B \in \mathcal{F}$, then we have $(f|_A)^{-1}(B) = f^{-1}(B) \cap A$, and by definition this set is in $\mathcal{E}_A$ (because $f^{-1}(B)$ is a measurable set in the original σ-algebra). □

So measurable functions restricted to a smaller σ-algebra are still measurable. The same is true for integrable functions — if we start with an integrable function and restrict, then $f|_A$ is still integrable in the smaller space. (This is an easy exercise.)

§8 October 1, 2024

In the previous lecture, we saw some important limit theorems, like Fatou's lemma, the monotone convergence theorem, and the dominated convergence theorem. We said we'll use these extensively to interchange between limits and integrations. Last time we finished the lecture by defining the restriction measure, an obvious way to define a new measure given a starting measure. In this lecture, the main focus will be on product spaces; we'll construct a product measure on the product σ-algebra. This immediately gives the construction of the Lebesgue measure in higher dimensions as well.

§8.1 The pushforward measure

We'll start by recalling the definition of the pushforward (or image) measure (which we saw earlier, but will now state in a more abstract way).

Definition 8.1 (Pushforward measure)**.** Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measure spaces, and suppose we have a function $f: E \to G$ which is measurable. Then if μ is a measure on E, we define the pushforward (or image) measure of μ as $\nu = \mu \circ f^{-1}$.

So in this way, if we have a measure μ and a measurable function f, we can always consider the pushforward measure, and this will always be a measure. So this is another naive way of constructing a new measure.

We mention this again because it has some applications to integration. In particular, we have the following result.

Proposition 8.2

If g is a nonnegative measurable function on G, then $\nu(g) = \mu(g \circ f)$.

This is an application of the pushforward measure — if we have two measurable functions f and g and we want to compute the integral with respect to μ, this is the same as the integral of g with respect to the pushforward.

To prove this for every nonnegative measurable function, we approximate by simple functions — we show this is true for nonnegative simple functions, and then take limits to get arbitrary g. To justify why we can take limits, we use the monotone convergence theorem.

§8.2 Density

Finally, another way to specify a measure is by specifying a *density*. We'll use this many times in the course.

Definition 8.3 (Density)**.** Let $(E, \mathcal{E})$ be a measure space with measure μ, and suppose that f is a nonnegative measurable function on E. Then we define $\nu(A) = \mu(f \cdot 1_A)$ for every $A \in \mathcal{E}$.

So we start with an original measure μ and nonnegative function f; and we define the measure of a set A with respect to ν as just the integral of f restricted to A.

Proposition 8.4

The function ν (as above) is indeed a measure.

Proof. First, we have $\nu(\emptyset) = \mu(f \cdot 1_\emptyset) = 0$ (since $1_\emptyset$ is the indicator function of the empty set, which is 0).

The second thing we need to check is countable additivity. Suppose that we take (A_n) to be a disjoint sequence in our σ-algebra; then $\mu(\bigcup_n A_n) = \mu(f 1_{\bigcup A_n})$. But because the A_n are disjoint, we have $1_{\bigcup A_n} = \sum 1_{A_n}$. And because we have nonnegative functions, we can interchange between summation and integration (this is an immediate convergence of the monotone convergence theorem). So this is just $\sum_n \mu(f 1_{A_n}) = \sum_n \nu(A_n)$. □

So this is another way to construct new measures.

What's the importance of this in the context of probability? Here ν plays the role of the *low* of a random variable, and μ plays the role of a Lebesgue measure. If a function f exists with these properties, then we say the random variable X has density f.

Definition 8.5 (Density of a random variable)**.** Let X be a random variable. We say that X has a density if μ_X has a density with respect to the Lebesgue measure — in other words, there exists a nonnegative measurable function f_X such that

$$\mathbb{P}[X \in A] = \int_A f_X(x)\,dx.$$

We call f_X the density of X.

Recall that μ_X is the low of X (which is a Borel measure). This is the special case of the earlier definition where ν is μ_X (recall that we write $\mu_X(A) = \mathbb{P}[X \in A]$, and $\int$ denotes the Lebesgue integral).

Most random variables we're familiar with have this property — e.g., the exponential or gamma distribution all have densities — and we'll see some of them later in this course.

Student Question. *Does every random variable have a density?*

Answer. No — there are random variables which cannot be expressed in this way. But most of the random variables we're interested in — e.g., the Gaussian, exponential, and gamma distribution — have lows of this form.

For instance, if the random variable takes discrete values, then this can't happen — if you integrate any function with respect to the Lebesgue measure, on a discrete set you will get 0 — because the Lebesgue measure of a discrete set is 0.

§8.3 Product measures

We've now seen a few ways to construct new measures. The next one we'll see is the product measure, defined on the product σ-algebra. We'll start by defining the product σ-algebra.

Definition 8.6 (Product σ-algebra)**.** Let $(E_1, \mathcal{E}_1, \mu_1)$ and $(E_2, \mathcal{E}_2, \mu_2)$ be finite measure spaces. Let

$$\mathcal{A} = \{A_1 \times A_2 \mid A_1 \in \mathcal{E}_1,\ A_2 \in \mathcal{E}_2\}$$

be the π-system consisting of rectangles. Then the product σ-algebra is defined as $\mathcal{E} = \sigma(\mathcal{A})$.

It's easy to see that $\mathcal{A}$ is a π-system on $E_1 \times E_2$, and the product σ-algebra is just the σ-algebra generated by this π-system. This is an abstract generalization of what we're doing in two dimensions (with $\mathbb{R} \times \mathbb{R}$).

What we'd like to do is use μ_1 and μ_2 to construct a new measure on the product σ-algebra.

First, we have powerful tools like Caratheodory's extension theorem; why can't we use that here? We could, but we'd like an explicit description of this measure, so we'll do it in a different way (Caratheodory doesn't give us a very explicit form — the definition involves complicated things like infimums and outer measures).

The idea is to define μ (on the product σ-algebra) such that

$$\mu(A) = \int_{E_1} \left(\int_{E_2} 1_A(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

So we have an inside integral with respect to the second measure, and then an outside integral with respect to the first. This is our candidate for the measure extending μ_1 and μ_2.

Before we do that, we need to make sure that this statement makes sense — in particular, that the function $\int_{E_2} 1_A(x_1, x_2)\mu_2(dx_2)$ is measurable and can be integrated and so on. That's the purpose of the second lemma — to make sure that this thing we've defined is really a well-defined set function.

Lemma 8.7

Let $E = E_1 \times E_2$, and suppose that $f: E \to \mathbb{R}$ is $\mathcal{E}$-measurable (i.e., measurable with respect to the product σ-algebra). Then the following conditions hold:

(i) For every $x_2 \in E_2$, the function on E_1 given by $x_1 \mapsto f(x_1, x_2)$ is $\mathcal{E}_1$-measurable.

(ii) If f is bounded or nonnegative, then the function on E_2 given by $f_2(x_2) = \int_{E_1} f(x_1, x_2)\mu_1(dx_1)$ is $\mathcal{E}_2$-measurable.

Once we have this lemma, it's easy to see our definition of μ from earlier makes sense — (ii) says that the function we have on the inside is measurable in E_1, and therefore it makes sense to integrate it. So (i) shows that the function on the inside is measurable in E_2 (which allows us to integrate it over E_2); and (ii) says that the resulting integral is measurable in E_1. So μ is a well-defined set function; and after we prove this lemma, we'll show it's the measure we're actually looking for.

Proof of (i). For (i), note that for fixed $x_2 \in E_2$, if we consider the map $i_1: E_1 \to E$ given by $x_1 \mapsto (x_1, x_2)$, then this map is measurable. So then $f \circ i_1$, which is exactly the function in (i), is $\mathcal{E}_1$-measurable — so (i) follows because i_1 is a measurable function and the composition of measurable functions is measurable.

For (ii), we need to show this property is true for any bounded or nonnegative measurable function. For that, we'll invoke a theorem which when we stated it was not clear why it was useful — the monotone class theorem, which stated that if you have a linear space of functions, then under certain conditions, it contains everything. Here we're trying to prove a property for all nonnegative measurable functions, so it's enough to show that the set of functions for which it's true satisfies these conditions.

So we define the set

$$V = \{f: E \to \mathbb{R} \text{ s.t. } f \text{ is measurable and } x_2 \mapsto \int_{E_1} f(x_1, x_2)\mu_1(dx_1) \text{ is } \mathcal{E}_2\text{-measurable}\}.$$

We want to show this satisfies the conditions of the monotone class theorem.

We first need to show that $1_E \in V$, and $1_A \in V$ for all A in a generating π-system, here $\mathcal{A}$. First 1_E is the constant number 1; then for each x_2, we just get a constant function, and constant functions are measurable. When $f = 1_A$, we still get the measure with respect to μ_1 of A_1 (where $A = A_1 \times A_2$). So this is clear.

The second thing is we want V to be a vector space. But this is clear by the linearity of the integral (which we have already shown).

So V satisfies the first two conditions for the monotone class theorem. The last condition of the theorem is that V should be closed under increasing limits of nonnegative sequences. So we consider a sequence (f_n) of nonnegative functions such that $f_n \uparrow f$ pointwise as $n \to \infty$; then $(x_2 \mapsto \int_{E_1} f_n(x_1, x_2)\mu_1(dx_1)) \uparrow (x_2 \mapsto \int_{E_1} f(x_1, x_2)\mu_1(dx_1))$ by the monotone convergence theorem. But all the functions on the left-hand side are measurable, and pointwise limits of measurable functions are also measurable. So the right-hand side is $\mathcal{E}_1$-measurable, which means $f \in V$.

Now by the monotone class theorem, we have that V contains all bounded measurable functions.

This proves (ii) is true for all bounded measurable functions. We also wanted to prove it for all *nonnegative* measurable functions, but this is not too hard to see. If f is a general nonnegative measurable function (but not necessarily bounded), then $f \wedge n$ (defined as $\min\{f, n\}$, where n is some positive integer) is bounded and measurable, which means $f \wedge n \in V$. And then by the monotone convergence theorem (the increasing limit statement from above), we also have $f \in V$. □

So this means μ is indeed a well-defined set function.

Student Question. *Why do we need bounded* or *nonnegative?*

Answer. The two conditions aren't comparable — a bounded function might be negative, and a nonnegative function might be unbounded.

The place we're using boundedness is in the statement of the monotone class theorem, which only gives that V contains all the *bounded* measurable functions. The monotone class theorem does not itself talk about nonnegative measurable functions; but we get it by again applying the monotone convergence theorem writing f as the limit of $f \wedge n$.

Now that we know μ is well-defined, we're ready to prove our theorem — that μ is the unique function extending μ_1 and μ_2 in our product space.

Theorem 8.8

Suppose that μ_1 and μ_2 are *finite* measures. Then there exists a unique measurable set function μ with the property that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for every $A_1 \times A_2 \in \mathcal{A}$.

Definition 8.9. We call this measure μ the product measure, and denote it by $\mu = \mu_1 \otimes \mu_2$.

Again, it's tempting to apply Caratheodory's extension theorem; but we can do something better because we have an *explicit* description of μ.

Proof. As before, we define

$$\mu(A) = \int_{E_1} \left(\int_{E_2} 1_A(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

This is a well-defined set function by the previous lemma. We need to check two things — that this is a measure, and that it satisfies the required property. (We also need to check a third thing, namely uniqueness.)

For the first condition of being a measure, we have $\mu(\emptyset) = 0$ because $1_\emptyset = 0$. Next, we need to check countable additivity. Suppose that (A_n) is a disjoint sequence in $\mathcal{E}$, and $A = \bigcup_n A_n$. Then writing down the definition of $\mu(A)$, we have

$$\mu(A) = \int_{E_1} \left(\int_{E_2} 1_A(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

And because the A_n's are disjoint events, we can write $1_A(x_1, x_2) = \sum_n 1_{A_n}(x_1, x_2)$; then we get a summation inside the second integral, so

$$\mu(A) = \int_{E_1} \left(\int_{E_2} \sum_n 1_{A_n}(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

Now we're going to use the limit theorems — we have nonnegative functions, so we can interchange the summation and integral twice; and then we obtain

$$\mu(A) = \sum_n \int_{E_1} \left(\int_{E_2} 1_{A_n}(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1) = \sum_n \mu(A_n).$$

So this means μ is countably additive, and therefore it is a measure.

Next, it remains to check that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. This is clear — we have

$$\mu(A_1 \times A_2) = \int_{E_1} \left(\int_{E_2} 1_{A_1 \times A_2}(x_1, x_2)\mu_2(dx_2) \right) \mu_2(dx_1).$$

Then we have $1_{A_1 \times A_2}(x_1, x_2) = 1_{A_1}(x_1)1_{A_2}(x_2)$, so we can pull out a factor of $\mu_2(A_2)$ and get

$$\int_{E_1} 1_{A_1}(x_1)\mu_2(A_2)\mu_1(dx_1) = \mu_1(A_1)\mu_2(A_2).$$

For uniqueness, μ must be finite (since μ_1 and μ_2 are), so it's determined by its values on a π-system. □

If the space is infinite, we don't have uniqueness anymore — for example, you can take any constant multiple of the Lebesgue measure in two dimensions.

Here we have a product measure. If we first integrated with respect to E_1 and then E_2, we'd still get a measure; and the point is this measure will still factorize, and by uniqueness they have to be the same. So this means if we swapped the order of integration — integrating over E_1 on the inside and E_2 on the outside — then we'd still get a set function with the given properties, which means by uniqueness that it must be the same. So you can interchange the order of integration.

And this can be generalized, with nonnegative measurable functions vs. bounded measurable functions. This is called Fubini's theorem; we're going to state but not prove it, but the proof is the same (using the theorems we've seen so far, such as the monotone and dominated convergence theorems).

Theorem 8.10 (Fubini's theorem)

(i) If f is a nonnegative measurable function on $E = E_1 \times E_2$, then

$$\mu(f) = \int_{E_1} \left(\int_{E_2} f(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

In particular, we have

$$\int_{E_1} \left(\int_{E_2} f(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1) = \int_{E_2} \left(\int_{E_1} f(x_1, x_2)\mu_1(dx_1) \right) \mu_2(dx_2).$$

(ii) If f is integrable and $A = \{x_1 \in E_1 \mid \int_{E_2} |f(x_1, x_2)|\, \mu_2(dx_2) < \infty\}$, then $\mu_1(E_1 \setminus A) = 0$. If we define

$$f_1(x_1) = \begin{cases} \int_{E_2} f(x_1, x_2)\mu(dx_2) & x_1 \in A \\ 0 & \text{otherwise,} \end{cases}$$

then we have $\mu_1(f_1) = \mu(f)$.

So (i) generalizes what we just did — we integrate with respect to the second coordinate and then the first coordinate. And it doesn't matter the order in which we integrate; we should still get the same thing (namely, $\mu(f)$). This is also sometimes known as *Tonneli's theorem*.

Then (ii) considers what happens if f is *integrable* (rather than nonnegative). Here the same interchange is still valid, but we have to be slightly more careful. We have the same thing as before, but the inside integral is not necessarily well-defined everywhere. But Fubini's theorem says that it's well-defined *almost* everywhere (A is the set on which it's well-defined); and if we integrate this, we still get the same result.

The proof is in the same spirit of the previous theorem, modulo some more limiting arguments (the monotone or dominated convergence theorems).

§8.4 Product spaces and independence

Now that we have the product measures, we'll see their applications in probability. Their applications usually lie in independence — somehow independence can be characterized using product spaces.

Proposition 8.11

Let $X_1, X_2, \ldots, X_n$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which take values in some measure spaces $(E_1, \mathcal{E}_1)$, $(E_2, \mathcal{E}_2)$, $\ldots$, $(E_n, \mathcal{E}_n)$. We define $E = E_1 \times \cdots \times E_n$ and $\mathcal{E}$ as the product σ-algebra $\mathcal{E}_1 \otimes \mathcal{E}_2 \otimes \cdots \otimes \mathcal{E}_n$. Also define $X = (X_1, \ldots, X_n)$. Then X is $\mathcal{E}$-measurable, and the following conditions are equivalent:

(i) $X_1, \ldots, X_n$ are independent.

(ii) $\mu_X = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$.

(iii) For any bounded measurable functions $f_1, \ldots, f_n$, we have

$$\mathbb{E}\left[\prod_{j=1}^{n} f_j(X_j)\right] = \prod_{j=1}^{n} \mathbb{E}[f_j(X_j)].$$

Remark 8.12. We've only defined the product σ-algebra of two spaces, but this can be iterated, and we can define the product σ-algebra of n spaces in the same way (by induction). Similarly, the product measure is defined by generalizing what we did in two dimensions, using induction.

So these are some nice conditions for random variables to be independent. Their values could be in abstract measure spaces; they're not necessarily taking real values. The third condition states that the expectation of products of any measurable images of these is the same as the product of the expectations.

Proof. First we'll go from (i) to (ii). To do this, we let $\nu = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$ be the product measure of the lows of the X_j's. We want to show that $\nu = \mu_X$. To do this, what we actually need to do is show that these two measures agree on a π-system generating the σ-algebra (they're certainly finite measures, since they have total mass 1). We define the obvious π-system

$$\mathcal{A} = \{A_1 \times \cdots \times A_n \mid A_1 \in \mathcal{E}_1, \ldots, A_n \in \mathcal{E}_n\}.$$

Then $\mathcal{A}$ is a π-system generating the product σ-algebra. Moreover, if we take any $A = A_1 \times \cdots \times A_n \in \mathcal{A}$, then we have

$$\mu_X(A) = \mathbb{P}[X \in A] = \mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n]$$

by definition. And now we use independence: we know the probability of the intersection of these events is the product of probabilities, so we get

$$\mu_X(A) = \mu_{X_1}(A_1)\mu_{X_2}(A_2)\cdots\mu_{X_n}(A_n).$$

But this is just $\nu(A)$. So this means the two measures agree on a π-system which generates $\mathcal{E}$, and therefore $\mu_X = \nu$. This means independence gives us this nice expression for the low of $X = (X_1, \ldots, X_n)$ — it's the product measure of the individual lows.

Now we'll go from (ii) to (iii) — now we know μ_X has this nice form, and we want to show the expectation of a product is the product of expectations.

By our assumption, we have

$$\mathbb{E}\left[\prod_{j=1}^{n} f_j(X_j)\right] = \int_E f_1(x_1)f_2(x_2)\cdots f_n(x_n)d\mu_X(x_1, \ldots, x_n)$$

(by the definition of the low of a random variable). But now we know this measure here is just the product measure — so we obtain that this is

$$\int_E f_1(x_1)f_2(x_2)\cdots f_n(x_n)d\mu_{X_1} \otimes \cdots \otimes \mu_{X_n}(x_1, \ldots, x_n).$$

And now we can apply Fubini's theorem; so this can also be written as

$$\int_E \prod_{j=1}^{n} f_j(x_j)\mu_{X_n}(dx_j).$$

And Fubini's theorem lets us interchange the order of integration and multiplication, to write this as

$$\prod_{j=1}^{n} \int_E f_j(x_j)\mu_{X_j}(dx_j).$$

And this is exactly $\prod_{j=1}^{n} \mathbb{E}[f_j(X_j)]$, which is what we wanted.

To complete the proof, we need to show (iii) implies (i). This is easy — we can apply the classical definition of independence of random variables. So we'll apply (iii) with the indicator function of measurable sets — we take $f_k = 1_{A_k}$ for arbitrary $A_k \in \mathcal{E}_k$. Then we have

$$\mathbb{P}[X_1 \in A_1, \dots, X_n \in A_n] = \mathbb{E}\left[\prod_{k=1}^{n} 1_{A_k}(x_k)\right].$$

But by our assumption, this is equal to

$$\prod_{k=1}^{n} \mathbb{E}[1_{A_k}(x_k)] = \prod_{k=1}^{n} \mathbb{P}[A_k].$$

So this means we have independence, since the A_k's were arbitrary measurable sets. This completes the proof, because it shows these random variables X_k are indeed independent. □

This is a very useful proposition because it gives some nice equivalent conditions for a collection of random variables to be independent. And it's a nice application of all this theory about product measures in the context of probability. We're going to use it multiple times in this course when we try to verify whether a collection of random variables is independent.

Student Question. *For independence, we know $X_1, \dots, X_n$ are independent if and only if an expectation of a product is a product of expectations for* all *measurable functions. But do we really need to check this for* all *measurable functions, or is it enough to check it for some smaller subset?*

Answer. It suffices to check it for indicator functions — we can see this from the above proof. (In fact, the statement that it holds for indicator functions implies it holds for all measurable functions, by the above proof; this is not *a priori* obvious.)

§8.5 L^p spaces

We'll now define L^p spaces, since we're going to talk about them when we talk about pth moments of random variables later on.

Definition 8.13 (L^p spaces)**.** Let $(E, \mathcal{E}, \mu)$ be a measurable space. Then for any $1 \le p < \infty$, we define $L^p(E, \mathcal{E}, \mu)$ to be the set of all measurable functions f such that

$$\|f\|_p = \left(\int_E |f|^p \, d\mu\right)^{1/p} < \infty.$$

We'll be interestsed in this kind of integral where μ is the low of a random variable; then this will be the pth moment.

We can also make sense of this when $p = \infty$. Here of course we have $1/\infty = 0$, so we can't use this definition, but we can modify it.

Definition 8.14 (L^∞)**.** We define $L^\infty(E, \mathcal{E}, \mu)$ as the space of measurable functions f such that

$$\|f\|_\infty = \inf\{\lambda > 0 \mid |f| \leq \lambda \ \mu\text{-almost everywhere}\} < \infty.$$

So if the infimum is a finite number, then we say f is in L^∞. Next lecture, we'll talk about some important inequalities like Markov and Chebyshev involving these spaces.

(We write L^p and L^∞ when the measure space is clear from context.)

§9 October 3, 2024

The previous lecture, we finished proving the basic properties of integration, including some useful limit theorems. We also mentioned Fubini's theorem, which is very useful; we're going to use it a few times throughout the course. We finished the previous lecture by defining L^p spaces. We're going to use L^p spaces quite often when talking about the convergence of random variables. This lecture, we'll mention a few important inequalities (e.g., Markov, Chebyshev, and so on); we'll state them without proof (since we don't have time), and we'll use them throughout the course (we can also use them on problem sets and the exam). Then we'll talk about some properties of L^p and Hilbert spaces (especially about orthogonal decompositions).

§9.1 Some inequalities

§9.1.1 Markov's inequality

The first inequality is Markov's, which is used very often in probability and statistics; its proof is elementary, but it's still very useful. We'll state it for an arbitrary measure space.

Proposition 9.1 (Markov's inequality)

Let f be a nonnegative measurable function on some measure space $(E, \mathcal{E}, \mu)$, and fix any $\lambda > 0$. Then

$$\mu(\{x \in E \mid f(x) \geq \lambda\}) \leq \frac{\mu(f)}{\lambda}.$$

When μ is a probability measure, this will bound the probability that our random variable is large — if we take λ to be large.

Proof. The proof is elementary — we just make the simple observation that

$$f \geq f \cdot \mathbf{1}_{f \geq \lambda} \geq \lambda \mathbf{1}_{f \geq \lambda}.$$

And now we can integrate — by the monotonicity of the integral, we have

$$\mu(f) \geq \mu(\lambda 1_{\{f \geq \lambda\}}) = \lambda \mu(1_{\{f \geq \lambda\}}) = \lambda \mu(\{f \mid \lambda\})$$

(by the linearity of the integral, and the fact that the integral of an indicator function is just the measure of the space). So then

$$\mu(\{f \geq \lambda\}) \leq \frac{\mu(f)}{\lambda}.$$

□

The proof is simple, but if we imagine μ to be a probability measure and f a random variable, then this inequality bounds the probability that the random variable is large. So that's why this is useful.

§9.1.2 Jensen's inequality

This is the first very useful inequality of the course, but there are other types of inequalities which might be a bit more complicated. Another inequality is Jensen's inequality, which is about convex functions.

Proposition 9.2 (Jensen's inequality)
Let X be an integrable random variable (in a probability space) with values in some interval $I \subseteq \mathbb{R}$. Then for any *convex* function $f: I \to \mathbb{R}$, we have

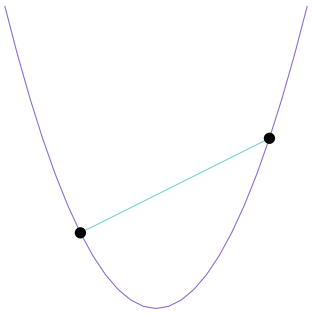
$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

This is a very useful way to get lower bounds on $\mathbb{E}[f(X)]$. For instance, if f is the function $x \mapsto x^2$, then we get $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$, which is a very standard and useful inequality.

Definition 9.3 (Convex function)**.** We say a function $f: I \to \mathbb{R}$ (for an interval I) is convex if for any $t \in [0,1]$ and $x, y \in I$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

We can interpret this graphically:



If we take any two points $(x, f(x))$ and $(y, f(y))$ on the graph of the function, then graphically convexity means that if we draw the segment connecting them, this segment lies above the graph of the function.

§9.1.3 Hölder's inequality

The third inequality, which is quite useful and which we've probably seen in real analysis, is Hölder's inequality. This involves conjugate exponents; we'll first define what this means.

Definition 9.4 (Conjugate)**.** Let $p, q \in [1, \infty]$. We say p and q are conjugate if $\frac{1}{p} + \frac{1}{q} = 1$.

As usual, we use the convention $\frac{1}{\infty} = 0$.

Proposition 9.5 (Hölder's inequality)
Let $p, q \in (1, \infty)$ be conjugates. Then for any measurable functions f and g, we have

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

(We're using the notation from last class.)

A special case of this is the Cauchy–Schwarz inequality — if $p = q = 2$, then you get the Cauchy–Schwarz inequality.

Student Question. *Don't we need to require f and g to be integrable, rather than measurable?*

Answer. No, because we're taking absolute values; the integral of a nonnegative measurable function is always defined, though it might be ∞.

§9.1.4 Minkowski's inequality

The last inequality we're going to use is Minkowski's inequality.

Theorem 9.6 (Minkowski's inequality)
Let $p \in [1, \infty]$, and suppose that f and g are measurable. Then $\|f+g\|_p \le \|f\|_p + \|g\|_p$.

This is a very useful inequality which comes from Hölder's inequality plus a clever trick.

In this course, we're not interested in the proofs of these inequalities, but they're very useful to know and be able to use.

Student Question. *How is $\mathbb{E}[X]$ defined?*

Answer. When we take the expectation of a random variable, you always live in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and you take the integral with respect to $\mathbb{P}$. So we actually define $\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega)$. We have X defined on some abstract measure space, and we're taking the integral with respect to that. (This is the advantage of the general notion of the integral from previous lectures, that we can take an integral in any abstract space.) This integral is the same as $\int_{\mathbb{R}} y \, d\mu_X(y)$ (this is equivalent by definition, since μ_X is the pushforward measure).

§9.2 Facts about L^p spaces

We'll now briefly review some facts about L^p spaces, without proof (these are typically discussed in functional analysis).

§9.2.1 Normed spaces

(We'll assume familiarity with vector spaces; we also briefly mentioned that when stating the monotone class theorem.)

Definition 9.7 (Norm on a vector space). Let V be a vector space. A norm on V is a function $\|\bullet\| : V \to \mathbb{R}_{\ge 0}$ such that:

(i) (Triangle inequality) $\|u+v\| \le \|u\| + \|v\|$ for all $u, v \in V$.

(ii) (Scaling) $\|au\| = |a| \, \|u\|$ for all $a \in \mathbb{R}$ and $u \in V$.

(iii) (Positive definiteness) If $\|u\| = 0$, then $u = 0$.

Last lecture, we defined

$$\|f\|_p = \left(\int_E |f|^p \, d\mu \right)^{1/p}.$$

Is this a norm? It satisfies (i) by Minkowsi's inequality, and it clearly satisfies (ii) (by the scaling properties of the integral). But it does not satisfy (iii). It is true that $\|0\|_p = 0$, but if $\|f\|_p = 0$, this does not necessarily imply $f = 0$. What it does imply is that $f = 0$ μ-almost everywhere.

So in order to make this be a norm, we need to modify the space we're considering. We modify it in the following way.

Definition 9.8. For $f, g \in L^p$, we say that $f \sim g$ (i.e., f is equivalent to g) if $f - g = 0$ μ-almost everywhere. For every $f \in L^p$, we define

$$[f] = \{g \in L^p \mid g - f = 0 \ \mu\text{-almost everywhere}\}$$

as the equivalence class of f under this equivalence relation.

Now we're ready to make L^p a normed space, by defining a slightly modified space.

Definition 9.9. We define $\mathcal{L}^p(E) = \{[f] \mid f \in L^p\}$.

In this modified space, the pth norm as defined last lecture is a norm — because if an element of L^p has 0 norm, then it belongs to the equivalence class of 0 in $\mathcal{L}^p$.

In this class, we won't really distinguish between $\mathcal{L}^p$ and L^p, because from a measure theoretic point of view they're the same (but from a set-theoretic point of view, they're not the same).

§9.2.2 Banach spaces

One important property of $\mathcal{L}^p$ is that it's complete — this means if we take any Cauchy sequence, it converges. We're going to review this quickly for reference. Specifically, complete vector spaces are also called *Banach spaces*. These are studied in functional analysis, but we won't go deeply into them here.

Definition 9.10 (Banach space)**.** A normed vector space $(V, \|\bullet\|)$ is called complete (or a Banach space) if every Cauchy sequence converges — in other words, if (v_n) is a sequence in V with the property that $\|v_n - v_m\| \to 0$ as $m, n \to \infty$ (i.e., (v_n) is a Cauchy sequence), then there exists $v \in V$ with the property that $\|v_n - v\| \to 0$ as $n \to \infty$ (i.e., v_n converges go v with respect to $\|\bullet\|$).

So if every Cauchy sequence has a limit, then our space is called complete.

Example 9.11

The space $\mathbb{R}$ is complete.

Example 9.12

The space $(0, 1)$ is not complete — the sequence $\frac{1}{n}$ is Cauchy, but it doesn't converge in this space (it tends to 0, which is not in $(0, 1)$).

We mention this because L^p is a Banach space, which is our next theorem; we'll state this without proof.

Theorem 9.13

Let $1 \leq p \leq \infty$. Then $\mathcal{L}^p$ is a Banach space.

There's been lots of research into these spaces, especially in the previous century (e.g., in functional analysis).

§9.3 The case $p = 2$ and Hilbert spaces

We'll now consider the case $p = 2$. Here the space has an exceptionally nice structure, because you can define an inner product (just as in $\mathbb{R}^2$), which is induced by the 2-norm.

Definition 9.14. For $p = 2$, we define the inner product for $f, g \in L^2$ as

$$\langle f, g \rangle = \int_E f \cdot g \, d\mu.$$

First, we might ask, is this a well-defined number? Does this actually make sense? The answer is yes, by the useful inequalities we've learned so far, e.g., Hölder's inequality — Hölder's inequality says that $\int |fg|$ is finite, because $\|f\|_2$ and $\|g\|_2$ are finite.

And this inner product induces the 2-norm, because $\|f\|_2 = \langle f, f \rangle^{1/2}$. This means in the special case $p = 2$, our norm is induced by an inner product; this does not happen when $p \neq 2$ (this is not obvious), which is why $p = 2$ is special. And this gives a very nice structure to $\mathcal{L}^2$ — in particular, it makes it a *Hilbert space*.

Definition 9.15 (Hilbert space)**.** A Hilbert space is a vector space with a complete inner product.

What we mean by a complete inner product is that you have your inner product, and it induces a norm by $\|f\| = \langle f, f \rangle^{1/2}$, and the space under this norm is complete.

So $\mathcal{L}^2$ is not just a Banach space, but also a Hilbert space.

§9.3.1 Orthogonal decomposition

Hilbert spaces are somehow nicer than Banach spaces because of the inner product. One very nice thing you can do there is orthogonal complements. First we'll define orthogonality, and then we'll talk about orthogonal complements.

Definition 9.16. Two functions $f, g \in \mathcal{L}^2$ are orthogonal if $\langle f, g \rangle = 0$.

This is a generalization of the Euclidean geometry fact that states that if we have two vectors with angle $90°$, then their inner product is 0. In math, people always start studying simple cases, and from there they build entire theories; that's the base of the theory behind Hilbert spaces.

Definition 9.17 (Orthogonal complements)**.** For any $V \subseteq \mathcal{L}^2$, the orthogonal complement of V is

$$V^\perp = \{f \in \mathcal{L}^2 \mid \langle f, v \rangle = 0 \text{ for all } v \in V\}.$$

So we can talk about orthogonal complements of entire sets of functions. Of course, this definition makes sense for any vector space with an inner product; you can also define this in spaces that aren't complete. But the point is that if you have completeness, then you have some very nice properties of orthogonal complements — in particular, we'll see that if you start with a closed subspace, then you can decompose your original space as a direct sum of the closed space and its orthogonal complement.

First, we'll define what we mean by a closed space.

Definition 9.18 (Closed subspace)**.** Let $V \subseteq L^2$ be a (vector) subspace. Then we say V is closed if whenever we take a sequence $(f_n) \subseteq V$ with the property that $f_n \to f$ for some $f \in L^2$ (with respect to $\|\bullet\|_2$), then there exists $v \in V$ with $v \sim f$.

This is consistent with the topological notion of closedness.

The main theorem we're going to prove is orthogonal decomposition.

Theorem 9.19

Let V be a closed subspace of L^2. Then each $f \in L^2$ has an orthogonal decomposition $f = u + v$, where $v \in V$ and $u \in V^\perp$. Moreover, $\|f - v\|_2 \leq \|f - g\|_2$ for all $g \in V$.

So every element of L^2 can be decomposed in the form $u + v$, and v is actually the element of V with minimum distance from f (in particular, this minimum distance is attained).

To prove this, we'll use two nice, simple identities — the Pythagoras identity and the parallelogram law (which can be generalized to abstract Hilbert spaces).

Lemma 9.20 (Pythagoras identity)

We have $\|f + g\|_2^2 = \|f\|_2^2 + \|g\|_2^2 + 2\langle f, g\rangle$.

This is a generalization of the simple facts that we have in two dimensions.

Lemma 9.21 (Parallelogram law)

We have $\|f + g\|_2^2 + \|f - g\|_2^2 = 2\|f\|_2^2 + 2\|g\|_2^2$.

We've probably seen the analogous identities if we imagine f and g as vectors in $\mathbb{R}^2$; but they hold for an arbitrary Hilbert space.

Now that we have these nice lemmas, we're going to prove the theorem.

Proof of Theorem. Fix $f \in \mathcal{L}^2$; now we're looking for $v \in V$ such that f has this decomposition. How do we look for this element? If we didn't have the second property we'd get lost, but this second property suggests which value of v we need.

First, we can take $(g_n) \subseteq V$ such that

$$\|f - g_n\|_2 \to \operatorname{dist}(f, V),$$

where by $\operatorname{dist}(f, V)$ we mean $\inf_{g \in V} \|f - g\|_2$. This is because we want v to be the element at which this infimum is attained; so we take such a sequence, and we're going to show that this sequence converges and define v as its limit (which will mean the infimum is attained).

In order to show that (g_n) converges, since we're in a complete vector space, it suffices to show that it is Cauchy. And we are going to do this using the lemmas above.

To do so, set $u_n = f - g_n$ for all $n \in \mathbb{N}$. We're going to show that (u_n) is Cauchy using the parallelogram law. By the parallelogram law, we have

$$\|u_n + u_m\|_2^2 + \|u_n - u_m\|_2^2 = 2(\|u_n\|^2 + \|u_m\|^2).$$

By the particular choice of u_n, this implies

$$\left\|2\left(f - \frac{g_n + g_m}{2}\right)\right\|_2^2 + \|g_m - g_n\|_2^2 = 2(\|f - g_n\|_2^2 + \|f - g_m\|_2^2).$$

Our goal was to bound $\|g_m - g_n\|_2^2$, so we'll rearrange this to get

$$\|g_n - g_m\|_2^2 = 2\left(\|f - g_n\|_2^2 + \|f - g_m\|_2^2\right) - 4\left\|f - \frac{g_n + g_m}{2}\right\|_2^2.$$

We wanted to show the left-hand side tends to 0, so it's enough to show that the right-hand side does. And to do this, the first two terms converge to $\operatorname{dist}(f, V)$. Meanwhile, the third term satisfies

$$\left\| f - \frac{g_n + g_m}{2} \right\|_2 \geq \operatorname{dist}(f, V).$$

This is because V is a vector space, so $g_n + g_m$ is too, and therefore so is $\frac{g_n+g_m}{2}$. So we get that

$$\|g_n - g_m\|_2^2 \leq 2(\|f - g_n\|_2^2 + \|f - g_m\|_2^2) - 4(\operatorname{dist}(f, V))^2.$$

And finally, $\|f - g_n\|_2^2$ and $\|f - g_m\|_2^2$ both tend to $\operatorname{dist}(f, V)^2$ as $m, n \to \infty$, which means this entire term tends to 0; and since $\|g_n - g_m\|_2^2$ is nonnegative, this means it must also tend to 0. So we get $\|g_n - g_m\|_2 \to 0$ as $n, m \to \infty$.

And since we've found a Cauchy sequence in a closed subspace, we obtain that there exists $v \in V$ such that $g_n \to v$. So now we have our candidate v. Moreover, we also know that $\|f - g_n\|_2 \to \|f - v\|_2$, because $g_n \to v$ so $f - g_n \to f - v$. So this means

$$\|f - v\|_2 = \operatorname{dist}(f, V).$$

In particular, we have found v such that the second statement is true. It remains to show that we get an orthogonal decomposition — i.e., that if we take $u = f - v$, then $u \in V^\perp$. (Note that we crucially used completeness in going from the fact that (g_n) is complete to saying that it converges.)

So we set $u = f - v$; and we need to show that $u \in V^\perp$ (and then we'll be done). In order to show this, we need to show that for any fixed $h \in V$, we have $\langle u, h \rangle = 0$. For that, we use the following trick: consider some $t \in \mathbb{R}$. Then we have

$$\operatorname{dist}(f, V)^2 \leq \|f - (v + th)\|_2^2 .$$

This is true for *every* t, by the definition of the distance (because $v + th$ is also in V, by the fact that V is a vector space). Now we can use the Pythagoras identity to expand this — this gives

$$\operatorname{dist}(f, V)^2 = \|f - v\|_2^2 + t^2 \|h\|_2^2 - 2t\langle f - v, h \rangle.$$

Now we have a nice quadratic, because we've shown that this inequality holds for *all* $t \in \mathbb{R}$. And a quadratic can be minimized — it's minimized when we set

$$t = \frac{\langle f - v, h \rangle}{\|h\|_2^2}.$$

But we also know that this quadratic is minimized only when $t = 0$, by the property that v is the element of V that minimizes $\|f - v\|$ (specifically, $\operatorname{dist}(f, V) = \|f - v\|_2^2$). So it follows that

$$\frac{\langle f - v, h \rangle}{\|h\|_2^2} = 0,$$

and therefore $\langle f - v, h \rangle = 0$.

And this is true for all $h \in V$, so $f - v \in V^\perp$, and that's the end of the proof. □

Student Question. *What does it mean when we write $n, m \to \infty$?*

Answer. Here when we write $\|g_n - g_m\| \to 0$ as $m, n \to \infty$ (in the definition of a Cauchy sequence), we mean that for every $\varepsilon > 0$, we can find $N \in \mathbb{N}$ such that $\|g_n - g_m\| < \varepsilon$ for every $m, n \geq N$. (So we want this distance to be small for *all* m and n which are sufficiently large.)

Why do we care about $p = 2$? This will be useful later when we study conditional expectations — conditional expectations has a nice geometric meaning using orthogonal decomposition, which we're going to elaborate on when we learn more about conditional expectations of random variables. All this stuff comes from functional analysis, and we don't have to know all of it; but it's nice to have a rough idea of what's going on.

Student Question. *Are we working in L^2 or $\mathcal{L}^2$?*

Answer. Probably $\mathcal{L}^2$. Measure-theoretically it doesn't make a difference, because from our point of view, if two functions differ in a set of measure 0 then they're the same for our purposes. But to be completely precise (e.g., when talking about completeness) we need to talk about the space with the norm.

§9.4 The L^1-convergence of random variables

We'll now talk about convergence of random variables in the space L^1.

Question 9.22. Suppose we have a sequence of random variables (X_n) such that $X_n \to X$ in probability as $n \to \infty$. Under what assumptions do we have $X_n \to X$ in $\mathcal{L}^1$, i.e., $\mathbb{E}[|X_n - X|] \to 0$?

(We defined convergence in probability earlier; this means $\mathbb{P}[|X_n - X| \geq \varepsilon] \to 0$ as $n \to \infty$, for every fixed $\varepsilon > 0$.)

Clearly if we have L^1 convergence, then we have convergence in probability by Markov's inequality — because

$$\mathbb{P}[|X_n - X| \geq \varepsilon] \leq \frac{\mathbb{E}[|X_n - X|]}{\varepsilon} \to 0$$

as $n \to \infty$ (if we assume L^1 convergence). So L^1 convergence is stronger than convergence in probability.

One might ask whether the converse is true. The answer is no, as seen by the following example.

Example 9.23

Take $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}((0,1)), \text{Lebesgue})$, and consider

$$X_n = n \cdot \mathbf{1}_{(0,1/n)}.$$

Then clearly $X_n \to 0$ almost surely as $n \to \infty$ (because if we fix any $x \in (0,1)$, for all sufficiently large n we have $x \notin (0, 1/n)$). In particular, this means X_n converges to 0 in probability.

But we do *not* have L^1 convergence — we have

$$\mathbb{E}[X_n] = n\mu((0,1/n)) = 1 \not\to 0.$$

So X_n does not converge to 0 in L^1.

This means L^1 convergence is stronger than convergence in probability in general — convergence in probability does not imply convergence in L^1. But under what conditions does it imply this?

One reasonable assumption one can use is if all the variables X_n are uniformly bounded (by some constant). If that's the case and you have convergence in probability, then you do have convergence in L^1. (In the above example, clearly they're not bounded, because $n \to \infty$.)

Theorem 9.24 (Bounded convergence theorem)
Suppose that X and (X_n) are random variables. Suppose that there exists some (fixed, finite) constant $C > 0$ with the property that $|X_n| \le C$ for all n. If $X_n \to X$ in probability as $n \to \infty$, then $X_n \xrightarrow{L^1} X$.

So if our random variables are bounded uniformly, then we do have L^1 convergence. In the previous example, we didn't have a uniform bound on the random variables — they could take arbitrarily big values as n grew.

Proof. The first step is to show that X satisfies the same bound $|X| \le C$. This is not very hard — fix $\varepsilon > 0$. Then

$$\mathbb{P}[|X| > C + \varepsilon] \le \mathbb{P}[|X - X_n| + |X_n| > C + \varepsilon]$$

by the triangle inequality (which states $|X| \le |X - X_n| + |X_n|$). And by the subadditivity of the measure, this is at most

$$\mathbb{P}[|X - X_n| > \varepsilon] + \mathbb{P}[|X_n| > C].$$

But the second term is 0, and the first term tends to 0 as $n \to \infty$ (because we have convergence in probability). In particular, this means $\mathbb{P}[|X| > C + \varepsilon] = 0$. And this is true for all $\varepsilon > 0$, so this implies $|X| \le C$ almost surely (i.e., with probability 1). So X is also bounded by C.

Now to complete the proof, we can somehow apply the dominated convergence theorem in order to extract L^1 convergence (that's why we needed this bound). So we fix $\varepsilon > 0$. Then we have

$$\mathbb{E}[|X_n - X|] = \mathbb{E}[|X_n - X| \cdot 1_{|X_n - X| \ge \varepsilon}] + \mathbb{E}[|X_n - X| \cdot \mathbf{1}_{\{|X_n - X| < \varepsilon\}}].$$

The first term is bounded by $2C\mathbb{P}[|X_n - X| \ge \varepsilon]$, and the second term is bounded by ε (because this term is only nonzero in which case $|X_n - X| < \varepsilon$, and in that case it's something at most ε multiplied by 1). So we get the bound

$$2C\mathbb{P}[|X_n - X| \ge \varepsilon] + \varepsilon,$$

which tends to ε as $n \to \infty$.

In particular, what we have shown is that

$$\limsup_{n \to \infty} \mathbb{E}[|X_n - X|] \le \varepsilon,$$

for every $\varepsilon > 0$. This implies that the lim sup is 0; and since this is a nonnegative random variable, the lim sup equals the lim inf equals 0. So we get $\mathbb{E}[|X_n - X|] \to 0$, as desired. □

So under the condition of having a uniform bound, convergence in probability and in L^1 are equivalent; but this is not true in general.

Next lecture, we're going to study uniformly integrable random variables.

Remark 9.25. In this proof, we only need $|X_n| \le C$ almost surely (but they're essentially the same for our purposes).

§10 October 8, 2024

Last lecture, we mentioned some basic definitions and facts about L^p spaces and Hilbert spaces. At the end of the lecture, we discussed the relationship between convergence in probability and in L^1. We mentioned convergence in L^1 implies convergence in probability, but the converse is not true; we gave a counterexample last time. Then we tried to come up with certain conditions that guarantee convergence in probability and

convergence in L^1 are independent. We saw that if we assume all the random variables are bounded uniformly, then this is true.

But we can do better than that. We don't need the functions to be bounded, we can instead say that they're not *concentrated* in arbitrarily small regions of our probability space. To make this rigorous, we need the notion of uniform integrability.

§10.1 Uniform integrability

Definition 10.1 (Uniform integrability)**.** Let $\mathcal{X}$ be a family of random variables. Given $\delta > 0$, define

$$I_{\mathcal{X}}(\delta) = \sup\{\mathbb{E}[X \cdot \mathbf{1}_A] \mid X \in \mathcal{X},\ A \in \mathcal{F} \text{ with } \mathbb{P}[A] < \delta\}.$$

We say that $\mathcal{X}$ is uniformly integrable if $\mathcal{X}$ is L^1-bounded (meaning $\|X\|_1$ is bounded over all $X \in \mathcal{X}$) and $I_{\mathcal{X}}(\delta) \to 0$ as $\delta \to 0$.

To explain the intuition behind the definition of $I_{\mathcal{X}}(\delta)$, we need a uniform bound over all elements of $\mathcal{X}$, of the expectation of our random variable when restricted to sets of small probability — we don't want our random variables to be concentrated on small-probability sets. We'll see this plays a role for the equivalence of convergence of probability and in L^1.

More explicitly, here's what we mean by L^1 (or more generally L^p) bounded random variables.

Definition 10.2 (L^p-bounded)**.** We say a family $\mathcal{X}$ of random variables is L^p-bounded if

$$\sup\{\|X\|_p \mid X \in \mathcal{X}\} < \infty.$$

In some sense, uniform integrability is the analogous of uniform continuity but for integration.

A simple result is the following (it's almost immediate from the definition, and is left as an exercise).

Definition 10.3. Finite unions of uniformly integrable families of random variables are also uniformly integrable (i.e., if $\mathcal{X}_1, \dots, \mathcal{X}_n$ are all uniformly integrable, then so is $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_n$).

So if we take any finite collection of families of random variables, then their union will also be uniformly integrable. For instance, if you have a random variable by itself and it is L^1-bounded, then it's uniformly integrable; so if you take finitely many L^1-bounded random variables, then it'll also be uniformly integrable. But we can also construct more elaborate families.

Proposition 10.4

Let $\mathcal{X}$ be a L^p-bounded family of random variables, for some $p > 1$. Then $\mathcal{X}$ is uniformly integrable.

So any family which is L^p-bounded will also be uniformly integrable; this lets us construct infinite families of random variables which are uniformly bounded.

Proof. The proof is Hölder's inequality, and in the spirit of the proof we've done so far. We define $C = \sup\{\|X\|_p \mid X \in \mathcal{X}\} < \infty$ (by the assumption $\mathcal{X}$ is L^p-bounded). We need to show $I_{\mathcal{X}}(\delta) \to 0$. To do so, suppose we fix some $X \in \mathcal{X}$ and some $A \in \mathcal{F}$. Then we have

$$\mathbb{E}[|X| \cdot \mathbf{1}_A] \le \mathbb{E}[|X|^p]^{1/p}(\mathbb{P}[A])^{1/q}$$

where q is the conjugate of p, by Hölder's inequality; and this is at most $C(\mathbb{P}[A])^{1/q}$.

This is a uniform bound depending only on $\mathbb{P}[A]$. And this means

$$I_{\mathcal{X}}(\delta) \leq C\delta^{1/q},$$

which tends to 0 as $\delta \to 0$. □

So this shows that L^p-bounded families of random variables are indeed uniformly integrable. But here it is crucial that $p > 1$; this wouldn't be true if $p = 1$.

Example 10.5

Consider the family of random variables $X_n = n \cdot \mathbf{1}_{(0,1/n)}$. Then (X_n) is clearly L^1-bounded, because $\mathbb{E}[|X_n|] = 1$ for all n. But (X_n) is *not* uniformly integrable. To see this, we have

$$I_{\mathcal{X}}\left(\frac{1}{n}\right) \geq \mathbb{E}[|X_n| \cdot \mathbf{1}_{(0,1/n)}] = 1,$$

which doesn't go to 0 as $\delta \to 0$.

So it is crucial that $p > 1$ in the previous proposition.

Student Question. *L^p bounds are monotonic, right — if a function is L^p bounded, then is it also L^q bounded for all $q \leq p$?*

Answer. Yes.

The main topic for today will be studying uniformly integrable families of random variables, which will help us get conditions for the equivalence of convergence in probability and in L^1.

For many purposes, it'll be useful to rephrase the definition of uniform integrability; the following lemma lets us do so.

Lemma 10.6

The family of random variables $\mathcal{X}$ is uniformly integrable if and only if

$$\sup\{\mathbb{E}[|X| \cdot \mathbf{1}_{\{|X|>k\}}] \mid X \in \mathcal{X}\} \to 0$$

as $k \to \infty$.

Proof. For the forwards direction, we assume uniform integrability, and we want a bound on this supremum. For that, we use Markov's inequality: fix $X \in \mathcal{X}$ and $k \in \mathbb{N}$. Then we have

$$\mathbb{P}[|X| > k] \leq \frac{\mathbb{E}[|X|]}{k}$$

by Markov's inequality. And this is at most

$$\frac{\sup\{\mathbb{E}[|Y|] \mid Y \in \mathcal{X}\}}{k}.$$

(This supremum exists because $\mathcal{X}$ is L^1-bounded.) In particular, this tends to 0 as $k \to \infty$.

Now gien any ε¡ choose δ small enough such that

$$\mathbb{E}[|X| \cdot \mathbf{1}_A] < \varepsilon$$

for all $X \in \mathcal{X}$ and $A \in \mathcal{F}$ with $\mathbb{P}[A] < \delta$. Then pick $k \in \mathbb{N}$ sufficiently large that

$$k\delta > \sup\{\mathbb{E}[|X|] \mid X \in \mathcal{X}\}.$$

This implies $\mathbb{P}[|X| > k] < \delta$ (by the above bound). And then $\mathbb{E}[|X| \cdot \mathbf{1}_{|X|>k}] < \varepsilon$ (for every $X \in \mathcal{X}$). This is exactly a rephrase of what we wanted to show.

(The crucial thing here is that we can first bound $\mathbb{P}[|X| > k]$ uniformly in X, and then we use the definition of uniform integrability.)

Now for the other direction, we assume that the stated condition holds, and somehow we need to extract uniform integrability. Here, we want to show two things: that $\mathcal{X}$ is L^1-bounded, and that $I_\chi(\delta) \to 0$ as $\delta \to 0$.

Let's first check that it's L^1-bounded. This is not hard to see — by the linearity of the integral, we can write

$$\mathbb{E}[|X|] = \mathbb{E}[|X| \cdot \mathbf{1}_{|X|\leq k}] + \mathbb{E}[|X| \cdot \mathbf{1}_{|X|>k}].$$

The first term is at most k; we'll leave the second term as it is. We know this term is always finite, because k is some number and the second term tends to 0 as $k \to \infty$, so in particular it is some finite number. Specifically, we have a uniform bound

$$\mathbb{E}[|X|] \leq k + \sup_{m\in\mathbb{N}}\{\mathbb{E}[|Y|\,\mathbf{1}_{|Y|>m}]\} < \infty.$$

Now for the second condition, we can perform a similar trick — we write

$$\mathbb{E}[|X|\,\mathbf{1}_A] = \mathbb{E}[|X|\,\mathbf{1}_{A\cap\{|X|>k\}}] + \mathbb{E}[|X|\,\mathbf{1}_{A\cap\{|A|\leq k\}}].$$

The first term is at most $\mathbb{E}[|X|\,\mathbf{1}_{\{|X|>k\}}]$, and the second term is at most $k\mathbb{P}[A]$. So for given $\varepsilon > 0$, we can pick k such that the first term is at most $\varepsilon/2$ (since this sequence converges to 0 uniformly as $k \to \infty$). And then we can pick $\delta > 0$ such that $k\delta < \varepsilon/2$. If we take any event A with $\mathbb{P}[A] < \delta$, then this choice of δ ensures the second term is also less than $\varepsilon/2$, and so overall we get $\mathbb{E}[|X| \cdot \mathbf{1}_A] < \varepsilon$ for all $X \in \mathcal{X}$ and A witih $\mathbb{P}[A] < \delta$. And since ε was arbitrary, this completes the proof. □

As a corollary, we have the following.

Corollary 10.7

Let $\mathcal{X} = \{X\}$ be a family consisting of a single random variable $X \in L^1$. Then $\mathcal{X}$ is uniformly integrable. Hence, any finite collection of L^1 random variables is uniformly integrable.

(The second statement follows from the first, because of the proposition that finite unions of uniformly integrable families are still uniformly integrable.)

Proof. We'll prove the first statement using the previous lemma. Here we have

$$\mathbb{E}[|X|\,\mathbf{1}_{|X|\geq N}] = \sum_{m\geq N} \mathbb{E}[|X|\,\mathbf{1}_{m\leq|X|<m+1}].$$

The point is that this tends to 0 as $N \to \infty$, because

$$\mathbb{E}[X] = \sum_{n\geq 1} \mathbb{E}[|X|\,\mathbf{1}_{\{n\leq|X|<n+1\}}].$$

(So this series converges because $X \in L^1$, which means its tail has to tend to 0 as $n \to \infty$.) This means the condition of the previous lemma is satisfied, so a single random variable is uniformly integrable. □

§10.2 Convergence in probability vs. L^1

Now we know that a collection of finitely many L^1 random variables is a uniformly integrable family. And we're ready to characterize when convergence in probability and in L^1 are equivalent — we'll show that this is the case if and only if we have uniform integrability. (This is the next theorem; we'll see applications when we study martingales in a few weeks.)

Theorem 10.8

Leet X and (X_n) be random variables. Then the following are equivalent:

(i) We have $X, X_n \in L^1$ and $X_n \to X$ as $n \to \infty$ in L^1.

(ii) (X_n) is uniformly integrable and $X_n \to X$ in probability.

So if we have a collection of random variables and we know they're uniformly integrable, then convergence in probability and L^1 are equivalent. This generalizes what we proved last time — we proved the equivalence when all X_n's are uniformly bounded, but any such family is also L^p-bounded and therefore uniformly integrable.

Proof. The forwards direction is a standard manipulation. The backwards direction is more tricky — we use uniform integrability to cut off X_n and X at a large value k, and then use the dominated convergence theroem.

For (i) to (ii), we'll first show convergence in probability. Here we'll apply the definition and Markov's inequality — if we fix any $\varepsilon > 0$, then we have

$$\mathbb{P}[|X_n - X| > \varepsilon] \le \frac{\mathbb{E}\,|X_n - X|}{\varepsilon} \to 0$$

as $n \to \infty$. So this implies $\mathbb{P}[|X_n - X| > \varepsilon] \to 0$ as $n \to \infty$, which implies the convergence in probability. (As you can see, we've extensively used Markov's and Hölder's inequality and several others.)

Now we need to prove uniform integrability. For this, we again fix some $\varepsilon > 0$. Then we take N sufficiently large such that $\mathbb{E}[|X_n - X|] \le \varepsilon/2$ for all $n \ge N$. (We can always do this because this sequence converges to 0 as $n \to \infty$, by assumption.)

Now we have that $\{X, X_1, X_2, \dots, X_{N-1}, X_N\}$ is a family of finitely many random variables, so we have uniform integrability. And since this family is uniformly integrable, we can pick some small number $\delta > 0$ with the property that for any $A \in \mathcal{F}$ with $\mathbb{P}[A] < \delta$, we have $\mathbb{E}[Y \cdot \mathbf{1}_A] < \delta$ for all Y in the family. In particular, this means $\mathbb{E}|X| \cdot \mathbf{1}_A, \mathbb{E}|X_n|\,\mathbf{1}_A \le \varepsilon/2$ for all $1 \le n \le N$.

Now we need to handle the case $n > N$. For this, by the triangle inequality

$$\mathbb{E}[|X_n|\,\mathbf{1}_A] \le \mathbb{E}[|X_n - X|\,\mathbf{1}_A] + \mathbb{E}[|X|\,\mathbf{1}_A].$$

The first term is at most $\mathbb{E}[|X_n - X|] < \varepsilon/2$ (by how we chose N), and the second is at most $\varepsilon/2$, so this is overall at most ε.

So this implies uniform integrability, that $I_{\mathcal{X}}(\delta) \to 0$.

Now we'll try to prove the other direction. So far, this argument is standard; we've used it a lot of times. But the argument for the other direction is less standard; but eventually we will also become familiar with that kind of argument.

The first step is to show that X is in L^1 — we've only assumed that X_n is uniformly integrable and $X_n \to X$ in probability, so it's not obvious that X is in L^1. If we had almost sure convergence we could use Fatou's lemma, but we don't have this.

But if you have convergence in probability, then you do have almost sure convergence along a subsequence. So we will work along that subsequence to prove that $X \in L^1$. In particular, we know there exists a subsequence of (X_n), which we call (X_{n_k}), with the property that $X_{n_k} \to X$ almost surely as $k \to \infty$. (We can always do this — it's not true in general that the entire sequence converges almost surely, but it is true along a subsequence.)

And now we can apply Fatou's lemma for that sequence. By Fatou's lemma, we have

$$\mathbb{E}[|X|] = \mathbb{E}[\liminf_{k\to\infty} |X_{n_k}|] \le \liminf_{k\to\infty} \mathbb{E}[|X_{n_k}|].$$

But this is finite, because we have assumed that X_n is uniformly integrable and therefore L^1-bounded. So this number is finite; this means X indeed is in L^1.

Now the second thing is to actually show that we have L^1-convergence. We fix $\varepsilon > 0$; we want to show that $\mathbb{E}\,|X_n - X| < \varepsilon$ for large n.

For that, first by uniform integrability, there exists k such that

$$\mathbb{E}[|X|\,\mathbf{1}_{\{|X|>k\}}], \mathbb{E}[|X_n|\,\mathbf{1}_{\{|X_n|>k\}}] \le \frac{\varepsilon}{3}$$

for all $n \in \mathbb{N}$ (by the fact that (X_n) is uniformly integrable, and an earlier lemma).

It's tempting to use the bounded convergence theorem, but it's unclear how to do this — we don't have a function which dominates everything which is in L^1. So this is a time when we need to perform a trick. And the trick is that we're going to cut off X_n at the value k. In particular, we define $X_n^k = (X_n \vee (-k)) \wedge k$, and likewise for X (where $\vee$ and $\wedge$ denote the maximum and minimum). Then we have $X_n^k \to X^k$ in probability for all k, because $X_n \to X$ in probability.

So now we have a sequence of variables which are bounded uniformly and converge in probability, and this implies $X_n^k \to X^k$ in L^1 (by the theorem we proved last lecture, that if we have a family of functions which is bounded uniformly and converge in probability, then they converge in L^1).

Now we need to combine this with the observation from earlier to complete the proof. First, this convergence says we can find $N \in \mathbb{N}$ such that $\mathbb{E}\left|X_n^k - X^k\right| \le \varepsilon/3$ for all $n \ge N$. This implies that for all $n \ge N$, we have

$$\mathbb{E}[|X_n - X|] \le \mathbb{E}[\left|X_n - X_n^k\right|] + \mathbb{E}[|X|\,\mathbf{1}_{|X|\ge k}] + \mathbb{E}[|X_n|\,\mathbf{1}_{|X_n|\ge k}]$$

by the triangle inequality. And now we're done — the first term is at most $\varepsilon/3$, and the second and third terms are also at most $\varepsilon/3$. So overall, this is less than ε, for all $n \ge N$. And that's the end of the proof. □

So we've characterized the equivalence of convergence in probability and L^1.

Student Question. *Why does $X_n^k \to X^k$ in probability?*

Answer. We have $\left|X_n^k - X^k\right| \le |X_n - X|$, so

$$\mathbb{P}[\left|X_n^k - X^k\right| \ge \varepsilon] \le \mathbb{P}[|X_n - X| \ge \varepsilon]$$

for any ε.

This will be especially useful when discussing martingales.

§10.3 The Fourier transform

There's two questions: does a Fourier transform exist? And when can we recover a function from its Fourier transform? The Fourier transform sort of encodes a function, and when the function is nasty, its Fourier transform is often much nicer (e.g., continuous). So that's why Fourier transforms are useful.

When proving things about the Fourier transform, it's useful to think about convolution — intuitively this lets you make a function more smooth.

We'll see that the theory of Fourier analysis will help us prove some very nice properties of random variables.

Definition 10.9 (Fourier transform). The Fourier transform of a function $f: L^1(\mathbb{R}^d)$, denoted $\widehat{f}: \mathbb{R}^d \to \mathbb{C}$, is given by

$$\widehat{f}(u) = \int_{\mathbb{R}^d} f(x) e^{i\langle u,x\rangle}\, dx,$$

where $\langle u, x\rangle = u_1x_1 + \cdots + u_dx_d$.

This is always well-defined, because the function on the inside is in L^1; so this integral will always give you some number (here a complex number).

Why introduce this concept? The point is that many computations are easier for $\widehat{f}$ than f — f itself might be very nasty, while $\widehat{f}$ is better. In particular, it usually behaves much better under differentiation; and there's also convolution, which will be relevant to sums of random variables.

More generally, you can also define the Fourier transform of a measure, not just a function.

Definition 10.10 (Fourier transform of a measure). For a finite measure μ on $\mathbb{R}^d$, its Fourier transform is the function $\widehat{\mu}: \mathbb{R}^d \to \mathbb{C}$ given by

$$\widehat{\mu}(u) = \int_{\mathbb{R}^d} e^{i\langle u,x\rangle}\, \mu(dx).$$

So we're taking the analogous integral, but integrating with respect to μ instead of the Lebesgue measure. This still makes sense because μ is finite, so it gives a finite mass on the entire space, which means the function on the inside is L^1. So this gives a well-defined function.

§10.3.1 Characteristic functions

Now let's see what this means in the context of probability. You can relate them by *characteristic functions*.

Definition 10.11 (Characteristic functions). Let X be a random variable. Then the characteristic function of X is the Fourier transform of its law μ_X, i.e.,

$$\phi_X(u) = \mathbb{E}[e^{i\langle u,x\rangle}] = \widehat{\mu_X}(u).$$

This is the first connection between Fourier analysis and probability — we're going to see all the information about the law of X is encoded into this characteristic function. This characteristic function is always well-defined, because the absolute value of this exponential is always 1 (even if X is not at all integrable, this characteristic function is well-defined). And studying the characteristic function of X allows us to prove some very nice properties.

§10.3.2 Some properties of the Fourier transform

Now we'll prove some easy but important properties of the Fourier transform.

Proposition 10.12

We have $\|\widehat{f}\|_\infty \le \|f\|_1$. Similarly, $\|\widehat{\mu}\|_\infty \le \mu(\mathbb{R}^d)$.

This is immediate from the definition (if we move the absolute values inside the integral). This is good news because we can apply e.g. the dominated convergence theorem.

Another nice property is that the Fourier transform is always continuous — even if you start with a nasty function, you end up with a continuous one.

Proposition 10.13

The Fourier transforms $\widehat{f}$ and $\widehat{\mu}$ are continuous.

Proof. If we have a sequence $u_n \to u$ as $n \to \infty$, then we also have $f(x)e^{i\langle u_n, x\rangle} \to f(x)e^{i\langle u, x\rangle}$ for every $x \in \mathbb{R}^d$. In other words, the functions inside the integral converge *pointwise*.

And now we can apply teh dominated convergence theorem, because we also have that $\left|f(x)e^{\langle u_n, x\rangle}\right| \le |f(x)|$, and we know $|f(x)|$ is in L^1. So we have a collection of functions $f(x)e^{i\langle u_n, x\rangle}$ all dominated by the L^1 function $|f(x)|$, and they converge pointwise to $f(x)e^{i\langle u, x\rangle}$; this means the integrals (over $x \in \mathbb{R}^d$) also converge, and therefore $\widehat{f}(u_n) \to \widehat{f}(u)$. □

You can imagine $\widehat{f}$ is a function encoding f, but with more regularity — for example, you start with an arbitrary function in L^1 and get a continuous function that encodes it.

§10.3.3 Convolutions

To do something useful with Fourier transforms, we also need to talk about convolutions. These will play a crucial role in discussing Fourier analysis, as well as when proving theorems about independence.

Definition 10.14 (Convolution of random variables)**.** Let μ and ν be probability measures. Then their convolution, denoted $\mu * \nu$, is defined as the law of $X + Y$ where X has law μ and Y has law ν, and X and Y are independent.

Explicitly, we have

$$(\mu * \nu)(A) = \mathbb{P}[X + Y \in A] = \int\int \mathbf{1}_A(x + y)\mu(dx)\mu(dy).$$

You can see convolution relies naturally because it's related to the law of sums of independent random variables; and we often want to study the long-range behavior of sums of independent random variables, so convolutions arise naturally. Surprisingly they also arise naturally in Fourier analysis, so they're a bridge between them.

§11 October 10, 2024

Last lecture, we started discussing Fourier transforms. Surprisingly, if you know the characteristic function of a random variable, then you can actually recover its law.

§11.1 Convolutions

We ended the previous lecture by defining the convolution of two measures — it corresponds to the law of the sum of two independent random variables. That's one reason we care about convolutions, because one purpose of this course is to study the sum of independent random variables.

Of course, we can also define a convolution between functions, or a convolution between measures and functions.

Definition 11.1 (Convolution of function with measure)**.** Let $f \in L^p$ (for $p \geq 1$), and let ν be a probability measure. Then the convolution of f with ν is the function $f * \nu$ defined by

$$(f * \nu)(x) = \int_{\mathbb{R}^d} f(x-y)\nu(dy).$$

By Hölder's inequality, this function $(f * \nu)$ is also in L^p.

And of course we can define the convolution between *functions* in a similar way.

Definition 11.2. We define the convolution of two 'nice' functions f and g as

$$(f * g)(x) = \int_{\mathbb{R}^d} f(x-y)g(y)\,dy$$

for each $x \in \mathbb{R}^d$.

Let's actually check that $f * \nu$ is actually in L^p. In fact, we'll prove the following stronger statement.

Proposition 11.3

For any $f \in L^p$ and ν a probability measure, we have $\|f * \nu\|_p \leq \|f\|_p$.

Proof. We have $\|f * \nu\|_p^p = \int_{\mathbb{R}^d} |\int_{\mathbb{R}^d} f(x-y)\nu(dy)|^p\,dx$. We can move the absolute value inside the integral to get

$$\|f * \nu\|_p^p \leq \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |f(x-y)\nu(dy)| \right)^p dx.$$

Now because ν is a probability measure, we can use Hölder's inequality to bring the pth norm inside — this is at most

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-y)|^p\,\nu(dy)\,dx.$$

And then we can apply Fubini's theorem (or even Tonelli's theorem, since this is a nonnegative function) to swap the order of integration. If we do so, then we obtain

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-y)|^p\,dx\,\nu(dy).$$

But the pth norm is translation invariant (so the translation of y doesn't affect it), which means we get

$$\|f\|_p^p \cdot \nu(\mathbb{R}^d) = \|f\|_p^p.$$

This means $f * \nu$ is indeed a L^p function, and its pth norm is bounded by that of f. □

§11.2 Fourier transform of a convolution

The point is that when we take a Fourier transform of the convolution, some very nice things happen.

Proposition 11.4

We have $\widehat{(f * \nu)}(u) = \widehat{f}(u)\widehat{\nu}(u)$ for all $u \in \mathbb{R}^d$.

So the Fourier transform of a convolution is just the product of Fourier transforms. The proof comes from just following the definitions, but the property is quite important.

Proof. By definition, we have

$$\widehat{f * \nu}(u) = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(x-y)\, \nu(dy) \right) e^{i\langle u,x\rangle}\, dx.$$

Now again we can apply Fubini's theorem to change the order of integration — so that we first integrate with respect to x, and then y. Then we get

$$\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(x-y) e^{i\langle u,x\rangle}\, dx \right) \nu(dy).$$

And now let's analyze what we have inside; we can rewrite this as

$$\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(x-y) e^{i\langle u,x-y\rangle} d(x-y) \right) \cdot e^{i\langle u,y\rangle}\, \nu(dy)$$

(by adding and subtracting y from the inner product). (We can go from dx to $d(x-y)$ by the translational invariance of the Lebesgue measure.)

The first term on the inside is, by definition, just $\widehat{f}(u)$ (because the Fourier transform is translation-invariant); so we get

$$\int_{\mathbb{R}^d} \widehat{f}(u) e^{i\langle u,y\rangle} \nu(dy).$$

But the $\widehat{f}(u)$ term is constant, and we end up with $\widehat{f}(u)\widehat{\nu}(u)$ (by the definition of $\widehat{\nu}$). □

So the Fourier transform has very nice behavior on convolutions.

This property also holds when we replace f by another probability measure; that's important for studying sums of independent random variables. We'll state it as a proposition (the proof is quite easy).

Proposition 11.5

Let μ and ν be probability measures, and let X and Y be independent random variables with laws μ and ν, respectively. Then we have

$$\widehat{\mu * \nu} = \widehat{\mu_{X+Y}}(u) = \widehat{\mu}(u)\widehat{\nu}(u).$$

Proof. Again, we just apply the definitions combined with independence. First, last lecture we saw that

$$(\widehat{\mu * \nu})(u) = \mathbb{E}[e^{i\langle u,X+Y\rangle}] = \mathbb{E}[e^{i\langle u,X\rangle} \cdot e^{i\langle u,Y\rangle}].$$

And since X and Y are independent random variables, this is just the product of expectations

$$\mathbb{E}[e^{i\langle u,X\rangle}]\mathbb{E}[e^{i\langle u,Y\rangle}].$$

But by definition, these tow terms are $\widehat{\mu}(u)$ and $\widehat{\nu}(u)$, respectively. □

So this boils down to just the definition of the Fourier transform and the independence of X and Y.

So now we can see why convolutions play an important role in probability and Fourier analysis — because they behave very nicely under Fourier transforms.

§11.3 Fourier inversion

These are the link between convolutions and Fourier analysis. The next topic we'll try to address is the Fourier inversion formula — one of the purposes of Fourier analysis is to start with a nasty function and encode it using a simpler function. That's the Fourier transform, but it's nontrivial to see that the Fourier transform *determines* the function — that if we have two functions with the same Fourier transform, they have to be the same Lebesgue-almost everywhere. In fact, we have something stronger — we have an explicit formula for how to recover the function.

Theorem 11.6 (Fourier inversion formula)
Let $f \in L^1$, and suppose that $\widehat{f} \in L^1$ as well. Then we have

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(u) e^{-i\langle u,x\rangle}\, du$$

for Lebesgue-almost every x.

(We only care about functions up to a set of Lebesgue measure 0.)

You can see that this means whenever we have a Fourier transform, we can recover the full function.

To prove this, we need several steps. We'll follow the same method we used with the Lebesgue integral — we first prove it for simpler functions, then approximate more complicated functions by simpler ones (using the convergence theorems from earlier to justify the change of integrals).

The first step is to show that this formula is true for *Gaussian* random variables. Then we'll show this holds for Gaussian *convolutions*. And the third thing is to show that *any* function is approximated by Gaussian convolutions. So it's a similar approach to integration, but the things we use to approximate are different.

§11.3.1 Gaussian densities

Definition 11.7 (Gaussian density). The Gaussian density with variance t (for any $t > 0$) is the function

$$g_t(x) = \frac{1}{(2\pi t)^{d/2}} \cdot e^{-|x|^2/2t}$$

(for every $x \in \mathbb{R}^d$).

When $t = 1$ and $d = 1$, we get the density of the standard normal distribution; so this is a generalization to higher dimensions. This is equivalently the density of $\sqrt{t} \cdot Z$, where Z is the d-dimensional vector consisting of d independent normal Gaussians — i.e., $Z = (Z_1, \ldots, Z_d)$ where $Z_i \sim \mathcal{N}(0,1)$ are independent. (So that's another way to interpret this function g_t.)

The idea is to first prove 11.6 holds for Gaussian densities. Then we'll show it holds for *convolutions* of such Gaussian densities. And then we'll approximate arbitrary functions by convolutions of functions of this type, and that'll finish the proof.

§11.3.2 Characteristic function of Gaussians

To prove the first step, we actually need to compute the Fourier transform of g_t; so that's the first step. We're first going to compute the characteristic function corresponding to a standard normal in one dimension.

Proposition 11.8

Let $Z \sim \mathcal{N}(0,1)$. Then its characteristic function is given by $\phi_Z(u) = e^{-u^2/2}$.

Proof. By definition, we have

$$\phi_Z(u) = \mathbb{E}[e^{iuZ}]$$

(in one dimension, the inner product is just the standard multiplication of real numbers). By the definition of the density, this is

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iux} \cdot e^{-x^2/2}\, dx.$$

How can we compute this? We know ϕ_Z has this form. One thing we can do is actually compute the integral (e.g., applying Cauchy's residue theorem or other tools from complex analysis). But there's another way — to prove that ϕ_Z is the unique solution to a differential equation, and then use theory from ODEs to find the exact form of that solution.

One observation is that we can differentiate with respect to u (because we have a bounded function inside the expectation, so we can apply the dominated convergence theorem). If we differentiate with respect to u, then we get

$$\phi_Z'(u) = \mathbb{E}[iue^{iuZ}] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} ixe^{ixu} e^{-x^2/2}\, dx.$$

But this is just $-u \cdot \phi_Z(u)$ (this comes from integration by parts). This means ϕ_Z satisfies the nice and simple differential equation

$$\phi_Z'(u) = -u \cdot \phi_Z(u).$$

But we know what the solutions to this ODE are — if this is true, then we know

$$\log(\varphi_Z(u)) = -\frac{u^2}{2} + C$$

(where $C \in \mathbb{C}$ is some constant, and log refers to the complex logarithm). This implies (taking the exponential of both sides) that $\phi_Z(u) = Ae^{-u^2/2}$, where $A \in \mathbb{C}$ is some constant.

Now in order to complete the proof, it suffices to find the constant A. But for this, we can just set $u = 0$; we have

$$A = \phi_Z(0) = \mathbb{E}[e^0] = 1.$$

And this implies $\phi_Z(u) = e^{-u^2/2}$, as desired. □

So we know the exact form of the characteristic function of a Gaussian. There are many other ways to do this proof; for example, you can also compute the integral directly using complex analysis (but this solution is cleaner).

Student Question. *How do we know ϕ_Z is differentiable?*

Answer. You can show this using the bounded convergence theorem.

Now that we have this form, we can actually compute the Fourier transform of g_t.

Proposition 11.9

Let $Z = (Z_1, \dots, Z_d)$ where $Z_i \sim \mathcal{N}(0,1)$ and $Z_1, \dots, Z_d$ are independent. Then $\sqrt{t} \cdot Z$ has density

$$g_t(x) = \frac{1}{(2\pi t)^{d/2}} e^{-|x|^2/2t},$$

and the Fourier transform of g_t is given by

$$\widehat{g_t}(x) = e^{-|u|^2 t/2}$$

for all $u \in \mathbb{R}^d$ and $t > 0$.

So the Fourier transform again has this nice form. We're going to do this by combining the previous result with independence.

Proof. By definition, we have

$$\widehat{g_t}(u) = \mathbb{E}[e^{i\langle u, \sqrt{t}Z\rangle}],$$

where Z is as given. And this is just

$$\mathbb{E}\left[\prod_{j=1}^{d} e^{iu_j\sqrt{t}Z_j}\right]$$

And since the Z_j's are independent, the expectation of the product is the product of expectations; this means

$$\widehat{g_t}(u) = \prod_{j=1}^{d} \mathbb{E}[e^{iu_j\sqrt{t}Z_j}] = \prod_{j=1}^{d} \phi_{Z_j}(\sqrt{t}u_j).$$

But we know this, since the Z_j's are all standard Gaussians; so this is

$$\prod_{j=1}^{d} e^{-tu_j^2/2} = e^{|u|^2 t/2}.$$

□

Remark 11.10. By definition $\widehat{g_t}(u) = \int_{\mathbb{R}^d} g_t(x) \cdot e^{i\langle u,x\rangle}\, dx$. But we also know g_t is the density of Z. So this is just $\mathbb{E}[e^{i\langle u, \sqrt{t}Z\rangle}]$ by the definition of the expectation.

Now let's prove the Fourier inversion theorem in the case of g_t; we can do this because we have an explicit description of the Fourier transform.

Proof. First, we can write $\widehat{g_t}$ as

$$\widehat{g_t}(u) = (2\pi)^{d/2} t^{-d/2} \cdot g_{1/t}(u),$$

just by rearranging the picture with teh explicit form of th Fourier tranfsorm of g_t. If we take the Fourier transform on both sides, this implies that

$$\widehat{\widehat{g_t}}(u) = (2\pi)^d g_t(u)$$

(just by doing some manipulation). This is not exactly the same as saying the Fourier inversion formula works, because there we're integrating against $e^{-\langle u,x\rangle}$, and not $e^{i\langle u,x\rangle}$. But this is fine — tbe Gaussian distribution has symmetry.

Writing down the above equality as an integral (so that we can see more clearly that it's almost the Fourier inversion formula), we have

$$g_t(u) = \frac{1}{(2\pi)^d}\widehat{\widehat{g}_t}(u) = \frac{1}{(2\pi t)^d}\int_{\mathbb{R}^d} \widehat{g}_t(x)e^{i\langle u,x\rangle\, dx}.$$

This is almost the Fourier inversion formula, except that the Fourier inversion formula has $-i$ instead of $+i$. But for that, we can just use symmetry — we have $g_t(u) = g_t(-u)$, and if we replace u with $-u$ in t his formula, then we just get

$$g_t(u) = g_t(-u) = -\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \widehat{f}_t(x)e^{-i\langle u,x\rangle}\, dx. \qquad \square$$

Writing down this as a lemma (which we just showed):

Lemma 11.11

The Fourier inversion formula holds for the Gaussian density function.

So that's the first step of our proof of the Fourier inversion formula for arbitrary L^1 functions.

§11.4 Theorem 11.6 for Gaussian convolutions

The next step is to show theinversion formula holds for $f * g_t$; then we'll prove the general case by approximating with functions of this form.

Definition 11.12 (Gaussian convolution)**.** Let $f \in L^1$. Then a Gaussian convolution of f is a function of the form $f * g_t$ for some $t > 0$.

So we're taking any g_t (as above) and convolving it with f. The point is that we're going to prove the Fourier inversion formula holds for functions of this form; and that any function f can be approximated by functions of this form.

Before we do this, we'll need some nice bounds on the L^1 and L^∞ norms of this function. First, we have

$$\|f * g_t\|_1 \le \|f\|_1$$

by definition. So $f * g_t$ is L^1-bounded. We also have pointwise bounds, in the sense that

$$|(f * g_t)(x)| = \int_{\mathbb{R}^d} f(x-y)e^{-|y|^2/2t} \cdot \frac{1}{(2\pi t)^{d/2}}\, dy.$$

But the right-hand side is at mot

$$(2\pi t)^{-d/2}\int_{\mathbb{R}^d} |f(x-y)|\, dy.$$

And by the translational invariance 0f the Lebesgue measure, this is just $(2\pi t)^{-1/2}\|f\|_1$. This in particular implies

$$\|f * g_i\|_\infty \le (2\pi t)^{-d/2}\|f\|_1.$$

As $t \to 0$, this bound gets worse and worse, as we can see from the formula. Why? As $t \to 0$, we'll se that this approximates the function f; so if f is unbounded, you cannot expect the upper bound on $|f * g_i|$ to go to 0 uniformly.

We have bounds for the convolution now, but we can also have similar bounds for the Fourier transform for the convolution — we have

$$\left\|\widehat{f * g_t}\right\|_1 = \left\|\widehat{f}\cdot\widehat{g}_t\right\|_1 \le (2\pi)^{d/2}t^{-d/2}\left\|\widehat{f}_1\right\|.$$

In an analogous way, we also have the same bound on the L^∞ norm.

Now we can actually prove that the Fourier iinversion formula holds for $f * g_t$. That's the second step we mentioned earlier.

Lemma 11.13

The Fourier inversion formula holds for Gaussian convolutions $f * g_t$.

(Again,t he strategy was to first show that the Fourier inversion formula holds for g_t, the density of a d-dimensional Gaussian with variance t. The second step is what we're doing now, to prove it for this convolution. And the third step is to take $t \to 0$, and say that this approximates f.)

Proof. This will be done by a direct computation. We have

$$(f * g_t)(x) = \int_{\mathbb{R}^d} f(x-y) g_t(y)\, dy.$$

But now we know the Fourier inversion formula holds for g_t, so we can replace g_t with its Fourier transform; then we get

$$\int_{\mathbb{R}^d} f(x-y) \left(\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{g}_t(u) e^{-i\langle u, y\rangle}\, du \right) dy.$$

Now we do the usual thing where we use Fubini's theorem to swap the order of integration; then we end up with

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x-y) \widehat{g}_t(u) e^{-i\langle u, y\rangle}\, dy\, du.$$

And on the inside integral we can separate the random variables — we multiply and divide by $e^{-i\langle u, x\rangle}$, and then this is just equal to

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(x-y) e^{-i\langle u, x-y\rangle}\, dy \right) \cdot \widehat{g}_t(u) e i\langle u, x\rangle\, du.$$

And now by the trnaslational invariance of the Lebesgue eausre, the first term is just the Fourier transform of f at u, so we get that this is

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(u) \widehat{g}_t(u) e^{i\langle u, x\rangle}\, du.$$

But the first two terms are just the Fourier transform of the convolution, so this is

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{(f * g_t)}(u) e^{-i\langle u, x\rangle}, du.$$

So we've proved that the Fourier inversion formula indeed holds for Gaussian convolutions. □

(We're using the bounds on magnitudes from earlier to justify Fubini's theorem — if we put absolute values on everything then we'd have a finite integral.)

§11.4.1 Approximating functions by Gaussian convolutions

To complete the proof, we're going to see that in a certain nice way, f is approximated by g_t as $t \to 0$. We're not going to prove this because the proof is quite techncial and out of scope, but we'll state the result and see how it's used to deduce the Fourier inversion formula. (The proof is probably in the notes on Canvas.)

Lemma 11.14

Suppose that $f \in L^p$ for $1 \leq p < \infty$. Then

$$\|(f * g_t) - f\|_p \to 0$$

as $t \to 0$.

So Gaussian convolutions are an extremely useful tool because they help us approximate functions in a certain nice way. Now we're going to somehow extract the Fourier inversion formula in general form, by just taking limits.

Proof of Fourier inversion formula. Set $f_t = f * g_t$. Then we first have that

$$\widehat{f_t}(u) = \widehat{f}(u)\widehat{g_t}(u) = \widehat{f}(u)e^{-|u|^2 t/2}.$$

We also know that $\|f_t - f\|_1 \to 0$ as $t \to 0$ (by the above statement).

This holds without making any assuptions on $\widehat{f}$. But at this point, we use the assumption that $\widehat{f} \in L^1$ — we'll see that this is useful to apply the dominated convergence theorem and change the order of integration and summation.

So now we apply the Fourier inversion formula on f_t to write

$$f_t(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \widehat{f}(u)e^{-|u|^2 t/2}e^{-i\langle u,x\rangle}\, du.$$

(We saw that the Fourier inversion formula is true for f_t, and we know the explicit form of the Fourier transform of f_t.)

We want to get exactly the same thing, but with f_t replaced by f. We need to justify this by taking limits as $t \to 0$; and this is where we use the fact that $\widehat{f} \in L^1$ (to see that this integral converges to the same integral with $t = 0$). For this, we know this is dominated by the L^1 function

$$\left|\widehat{f}(u)e^{-|u|^2 t/2}e^{i\langle u,x\rangle}\right| \leq \left|\widehat{f}\right|.$$

So we can apply the dominated convergence theorem, which tells us that

$$f_t(x) \to (2\pi)^{-d} \int \widehat{f}(u)e^{-i\langle u,x\rangle}\, du.$$

So $f_t(x)$ converges to the integral on the right-hand side as $t \to 0$.

Now, we know that f_t converges to f in L^1, but we *don't* know it converges almost everywhere. But we do know that it converges almost everywhere along a *subsequence*. In words, because $\|f_t - f\|_1 \to 0$, we can find a subsequence (t_n) such that $t_n \to 0$ as $n \to \infty$ and $f_{t_n} \to f$ as $n \to \infty$ Lebesgue-almost everywhere. But we also know that f_{t_n} converges Lebesgue-almost everywhere to the above function. So f has to be equal to that function Lebesgue-almost everywhere — i.e.,

$$f(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \widehat{f}(u)e^{-i\langle u,x\rangle}\, dx$$

for Lebesgue-almost every $x \in \mathbb{R}^d$. And that's exactly what we wanted to prove. □

So we've proven the Fourier inversion formula; this is very useful because if you know the Fourier transform, then you can recover the function.

Student Question. *We know this holds for L^1 functions, but is there a version of this for measures — a notion of recovering a measure from its Fourier transform?*

Answer. Yes. This is more or less equivalent to saying that the characteristic function determines the law.

Student Question. *When this equality fails, does it fail because of divergence?*

Answer. It could, but really the function could be whatever you imagine on a set of measure 0, and you'd still get the same Fourier transform.

Next lecture, as we can imagine, we're going to focus on the Fourier transform in L^2. There, there are some nice things going on — we'll see that the Fourier transform for L^2 functions behaves very nicely.

§12 October 17, 2024

In the previous lecture, we completed the Fourier inversion formula. This is a very important theorem because it gives a way of recovering the original function — we start with a function that can be very nasty (but is in L^1), and as long as its Fourier transform is in L^1 as well, we can recover the function. And the Fourier transform may satisfy much better properties (e.g., continuity or even differentiability).

We're not going to go very deep into Fourier analysis, but we will use the Fourier inversion formula to deduce properties of random variables.

§12.1 Fourier transforms in L^2

Before we do this, we'll talk a bit about L^2 — L^2 spaces are special because there's a notion of inner products. And there we'll see that the Fourier transform extends to a Hilbert space automorphism.

Theorem 12.1

There exists a Hilbert space automorphism $F: L^2 \to L^2$ such that $F(f) = (2\pi)^{-d/2}\hat{f}$ for every $f \in L^1 \cap L^2$.

(Here we're working in $\mathbb{R}^d$.)

So if you take the Fourier transform in L^2 and evaluate it at functions f which are in both L^1 and L^2 (the Fourier transform only makes sense for functions in L^1), then this extends into a Hilbert space isomorphism.

Student Question. *Isn't L^2 contained in L^1?*

Answer. No, because here we're not in a finite measure space. If the space is finite, you can apply Hölder's inequality to show $L^2 \subseteq L^1$, but in general (in infinite-measure spaces) this is not the case. (Here we're working with $\mathbb{R}^d$ with the Lebesgue measure.)

Proof. (This uses some basic knowledge from topology and functional analysis, but we will just state the results we need.)

We define a function $f_0: L^1 \cap L^2 \to L^2$ by $f \mapsto (2\pi)^{-d/2}\hat{f}$; this is a well-defined function.

The *Plancherel identity* states that f_0 preserves L^2-norm — we have $\|f_0(f)\|_2 = \|f\|_2$ for all $f \in L^1 \cap L^2$. (We'll take this for granted — the L^2 norm is somehow preserved under Fourier transforms, modulo some constant.) This means f_0 is an isometry, in the sense that it preserves norm.

We also know that $L^1 \cap L^2$ is dense in L^2. This can be taken as a black box, but it can also be proven using what we know — any function in L^2 can be approximated by simple functions, and simple functions are in both L^1 and L^2.

So by a result from functional analysis, this means f_0 extends uniquely to an isometry $F: L^2 \to L^2$, which is the isometry we want. In general, if you're in a Hilbert space and you have an isometry defined on a dense space, it can be extended to an isometry of the entire space. More generally, this also works with complete metric spaces.

So this is our function F; to complete the proof, we need to show that F is onto, meaning that for every $g \in L^2$, we can find $f \in L^2$ such that $F(f) = g$.

For this, we apply Fourier inversion — by the Fourier inversion formula, we have that $F(V) = V$ where

$$V = \{f \in L^2 \mid f, \hat{f} \in L^1\}.$$

So this space V is invariant under F. More specifically, we have that if $f \in V$, then $F^4(f) = f$. (Note that applying it twice gives the function $F^2(f)(x) = f(-x)$ by the definition of the Fourier transform and the Fourier inversion formula (where you swap signs in the exponential), so doing it four times gets f.)

This gives that $V \subseteq F(L^2)$. And since V is dense in L^2 (this can be shown by showing that V contains all Gaussian convolutions, and we have shown that Gaussian convolutions are dense in L^2), we deduce that $\overline{V} = L^2$ is contained in $\overline{F(L^2)}$ as well. And since F is an isometry, this is just $F(L^2)$. So we get that $F(L^2) = L^2$, which completes the proof (our transformation is also a bijection). □

Here we have crucially used the structure of L^2 to say we have an isometry (the Plancherel identity is only true in L^2, not for other values of p). So we have a very nice identity that lets us extend the Fourier transform to the entire space, even though it's not originally defined on the whole space.

§12.2 Characteristic functions

We'll now see a first application in probability theory. The main applications lie in characteristic functions of random variables.

Theorem 12.2

The characteristic function ϕ_X of a distribution μ_X of a random variable X determines μ_X — i.e., if X and X' are random variables and $\phi_X = \phi_{X'}$, then $\mu_X = \mu_{X'}$.

So this says that if you know the characteristic function, then you know the distribution. This will follow from Fourier analysis. (We don't need to know the proof, only the result, but we'll briefly sketch the proof.)

Proof sketch. First, if we want to determine μ_X, we want to determine $\mu_X(A) = \mathbb{P}[X \in A]$ for all $A \in \mathcal{B}(\mathbb{R})$ (we can also work in $\mathbb{R}^d$, it doesn't matter). One way to write this is as $\mathbb{E}_{\mu_X}[\mathbf{1}_A]$. It suffices to do this in the case where A is an interval, since if you have two Borel measures on $\mathbb{R}$ that agree on intervals, they have to agree everywhere (it's a π-system generating the Borel σ-algebra).

So everything boils down to how to approximate the indicator variable of an interval. And you can do that using continuous functions. Suppose we have an interval $[a, b]$; the indicator function looks like a straight line here, and 0 everywhere else.

To approximate this function, we can just add a very small continuous function close to a and b — we can take the points $(a - 1/n, 0)$ and $(b + 1/n, 0)$, and draw a straight line from $(a - 1/n, 0)$ to $(a, 1)$ and so on. Now this picture is the graph of a continuous function — it's 0 outside the interval $[a - 1/n, b + 1/n]$, it's linear on the first segment, constant on the middle segment, and linear on the last. So this is some

continuous function g_n; and we have $g_n \to \mathbf{1}_{(a,b)}$ almost everywhere as $n \to \infty$. And everything is bounded uniformly, so by the dominated convergence theorem we obtain

$$\mu_X((a,b)) = \lim_{n\to\infty} \mathbb{E}_{\mu_X}[g_n].$$

This means if we know $\mathbb{E}_{\mu_X}[g_n]$ (the integral with respect to μ_X, which is the same as $\mathbb{E}[g_n(X)]$), then we know this measure μ_X.

And how can we obtain this value? We can just apply Fourier inversion. If we do this, we get that ϕ_X determines $\mathbb{E}[g_n(X)]$ — if two characteristic functions are equal, then these quantities $\mathbb{E}[g_n(X)]$ will be equal (for the two random variables). Why? One way to write $\mathbb{E}[g_n(X)]$ is as

$$\mathbb{E}[g_n(X)] = \int_{\mathbb{R}} g_n(y)\, d\mu_X(y).$$

If you apply Fourier inversion, you can recover this quantity by the characteristic function. The point is that the Fourier inversion formula also works for measures as well. And that's the idea of the proof. You apply Fourier inversion to determine the values $\mathbb{E}[g_n(X)]$, and this also determines $\mu_X((a,b))$. And since the set of intervals is a π-system, if two measures agree on a π-system they have to agree everywhere.

We've hidden some details, specifically how exactly the Fourier transform of the measure determines the measure, which is not obvious. But it's exactly the same thing we've done so far — in the same way that the Fourier transform encodes the function, the Fourier transform of a measure encodes the measure. So saying the characteristic functions are the same means that the Fourier transforms of the measures are the same. □

Another application of Fourier analysis is the following.

Theorem 12.3

If ϕ_X is integrable (where X is a random variable in $\mathbb{R}^d$), then μ_X has a bounded, continuous density function, given by

$$f_X(y) = (2\pi)^{-d} \int_{\mathbb{R}^d} \phi_X(u) e^{-i\langle u,y\rangle}\, du.$$

In general, we mentioned in some previous lecture that it's not true in general that distributions of random variables have a density function; in general it's not true. But if you assume the characteristic function is integrable, then there is a density function, it's continuous, and it has this form.

Proof sketch. We start with a random variable X, and the idea is to approximate it. To do so, we consider the random variable $X + \sqrt{t}Z$, where $Z \sim \mathcal{N}(0,1)$. As $t \to 0$, this will converge to X. And we commented in the previous lecture that adding Z is the same as modifying the density of X, if it exists (but we don't yet know it exists). So what we're going to do is the following.

First, we know $X + \sqrt{t}Z$ has a bounded, continuous density function. Then if we apply Fourier inversion to the density function of $X + \sqrt{t}Z$, it's given by

$$f_t(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \phi_X(u) e^{-|u|^2 t/2} e^{-i\langle u,x\rangle}\, du.$$

Here we've just applied the Fourier inversion formula. And now we can take limits. We know we can bound the function on the inside by $|\phi_X|$, and we know this is integrable (because we assumed ϕ_X is integrable). So that means this quantity has a limit. The integral pointwise converges to $\phi_X(u)e^{-i\langle u,x\rangle}$, so this implies $f_t(x) \to f(x)$ pointwise, where

$$f(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \phi_X(u) e^{-i\langle u,x\rangle}\, du.$$

So what we've done is we don't know the density of X exists, but we know the density of $X + \sqrt{t}Z$ exists and is very nice. And we can apply Fourier inversion and get this expression. And then we prove that this has a limit, and deduce that the limit exists. And the limit has to be the density function of X. □

So we have a criterion that ensures a random variable actually has a density function, which in general is not true (if ϕ_X is not integrable, there are examples where you cannot find the density function).

Student Question. *How do we know $X + \sqrt{t}Z$ has this density?*

Answer. They're independent. And the law of the sum of independent random variables is the convolution of the measures; so we convolve the law of $\sqrt{t}Z$ (which is very explicit) with the law of X. The convolution of a function (here $\sqrt{t}Z$) with a measure (here μ_X) has a density, and you can compute it explicitly.

(We've also skipped some of the details, which we can study on our own; we won't be asked for the proofs, but it's nice to have a rough idea.)

§12.3 Weak convergence

Weak convergence is another notion of convergence that also arises naturally.

Definition 12.4 (Weak convergence of measures)**.** Let μ and (μ_n) be Borel probability measures. We say that $\mu_n \to \mu$ weakly as $n \to \infty$ if $\mu_n(g) \to \mu(g)$ as $n \to \infty$ for all bounded continuous functions g.

So we have weak convergence if and only if the integrals of bounded continuous functions converge. Immediately, you get the analogous definition for weak convergence of random variables (the random variables converge weakly if and only if their distributions converge weakly).

Definition 12.5 (Weak convergence of random variables)**.** Let X and (X_n) be random variables taking values in $\mathbb{R}^d$. We say that $X_n \to X$ weakly as $n \to \infty$ if and only if $\mu_{X_n} \to \mu_X$ weakly.

So this is the notion of weak convergence for random variables — they converge weakly if and only if their distributions do. So far we've seen almost sure convergence, L^p convergence, and convergence in distribution; and now we have weak convergence.

It's tempting to ask for the relationship between convergence in distribution and weak convergence. In one dimension they're actually equivalent (this is not obvious, and may be on a problem set). But the advantage of weak convergence is that it makes sense in higher dimensions as well; that's why sometimes we prefer to work with weak convergence of random variables, because often the random variables we're dealing with are multidimensional (e.g., we might have a multidimensional Gaussian vector).

In general, it's quite useful theoretically. One reason is it's related to convergence of characteristic functions — somehow weak convergence of random variables is encoded into the behaviour of the characteristic functions. We'll state this and briefly sketch the proof (we don't have to know the proof, but it'll be useful if we want to get more familiar with the techniques).

Theorem 12.6

Let X and (X_n) be random variables in $\mathbb{R}^d$. If $\phi_{X_n}(u) \to \phi_X(u)$ for all $u \in \mathbb{R}^d$ (i.e., the characteristic functions converge pointwise), then $\mu_{X_n} \to \mu_X$ weakly.

So if we have pointwise convergence of the characteristic functions, then we have weak convergence. This is important, because it'll be important later when we prove the central limit theorem — we're going to use this characterization.

Proof sketch. We'll apply the same method as before: it suffices to prove that $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ as $n \to \infty$ in the case where g is a smooth and compactly supported function in $\mathcal{C}^\infty$. (Here we want our function to be smooth, which means it can be differentiated infinitely many times; and we want it to be 0 outside a large cube.) This is not exactly obvious, but we can see a picture in one dimension: if you again take an interval and your indicator function $1_{(a,b)}$, and you again write down $a - 1/n$ and $b + 1/n$, before we just considered linear segments. But we can always make them smooth (by drawing a smooth curve instead of a straight line). (This is an approximation in L^1.)

To do this, we again use the same random variable $X + t\sqrt{Z}$ from before (where $t > 0$) to approximate X, where $Z \sim \mathcal{N}(0,1)$ is independent from X. Then if we apply Fourier inversion for this new random variable and the convergence of the characteristic functions ϕ_{X_n} to ϕ_X, we get $\mathbb{E}[g(X_n + \sqrt{t}Z)] \to \mathbb{E}[g(X + \sqrt{t}Z)]$. So you have this convergence for all t.

Finally, you check that

$$\left|\mathbb{E}[g(X_n + \sqrt{t}Z)] - \mathbb{E}[g(X_n)]\right|$$

is small (uniformly in n), using the fact that g is smooth — specifically, we get a bound that only depends on t and that tends to 0 as $t \to 0$.

So the idea is to first prove the convergence with $t\sqrt{Z}$ added, adn then show that these things are very close to the one without t (uniformly in n, as $t \to 0$). □

(This property will be useful when we'll prove the central limit theorem, about sums of independent random variables.)

> **Remark 12.7.** The reverse is also true (and is fairly obvious), because the characteristic function is the integral of a bounded function.

§12.4 Gaussian random variables

Gaussian random variables are very important in probability, so we'll now look at them more carefully. They appear in central limit theorems and generally whenever you have sums of independent random variables — that's why they're quite important and they appear in nature everywhere.

> **Definition 12.8** (Gaussian random variable)**.** A random variable X (in one dimension) is Gaussian if there exists some $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ such that the density of X is given by
>
> $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$
>
> We write this as $X \sim \mathcal{N}(\mu, \sigma^2)$.

> **Proposition 12.9**
>
> If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2$. Moreover, for all $a, b \in \mathbb{R}$, we have that $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Finally, we have
>
> $$\phi_X(x) = e^{iu\mu - u^2\sigma^2/2}.$$

So we have some kind of linearity, and we have explicit formulas for its mean, variance, and characteristic function.

The first two statements are fairly standard manipulations with integrals (you can find them on Wikipedia or in the lecture notes); we're only going to prove the last statement, which is less obvious.

Proof. The main observation is that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then by the linearity property, we can write $X = \sigma Z + \mu$ where $Z \sim \mathcal{N}(0,1)$. And we know the characteristic function for Z — we computed earlier that

$$\phi_Z(u) = e^{-u^2/2}.$$

This implies that

$$\phi_X(u) = \mathbb{E}[e^{iu(\sigma z+\mu)}] = e^{iu\mu}\mathbb{E}[e^{i(\sigma u)Z}] = e^{iu\mu} \cdot \phi_Z(u\sigma) = e^{iu\mu - u^2\sigma^2/2},$$

which is exactly what we need. □

(This is straightforward because we already did the hard work of computing ϕ_Z in some previous lecture.)

§12.5 Multidimensional Gaussians

We're going to state and prove similar properties for Gaussians in higher dimensions, because we'll need those. Before we do that, we should define them.

Definition 12.10 (Gaussian random variable)**.** We say that X is a Gaussian on $\mathbb{R}^n$ if $\langle u, X\rangle$ is a Gaussian on $\mathbb{R}$ for all $u \in \mathbb{R}^n$.

Now we're going to actually compute the expectation and variance and characteristic function for an arbitrary Gaussian in many dimensions.

Theorem 12.11

Let X be a Gaussian in $\mathbb{R}^n$, and let A be an $m \times n$ matrix and $b \in \mathbb{R}^n$. Then:

(i) $AX + b$ is a Gaussian on $\mathbb{R}^n$.

(ii) $X \in L^2$, and its law μ_X is determined by its mean $\mu = \mathbb{E}[X]$ and its covariance matrix

$$V = \operatorname{Var}[X] = (\operatorname{Cov}(X_i, X_j))_{i,j\in[n]}.$$

(iii) We have

$$\phi_X(u) = e^{i\langle u,\mu\rangle - \langle u, Vu\rangle/2}.$$

(iv) If V is invertible, then X has a density, given by

$$f_X(x) = (2\pi)^{-n/2} \det(V)^{-1/2} \cdot \exp\left(-\frac{1}{2}\langle x-\mu, V^{-1}(x-\mu)\rangle\right).$$

(v) If $X = (X_1, X_2)$, then X_1 and X_2 are independent if and only if $\operatorname{Cov}(X_1, X_2) = 0$.

For (i), linearity in one dimensions means we multiply by a number and add a number; in many dimensions we multiply by a *matrix* (and add a vector). For (ii), in one dimension the covariance matrix consisted of a single number, namely, the variance; here you replace that by the covariances of pairs of entries.

And (v) is a really important property of Gaussians — they're independent if and only if their covariance is 0.

Here we have a bunch of properties; we're going to prove them, and we're going to use most of them in this course.

Proof of (i). If we fix $u \in \mathbb{R}^m$, then we have

$$\langle AX + b, u\rangle = \langle AX, u\rangle + \langle b, u\rangle = \langle X, A^\intercal u\rangle + \langle b, u\rangle.$$

And this is Gaussian — $A^\intercal u$ is just a vector, so we know by assumption that the first term is a Gaussian, and the second is a constant; if we add a constant to a Gaussian, then we still get a Gaussian. So $\langle AX+b, u\rangle$ is indeed a Gaussian. □

Proof of (ii). If we write $X = (X_1, \dots, X_n)$, then we have $|X|^2 = |X_1|^2 + \cdots + |X_n|^2$ by definition of the norm. This means

$$\mathbb{E}[|X|^2] = \sum \mathbb{E}[|X_j|^2].$$

But this is always finite, because we know one-dimensional Gaussians are in L^2 (and the finite sum of finite quantities is also a finite quantity.) □

Proof of (iii). Let $\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])$ and $V = \mathrm{Var}(X) = (\mathrm{Cov}(X_i, X_j))$. Then for $u \in \mathbb{R}^n$, we have

$$\mathbb{E}[\langle u, X\rangle] = \langle u, \mu\rangle,$$
$$\mathrm{Var}(\langle u, X\rangle) = \langle u, Vu\rangle.$$

(This is not hard to check, and follows by definition of the covariance matrix.) By the one-dimensional case, this implies $\langle u, x\rangle \sim \mathcal{N}(\langle u, \mu\rangle, \langle u, Vu\rangle)$. And this means we know $\chi_X(u) = \mathbb{E}[e^{i\langle u, x\rangle}]$, because $\langle u, X\rangle$ is a one-dimensional Gaussian and we know the characteristic function of such a thing; so we get

$$\chi_X(u) = e^{i\langle u,\mu\rangle - \langle u, Vu\rangle/2}.$$

And since this is true for every u, we get that μ and V determine the characteristic function of X, so they determine the law of X as well. □

§13 October 22, 2024

§13.1 Gaussian random variables

Last class, we stated a theorem about several important properties of the Gaussian random variable. Importantly, it's closed under linear transformations — if X is Gaussian, then $AX + b$ is also a Gaussian random variable, though with a different variance and mean. We also saw that if you have a Gaussian random variable and you know its mean and variance, then you can recover its law.

Theorem 13.1

Let $X = (X_1, \dots, X_n)$ be a Gaussian random variable on $\mathbb{R}^n$. Let $\mu = \mathbb{E}[X]$ be the mean of X and V be the covariance matrix of X, defined as $V_{ij} = \mathrm{Cov}[X_i, X_j]$.

(iv) If V is invertible, then X has a density, given by

$$f_X(y) = (2\pi)^{-n/2} \det(V)^{-1/2} \exp\left(-\frac{1}{2}\langle x - \mu, V^{-1}(x - \mu)\rangle\right).$$

(v) If $X = (X_1, X_2)$ (i.e., $n = 2$), then $\mathrm{Cov}[X_1, X_2] = 0$ if and only if X_1 and X_2 are independent.

If X is Gaussian, then all its components X_j are also Gaussian, but they're not necessarily independent — there could be correlations between them. In general, it's not easy to find the density of X — it has a complicated form. But when V is invertible, it does have a nice form.

The last property is really important — if you want to check whether two Gaussians are independent, it's enough to just check their covariance. One direction is obvious — if they're independent, then the covariance has to be 0, because $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$. But the point is that the converse is true as well.

Remark 13.2. This is true only for Gaussians — it's not true for other random variables.

Proof of (iv). The idea is to start with a boring Gaussian vector — one whose components are independent Gaussians — and somehow construct X as an appropriate linear combination. In particular, let $Y = (Y_1, \dots, Y_n)$ where $Y_j \sim \mathcal{N}(0,1)$ are all independent. Then we know the density of Y — we've seen how to do this in previous lectures — and it's given by

$$f_Y(x) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\|x\|^2\right)$$

for every $x \in \mathbb{R}^n$. (This is something we've shown.) Now we need to somehow construct X given Y, as an appropriate linear combination. For this, we define the random variable

$$\hat{X} = V^{1/2}Y + \mu$$

(note that $V^{1/2}$ makes sense because V is nonnegative definite); this is some random variable in $\mathbb{R}^n$. We've seen that linear combinations of Gaussians are Gaussian, so $\hat{X}$ is a Gaussian. And we also gave an explicit description of the mean and variance of linear combinations of a Gaussian vector; we have $\mathbb{E}[Y] = 0$, so $\mathbb{E}[\hat{X}] = \mu$, while $\mathrm{Var}(\hat{X}) = V$. So this means $\hat{X}$ has the same mean and same covariance matrix as X; and this means $\hat{X}$ has the same law as X.

Student Question. *Why can we take $V^{1/2}$?*

Answer. It's a nonnegative definite matrix. You don't need invertibility here; we'll see where invertibility comes in later.

So this means $\widehat{X}$ has the same distribution as X. In particular, this means $f_X = f_{\widehat{X}}$.

Now we have to compute integrals — we have a transformation $Y \mapsto V^{1/2}Y + \mu$, so we have to change coordinates using this linear transformation and apply the Jacobian and stuff. If you do this, you end up with the formula we have (it's the Jacobian which is where you get the determinant). Eventually, what you do is get that

$$\mathbb{P}[\widehat{X} \in A] = \mathbb{P}[Y \in V^{-1/2}(A - \mu)],$$

and this probability is explicit — it's

$$(2\pi)^{-n/2} \int_{V^{-1/2}(A-\mu)} f_Y(x)\, dx.$$

And now you change variables here, because you have a linear transformation, and apply the Jacobian. If you change coordinates here, then you obtain that this probability is the integral of the original function over A.

Student Question. *Could you also use the pushforward measure?*

Answer. Yes; that's exactly what's going on (when you change variables) But there's also a formula in n dimensions for how to transform integrals using the Jacobian and something like that.

(We will omit the computations, but the point is that the main idea of the proof is to start with a normal Gaussian vector and take linear combinations to obtain a Gaussian with the same law as X.) □

Proof of (v). One direction is immediate — if X_1 and X_2 are independent, then $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$, so $\operatorname{Cov}[X_1, X_2] = 0$.

What's less obvious is the other direction — that if $\operatorname{Cov}[X_1, X_2] = 0$ then they're independent. The main point is that covariance 0 means the non-diagonal entries of the covariance matrix are 0, so

$$V = \begin{bmatrix} v_{11} & 0 \\ 0 & v_{22}. \end{bmatrix}$$

This implies if we take $u = (u_1, u_2) \in \mathbb{R}^2$, then we have

$$\langle u, Vu\rangle = u_1 v_{11} u_1 + u_2 v_{22} u_2.$$

And here $v_{11} = \operatorname{Var}[X_1]$ (by the definition of the covariance matrix, since $\operatorname{Cov}[X_1, X_2] = \operatorname{Var}[X_1]$), and $v_{22} = \operatorname{Var}[X_2]$. So we can also write this inner product as

$$\langle u, Vu\rangle = u_1^2 \operatorname{Var}[X_1] + u_2^2 \operatorname{Var}[X_2].$$

And this shows the characteristic functions factorize; if you manage to prove that $\phi_X = \phi_{X_1}\phi_{X_2}$, then you have independence. (There are many different ways to characterize independence, and one is to prove the characteristic function factorizes; this is what we're going to use.)

We've shown in a previous lecture that

$$\phi_X(u) = e^{i\langle u,\mu\rangle - \frac{1}{2}\langle u, Vu\rangle}.$$

And then we can just expand using this, and we get

$$e^{iu_1\mu_1 - \frac{1}{2}u_1^2 \operatorname{Var}[X_1]} e^{iu_2\mu_2 - \frac{1}{2}u_2^2 \operatorname{Var}[X_2]}.$$

But the two terms here are just $\phi_{X_1}(u_1)$ and $\phi_{X_2}(u_2)$; so we get

$$\phi_X(u) = \phi_{X_1}(u_1)\phi_{X_2}(u_2).$$

So we've shown the characteristic function factorizes, and this implies X_1 and X_2 are independent. □

We're going to use these properties later when proving the central limit theorem, which states the long-range behavior of sums of independent random variables roughly behaves like a Gaussian.

§13.2 Ergodic theory

This concludes the chapter on Fourier analysis and Gaussians. Now we're moving towards the proofs of the main limit theorems. But we're not exactly done, because we need a few more other tools. These tools come from ergodic theory. We shouldn't be terrified by this phrase, because we won't do much with it; we'll only focus on the tools we need. Ergodic theory is an entire branch of math with lots of interaction with other things (number theory, analysis, combinatorics, probability), but we'll focus on things.

Intuitively, ergodic theory is the study of long-run behavior of a system under the application of a map (so you have a map and apply it multiple times and see what happens). For example, you might look at the evolution of the position or velocity of a particle.

We're going to state and prove some ergodic theorems which will be very useful when proving the strong law of large numbers.

§13.3 The setup

In this setup, we have a measure space $(E, \mathcal{E}, \mu)$ (where E is a set, $\mathcal{E}$ a σ-algebra on E, and μ a (not necessarily probability) measure). And we also have a measurable map $\theta: E \to E$ which is *measure-preserving*.

Definition 13.3. A map θ is measure-preserving if $\mu(A) = \mu(\theta^{-1}(A))$ for all $A \in \mathcal{E}$.

We're going to study particular systems of this form, and we're going to see how and why they're useful.

Example 13.4

Take $(E, \mathcal{E}, \mu) = ([0,1), \mathcal{B}([0,1)), \text{Lebesgue})$ (in this case, the measure is a probability measure). Then for every $a \in [0,1)$, we can define the map

$$\theta_a: x \mapsto x + a \pmod 1.$$

(What we mean by mod 1 is $x + a - \lfloor x + a \rfloor$.) By the translation invariance of the Lebesgue measure, θ_a is indeed measure-preserving.

In general, the goal of ergodic theory is to understand long-run averages of the system when we apply θ many times. In particular, we'll be interested in the following case. Suppose we have a function $f: E \to \mathbb{R}$ which is measurable, and let

$$S_n(f) = f + (f \circ \theta) + (f \circ \theta^2) + \cdots + (f \circ \theta^{n-1}).$$

The main goal of ergodic theory is to know the following thing:

Question 13.5. How does the *ergodic average* $\frac{1}{n} S_n(f)$ behave as $n \to \infty$?

In this course, we'll be interested in a particular choice of f and θ — you can see that this looks like averages of random variables, and that's what we're interested in, so that's how ergodic theory comes into play.

We're going to see that under certain conditions, this quantity converges in some sense.

§13.4 Invariance

Before trying to analyze the objects we're interested in, we'll talk about some notions of invariance in such a space.

Definition 13.6 (Invariant function)**.** A measurable function $f: E \to \mathbb{R}$ is invariant if $f = f \circ \theta$.

(This is an analogous notion of the notion that $\mu(A) = \mu(\theta^{-1}(A))$.)

Definition 13.7 (Invariant subsets)**.** We say that a set $A \in \mathcal{E}$ is invariant for θ if $A = \theta^{-1}(A)$.

We'll see that these concepts are quite important in ergodic theory, because we'll analyze ergodic averages of this kind of functions.

One might wonder what kinds of structure the space of invariant subsets has. It turns out that it's a σ-algebra.

Definition 13.8. We write $\mathcal{E}_\theta = \{A \in \mathcal{E} \mid A \text{ is invariant}\}$.

Fact 13.9 — The collection $\mathcal{E}_\theta$ is a σ-algebra, and a function $f: E \to \mathbb{R}$ is invariant if and only if it is $\mathcal{E}_\theta$-measurable.

(This is fairly easy to see, so we can try to do it as an exercise.)

The most important definition is the following.

Definition 13.10 (Ergodicity). We say θ is ergodic if for every $A \in \mathcal{E}_\theta$, we have $\mu(A) = 0$ or $\mu(E \setminus A) = 0$.

In probability spaces, this is equivalent to saying that $\mu(A) \in \{0, 1\}$.

Example 13.11

Let $(E, \mathcal{E}, \mu)$ be $[0, 1)$ with the Lebesgue measure, as before. Then θ_a is ergodic if and only if $a \in \mathbb{R} \setminus \mathbb{Q}$.

This is a nice exercise; it's not super difficult, but it's not obvious.

§13.5 Integrals

Now let's see how integrals behave under invariant functions.

Proposition 13.12

If f is integrable and θ is measure-preserving, then $f \circ \theta$ is also integrable, and $\int_E f \, d\mu = \int_E (f \circ \theta) \, d\mu$.

So integrals are preserved under measure-preserving mappings; that makes intuitive sense. The way you prove this is you want to prove a property for a general class of functions, so you first prove it for simple functions, then nonnegative measurable functions (by approximating them with simple functions, using e.g., the monotone convergence theorem to change between integration and taking limits); and finally you prove it for arbitrary integrable functions by writing them as the difference of two nonnegative ones. We won't go through the details because we've seen this argument several times, and the method is quite standard.

Question 13.13. If we start with an ergodic transformation θ, how many invariant functions are there?

The answer is that there are very, very few such functions. In particular, we have the following (which will probably be an exercise on the problem set — it's nice to try to do it on your own):

Proposition 13.14

If a measure-preserving map θ is ergodic and f is invariant under θ, then there exists a constant $c \in \mathbb{R}$ such that $f = c$ μ-almost everywhere.

So the only mappings which are invariant under ergodic transformations are constant. This is important — if you want to show the ergodic average $\frac{1}{n} S_n(f)$ is constant, then you show that its limit exists and is invariant for some ergodic map, and then you deduce that it has to be constant.

These are the general concepts from ergodic theory, and for the rest of the lecture we're going to spend some time studying a particular example of a measure space and ergodic mapping.

§13.6 The Bernoulli shift

Now we'll spend some time studying the Bernoulli shift, because we're going to use it to prove the strong law of large numbers — we're going to apply all the theorems we'll hopefully state today to this particular example of an Ergodic system.

Let m be a Borel probability measure on $\mathbb{R}$. Then we know there exists an independent and identically distributed sequence of random variables Y_1, Y_2, ... with law m. We proved this an earlier lecture (we constructed it in a funny way). Now we're going to work with these more naturally, using Bernoulli shifts.

First, we need a space E. This space will be a bit fancy — it'll be $E = \mathbb{R}^{\mathbb{N}}$. This is a weird-looking symbol, but it's just the set of all real-valued sequences.

Definition 13.15. We define $\mathbb{R}^{\mathbb{N}}$ as the collection of all sequences $(x_n) \subseteq \mathbb{R}$.

This might look a bit weird, but the idea behind the notation is that a sequence is just a function $x: \mathbb{N} \to \mathbb{R}$ (we just take a positive integer n, and then output x_n). So we're considering the set of all functions $\mathbb{N} \to \mathbb{R}$, and we denote it by $\mathbb{R}^{\mathbb{N}}$.

But we need a σ-algebra on this space. The natural one is the σ-algebra generated by *projections*.

Definition 13.16. For each n, let $\pi_n: \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ be the projection map $x \mapsto x_n$, which takes in every sequence $(x_m)_{m \in \mathbb{N}}$ and outputs its nth element.

We've seen we can define σ-algebras as the σ-algebra generated by a collection of functions; the one we'll define on this space is the smallest σ-algebra which makes all these π_n's measurable. (We've seen similar σ-algebras before.)

Definition 13.17. We define $\mathcal{E}$ to be the σ-algebra on $\mathbb{R}^{\mathbb{N}}$ generated by all the maps π_n.

(Recall that this is the intersection of all the σ-algebras with respect to which π_n is measurable for all n.)

So now we have our space and σ-algebra; what's missing is a measure. For this, we're going to use the construction of i.i.d. random variables (Y_n) from before. But before we do this, there's another way to define the σ-algebra, as the σ-algebra generated by a certain π-system: let $\mathcal{A}$ be the product of all cylinders, i.e.,

$$\mathcal{A} = \left\{ \prod_{n \in \mathbb{N}} A_n \mid A_n \in \mathcal{B}(\mathbb{R}) \text{ for all } n \text{ and } A_n = \mathbb{R} \text{ for all sufficiently large } n \right\}.$$

This is a π-system, where elements have the form $A_1 \times A_2 \times \cdots \times A_n \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \cdots$. So in other words, this is the set of all sequences such that the first n elemnets belong to the sets A_j, and after that we don't care (they can be whatever they want).

Remark 13.18. Here we're abusing notation a bit — we're really looking at all sequences x such that $x_j \in A_j$ for all $1 \leq j \leq n$, and after that we can have anything we want.

Then we have $\mathcal{E} = \sigma(\mathcal{A})$ — so this is an equivalent description of this σ-algebra.

Now we have our space, and we have our σ-algebra; so now we need to construct a measure. To construct a measure, we'll consider a random variable taking values in E. Specifically, we let $Y: (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}^{\mathbb{N}}$ be the random variable (i.e., map $\omega \mapsto Y(\omega)$) which maps $n \mapsto Y_n(\omega)$. In other words, we just have

$$Y = (Y_1, Y_2, \ldots).$$

(This makes sense because the Y_js all live in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, so we can define this.)

Now the natural measure is just the pushforward:

Definition 13.19. We define $\mu = \mathbb{P} \circ Y^{-1}$.

In other words, μ is just the law of Y (from the original probability space). Recall that this means $\mu(A) = \mathbb{P}[Y \in A]$ for all $A \in \mathcal{E}$.

This is maybe a bit abstract, but it's not counterintuitive. What's maybe hard to digest is that so far we've been considering random variables in 1 dimension or maybe n dimensions; it's somehow a bit weird to consider random variables taking values in some infinite-dimensional space. But apart from that, we haven't done anything too complicated.

Student Question. *How do we know the pre-image of any measurable set in $\mathbb{R}^{\mathbb{N}}$ is measurable in the original space?*

Answer. You can consider sets $A \in \mathcal{A}$, for which this is true; and then it's a π-system, so you get everything in $\mathcal{E}$ as well.

Now we'll make a few observations about this measure. First, what's the form of μ when restricted to our π-system $\mathcal{A}$?

Fact 13.20 — For $A \in \mathcal{A}$, we have $\mu(A) = m(A_1) \cdots m(A_n)$ (where $A = A_1 \times \cdots \times A_n \times \mathbb{R} \times \cdots$).

This is by the independence of the Y_i. (Note that we're abusing notation here — by this we really mean $A = \{x \mid x_j \in A_j \text{ for } 1 \le j \le n\}$. We may abuse notation further by simply writing $A_1 \times \cdots \times A_n$ and omitting the $\mathbb{R}$'s.)

Now we're ready to define the *shift map*.

Definition 13.21. We define the shift map $\theta : E \to E$ as $(x_1, x_2, x_3, \ldots) \mapsto (x_2, x_3, x_4, \ldots)$.

In other words, θ takes in a sequence and shifts its elements one to the left. If you think of x as a function $\mathbb{N} \to \mathbb{R}$ (written $n \mapsto x(n)$), then $\theta(x)$ is the sequence $\mathbb{N} \to \mathbb{R}$ which maps $n \mapsto x(n+1)$. So we've just shifted the sequence one position to the left.

So now we have our system — we have our space, our σ-algebra, a measure, and a map. What we're going to do from now on in order to prove some of the nice theorems of the course is to apply ergodic theory to this particular example.

Why is this system important? Suppose we take the function $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ defined by $(x_n)_{n \in \mathbb{N}} \mapsto x_1$ (the projection onto the first coordinate). Then we can write

$$S_n(f)(x) = x_1 + x_2 + x_3 + \cdots + x_n.$$

So this means we can express the sum of x_j's in the form $S_n(f)$. And eventually we're going to study quantities of this form, because we want to study averages in general — we want to study expressions

$$\frac{x_1 + \cdots + x_n}{n}$$

for i.i.d. random variables (because we're interested in averages of random variables), and this can be expressed as the ergodic average $\frac{1}{n} S_n(f)(x)$. So that's the reason we want to study this particular Bernoulli shift — because then we can express averages of random variables as ergodic averages.

We'll now come back to something mentioned earlier, a relationship between ergodicity and Kolgomorov's $\{0, 1\}$ law.

Theorem 13.22

The shift map θ is an ergodic measure-preserving transformation.

First, the fact that θ is measure-preserving is quite easy — you can first show that it preserves measures of sets $A \in \mathcal{A}$ (which is quite easy to see). And then this means it preserves everything.

What's nontrivial is ergodicity, and we'll see that its proof is very similar to Kolgomorov's $\{0,1\}$ law from the beginning of the course.

Proof. To prove θ is ergodic, we'll consider the tail σ-algebra (as we did in the proof of Kolgomorov's $\{0,1\}$ law). For each $n \in \mathbb{N}$, we define

$$\mathcal{T}_n = \sigma(T_m \mid m \geq n+1),$$

and the *tail σ-algebra* is $\mathcal{T} = \bigcap_n \mathcal{T}_n$.

We're going to focus on elements in the π-system $\mathcal{A}$, as before. Suppose that we fix some $A = \prod_{n\in\mathbb{N}} A_n \in \mathcal{A}$. Then if we consider its inverse under θ n times, we get

$$\theta^{-n}(A) = \{x \in \mathbb{R}^{\mathbb{N}} \mid \pi_{n+k}(x) \in A_k \text{ for all } k\}$$

(just by definition, because we've shifted everything by n positions). In particular, this means $\sigma^{-n}(A) \in \mathcal{T}_n$, because it only depends on the values of π_m with $m \geq n+1$.

So $\theta^{-n}(A) \in \mathcal{T}_n$.

This implies that if $A \in \mathcal{E}_\theta$ is invariant under θ, meaning that $A = \theta^{-1}(A)$, then we have $A \in \mathcal{T}_n$ for all n (because if $A = \theta^{-1}(A)$, then we also have $A = \theta^{-n}(A) \in \mathcal{T}_n$), and therefore $A \in \mathcal{T}$. In particular, we have that $\mathcal{E}_\theta \subseteq \mathcal{T}$ (the σ-algebra consisting of invariant sets is contained in the tail σ-algebra).

And what we wanted to show was that every element of $\mathcal{E}_\theta$ has measure 0 or 1. But we know this by Kolgomorov's $\{0,1\}$-law — we know every $A \in \mathcal{T}$ has $\mu(A) \in \{0,1\}$ by Kolgomorov's $\{0,1\}$ law, which implies the same is true for every invariant set $A \in \mathcal{E}_\theta$. But that's exactly the definition of ergodicity. □

So we have a nice ergodic system, and the purpose of proving the limit theorems is to apply the ergodic theorems to this ergodic system.

Student Question. *We're only considering sets in the π-system; how do we know this holds even if your invariant set isn't in the π-system?*

Answer. We want to prove a property is true for every set in the σ-algebra, and to do so it suffices to show that for a π-system (by something like Dynkin's π-system lemma). What we've shown is that any set in the π-system which is invariant under θ has measure 0 or 1; and then you use e.g. Dynkin's lemma to generalize that this property is true for every element of the σ-algebra.

The next step is to state and prove some ergodic theorems, and then we're going to apply them to prove nice laws.

§14 October 24, 2024

Last class, we discussed a bit about ergodic theory and some basic notions from general ergodic theory. We also constructed the *Bernoulli shift* ergodic system and the shift map. We're going to use ergodic theory applied to this specific ergodic system in order to prove some important limit theorems of the course, like the strong law of large numbers. We also proved that this system is ergodic.

In this lecture, we'll first state Birkhoff's ergodic theorem, which we don't have time to prove. But we'll prove von Neumann's ergodic theorem, which comes from it. And we'll see how to apply these to the Bernoulli shift ergodic system in order to prove the first basic theorems of the course.

§14.1 Two ergodic theorems

Theorem 14.1 (Birkhoff's ergodic theorem)

Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space and $\theta: E \to E$ a μ-preserving map, and let $f: E \to \mathbb{R}$ be integrable. Then there exists an invariant function $\overline{f}$ with the property that $\mu(|\overline{f}|) \leq \mu(|f|)$, and

$$\frac{S_n(f)}{n} \to \overline{f} \, \mu\text{-almost everywhere.}$$

In our case, we'll actually have a probability space, so our space will be σ-finite. We'll also always have a map θ, and when we say $\overline{f}$ is invariant, we mean with respect to θ. So if you have a general σ-finite measure space and a measure-preserving function θ, and you take any integrable function f, then its ergodic average will always converge to some invariant function $\overline{f}$.

Here we don't need the assumption that θ is ergodic. But if we do make that assumption, then in particular $\overline{f}$ is constant μ-almost everywhere.

We're going to apply this theorem in the case where θ is ergodic; it'll just be the shift map in the Bernoulli ergodic system.

This is very important because we can express the average of the sum of independent random variables as the ergodic averages in the Bernoulli shift system (for an appropriate function f).

This is a very strong theorem. We don't have enough time to go to its proof, but the main point is that the ergodic average converges to some function, and that function is invariant.

One question one can ask is:

Question 14.2. How can we identify the function $\overline{f}$. For example, if θ is ergodic, then $\overline{f}$ is a constant function, but can we identify it?

If we could apply the dominated convergence theorem, then we could say $\overline{f} = \mu(f)$ — if we also have L^1-convergence, then this has to be true. But it's not always clear if we can apply the dominated convergence; a bit of work is needed to get L^1-convergence. And that's the content of the next ergodic theorem, which we're going to prove.

Student Question. *Without assuming θ is ergodic (where we just show that the averages converge to $\overline{f}$), is the proof constructive?*

Answer. No, the proof is unfortunately more abstract. In general it's not easy to identify this function $\overline{f}$. When you have ergodicity and L^1 convergence, then it'll just be the integral of f; but in general it's not easy to identify. But for our purposes it really will just be the integral of f.

Theorem 14.3 (Von Neumann's ergodic theorem)

Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space, and θ a measure-reserving map. Let $p \in [1, \infty)$, and assume further that $f \in L^p$ and that $\mu(E) < \infty$. Then there is some invariant function $\overline{f} \in L^p$ such that $\frac{1}{n} S_n(f) \to \overline{f}$ in L^p.

So if we make the further assumption that f is in L^p, then we actually have L^p-convergence.

We're going to prove this, and this will use Birkhoff's ergodic theorem (to guarantee the existence of $\overline{f}$).

In particular, if θ is ergodic (so $\overline{f}$ is constant), in a probability space, if we have L^p convergence with $p > 1$, then we also have L^1-convergence, which means $\mathbb{E}[S_n(f)/n] \to \mathbb{E}[\overline{f}]$. But the expectation of the left is $\int f$, and the integral of the right is a constant.

Proof. First note that $\|f\circ\theta\|_p^p = \int_E |f\circ\theta|^p\,d\mu$. And because θ is measure-preserving, this integral has to be equal to $\int_E |f|^p\,d\mu = \|f\|_p^p$. So this means $\|f\circ\theta\|_p^p = \|f\|_p^p$, which implies that

$$\left\|\frac{S_n(f)}{n}\right\|_p = \frac{1}{n}\|f + f\circ\theta + \cdots + f\circ\theta^{n-1}\|_p \le \frac{\|f\|_p + \|f\circ\theta\|_p + \cdots + \|f\circ\theta^{n-1}\|_p}{n} = \|f\|_p$$

(the middle step is Minkowski's inequality). Somehow we need to apply the dominated convergence theorem, but it's not entirely clear how to apply it, because here we don't have a function that dominates all terms $\frac{1}{n}S_n(f)$. We have a sequence that's L^p-bounded, but we don't have a function that dominates the ergodic averages. So we need to somehow truncate this function.

To do so, we fix some $\varepsilon \in (0,1)$ and take $M > 0$ large, such that if we define

$$g = (f \vee (-M)) \wedge M$$

(so that g is f cut off to be between $-M$ and M), then we have $\|f-g\|_p < \frac{\varepsilon}{3}$. We can always do this, because if we take this function and let $M \to \infty$, then this function converges increasingly (in absolute value) to $|f|$; so here we can apply the monotone convergence theorem. So we can always find such a function g with this property.

Now we can apply Birkhoff's theorem — g is also in L^1 (because it's a bounded function in a finite measure space — here we're using the fact that $\mu(E) < \infty$). This means we can apply Birkhoff's theorem to g, and this implies there exists some $\overline{g} \in L^1$ which is invariant, and with the property that

$$\frac{S_n(g)}{n} \to \overline{g} \ \mu\text{-almost everywhere.}$$

The point is that here we *can* apply the dominated convergence theorem, because all the functions on the left are dominated by M — we have

$$\left|\frac{S_n(g)}{n}\right| \le M$$

for all $n \in \mathbb{N}$, so we can apply the dominated convergence theorem to get that

$$\left\|\frac{S_n(g)}{n} - \overline{g}\right\|_p \to 0$$

as $n \to \infty$. This means you can find $N \in \mathbb{N}$ such that

$$\left\|\frac{S_n(g)}{n} - \overline{g}\right\|_p < \frac{\varepsilon}{3} \text{ for all } n \ge N.$$

That was the whole point of introducing g — to apply the dominated convergence theorem (f is not necessarily bounded, so you can't necessarily dominate all the ergodic averages).

Now this means we can write $\|\overline{f} - \overline{g}\|_p$ (where $\overline{f}$ is the function coming from Birkhoff's ergodic theorem) as

$$\|\overline{f} - \overline{g}\|_p^p = \int_E \liminf_{n\to\infty} \left|\frac{S_n(f-g)}{n}\right|^p d\mu.$$

And now we can apply Fatou's lemma to get that this is at most

$$\|\overline{f} - \overline{g}\|_p^p \le \liminf_{n\to\infty} \left(\int_E \left|\frac{S_n(f-g)}{n}\right|^p d\mu\right) \le \liminf_{n\to\infty} \|f-g\|_p^p.$$

And this then implies that

$$\left\|\frac{S_n(f)}{n} - \overline{f}\right\|_p \le \left\|\frac{S_n(f-g)}{n}\right\|_p + \left\|\frac{S_n(g)}{n} - \overline{g}\right\|_p + \|\overline{f} - \overline{g}\|_p$$

by the triangle inequality. But the first term is at most $\varepsilon/3$ (by the choice of g, so that $\|f-g\|_p \le \varepsilon/3$); the second term is again at most $\varepsilon/3$ by the choice of N; and the third is again at most $\varepsilon/3$ by what we did above. So this is at most ε for all $n \ge N$.

And this shows what we wanted — we showed that

$$\left\| \frac{S_n(f)}{n} - \overline{f} \right\|_p < \varepsilon \quad \text{for all } n \ge N,$$

which means that $\frac{1}{n}S_n(f) \to \overline{f}$ in L^p as $n \to \infty$. So we in fact have convergence with respect to the pth norm if we assume that $f \in L^p$. □

And this is the main theorem we're going to use in order to prove the first main limit theorems of the course.

Student Question. *Are we using the fact that $S_n(f-g) = S_n(f) - S_n(g)$?*

Answer. Yes, this is true by linearity. Even if θ is nonlinear, this doesn't matter — we have $(f-g)\circ\theta = (f\circ\theta) - (g\circ\theta)$.

Student Question. *Why is $\|\overline{f}-\overline{g}\|_p^p$ equal to this liminf?*

Answer. Because $\overline{f}$ is the pointwise limit of the ergodic average of f, and likewise with $\overline{g}$ and g. So you have $\frac{1}{n}S_n(f) \to \overline{f}$ pointwise, and $\frac{1}{n}S_n(g) \to \overline{g}$; this implies that

$$\left| \frac{S_n(f)}{n} - \frac{S_n(g)}{n} \right|^p \to \left|\overline{f} - \overline{g}\right|^p.$$

So the function on the right is the limit of the function on the left, which is the same as the $\liminf$. (Here we actually have a limit, not just a $\liminf$; we just wrote it with a $\liminf$ to make it clear how we're applying Fatou's lemma.)

These are the main ergodic theorems we're going to use (ergodic theory is of course much deeper, but these are just the essentials). The lecture notes may give a proof of Birkhoff's ergodic theorem (it's also in the textbooks), but it's a bit technical, so out of the scope of this course.

§14.2 Weak and strong law of large numbers

Now we'll try to prove some big theorems of this course. We'll first prove the strong law of numbers. We'll first do this assuming the fourth moments are bounded from above and finite, and then we'll do the general case.

The setup is that we have a sequence of independent identically distributed random variables $(X_n)_{n\in\mathbb{N}}$, and the're in L^1 with some common mean $\mathbb{E}[X_i] = \mu$. And we let $S_n = X_1 + \cdots + X_n$.

Here we have two things. The first is the *weak* law of large numbers. (This is a consequence of the strong law, but we will also state it separately.)

Theorem 14.4 (Weak law of large numbers)
We have $\frac{S_n}{n} \to \mu$ in probability, provided that $\mathbb{E}[X_i^2] < \infty$.

The strong law of large numbers is stronger than this statement, and we're going to prove it.

Theorem 14.5 (Strong law of large numbers)
We have $\frac{S_n}{n} \to \mu$ almost surely, providedthat $\mathbb{E}[|X_i|] < \infty$ (i.e., that the random variables are in L^1).

(Note that this is a weaker assumption than that $\mathbb{E}[X_i^2] < \infty$.) This is stronger than the weak law of large numbers, because convergence in probability is implied by almost sure convergence.

We're going to prove two versions with different assumptions, but this is the one we care about. We're going to use Birkhoff's ergodic theorem to extract the existence of the limit, and then use von Neumann's ergodic theorem to say that the limit has to be the mean.

§14.2.1 Strong LLN with finite fourth moments

We'll first prove a version of the strong law of large numbers assuming finite fourth moments.

Theorem 14.6 (Strong LLN with finite fourth moments)
Let (X_n) be a sequence of independent random variables with the property that there exists $\mu \in \mathbb{R}$ and $M > 0$ such that $\mathbb{E}[X_n] = \mu$ and $\mathbb{E}[X_n^4] \leq M$ for all $n \in \mathbb{N}$. Then letting $S_n = X_1 + \cdots + X_n$, we have that $\frac{S_n}{n} \to \mu$ almost surely as $n \to \infty$.

Here we have to be a bit careful — the assumption is not the same as before, and neither is stronger than the other. Here we assumed we have independence, but not necessarily *identical laws*. In the strong LLN, we assume the law is the same (the random variables are i.i.d.). Here we don't assume the laws are the same (just that they're independent and have the same mean), but we also assume a uniform bound on the fourth moments.

The proof is actually completely elementary — we're not going to use tools from ergodic theory here, it's more or less a combinatorial argument.

Proof. First, we can reduce the problem to the case that $\mu = 0$ by defining $Y_n = X_n - \mu$ (so we're centering the random variables) — if we do this, then the Y_ns are still independent, and $\mathbb{E}[Y_n] = \mathbb{E}[X_n] - \mu = 0$ by linearity of expectation; and we have

$$\mathbb{E}[Y_n^4] \leq 2^4 \mathbb{E}[X_n^4 + \mu^4] \leq 16(\mu^4 + M)$$

for all n. So the Y_ns satisfy our assumptions, which means if we know the claim in the case $\mu = 0$, then we can extract the claim in general (it'd say that $X_n - \mu \to 0$ almost surely, so $X_n \to \mu$.)

So from now, we assume that $\mu = 0$. Now we just expand S_n^4; if we do this expansion, we get

$$\mathbb{E}[S_n^4] = \mathbb{E}\left[\sum_k X_k^4\right] + 6\sum_{i<j} \mathbb{E}[X_i^2 X_j^2] + A\sum_{i<j} \mathbb{E}[X_i X_j^3] + B\sum_{i<j<k} \mathbb{E}[X_i X_j X_k^2]$$

(for some constants). And now the point is that the last two terms are 0 — this is because we can factorize them out as $\mathbb{E}[X_i X_j^3] = \mathbb{E}[X_i]\mathbb{E}[X_j^3]$, and $\mathbb{E}[X_i] = 0$. The same is true for the final term — we have $\mathbb{E}[X_i X_j X_k^2] = \mathbb{E}[X_i]\mathbb{E}[X_j]\mathbb{E}[X_k^2] = 0$. So it remains to estimate the first two terms.

The first term is bounded by our assumption — by linearity of expectation we have $\mathbb{E}[\sum_k X_k^4] = \sum_k \mathbb{E}[X_k^4] \leq nM$ (because we assumed $\mathbb{E}[X_k^4] \leq M$ for all k).

Now for the second term, we can apply Hölder's inequality or Jensen's inequality to say that for $i \neq j$, we have

$$\mathbb{E}[X_i^2 X_j^2] = \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] \leq \sqrt{\mathbb{E}[X_i^4]\mathbb{E}[X_j^4]} \leq M$$

by Jensen's inequality. So we have some nice bounds on the first and second terms, and this implies the expectation of the second term is

$$\mathbb{E}\left[6\sum_{i<j} X_i^2 X_j^2\right] \leq 6 \cdot \frac{n(n-1)}{2} \cdot M = 3n(n-1)M.$$

So overall, we have

$$\mathbb{E}[S_n^4] \leq nM + 3n(n-1)M \leq 3n^2 M,$$

which implies that

$$\mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] \leq \frac{3M}{n^2}.$$

And this implies that we have

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right] \leq \sum_{n=1}^{\infty}\frac{3M}{n^2} < \infty.$$

And if the expectation of a random variable is finite, that random variable has to be finite almost surely; so this means $\sum(\frac{S_n}{n})^4 < \infty$ almost surely. And if the sum of a sequence is finite, then the sequence has to converge to 0; so we get that $\frac{S_n}{n} \to 0$ almost surely, which is exactly what we wanted. □

The advantage of this proof is that here we haven't assumed the variables have the same law, because that's not necessary. We just assumed independence and equal means, and a uniform bound on the fourth moments. That's the first version of the strong law of large numbers, and it's not the same as the first version (which assumes they have the same law and doesn't have the fourth moments assumption).

This is the first main limit theorem of the course.

Student Question. *Is there something special about the fourth moment?*

Answer. If you used second moments instead, then the bound that you would get on the second moment would give you a sum of the form $\sum 1$ or $\sum \frac{1}{n}$; and this wouldn't converge. The point of the 4th moments is to get a finite sum. You can also use higher moments, but if higher moments are bounded, then so are the 4th moments.

§14.2.2 The strong law of large numbers

Now we'll prove the first form of the strong law of large numbers. Here we don't have the assumption on the fourth moments, so we're going to need Birkhoff. This is the second main limit theorem of the course; the third is the central limit theorem, which we'll prove either this lecture or the next.

Theorem 14.7 (Strong law of large numbers)

Let (Y_n) be an i.i.d. sequence of integrable random variables with mean $\nu \in \mathbb{R}$. Then letting $S_n = Y_1 + \cdots + Y_n$, we have

$$\frac{S_n}{n} \to \nu \text{ almost surely.}$$

We're going to use ergodic theory to prove this (there are other proofs in the textbook, but ergodic theory gives a clean one; the main thing is to just set things up in the right way that we can apply ergodic theory, and we've already described how to do the setup using Bernoulli shifts).

Proof. Let m be the law of Y_1 (which is the same as the law of Y_n for all n), and consider the random variable $Y = (Y_1, Y_2, \dots)$, which is a random variable $\Omega \to \mathbb{R}^{\mathbb{N}} = E$ (as defined last class — the space of all sequences of real numbers). We then let $(E, \mathcal{E}, \mu)$ be the canonical space associated with m, as defined last class (the space corresponding to the Bernoulli shift — the *canonical space* is another name of the space we constructed last time), where $\mu = \mathbb{P} \circ Y^{-1}$ is the pushforward of the original probability space where everything lives, with respect to the random variable Y.

Last time, we defined the function θ that shifts the sequence one place to the right, and we defined $f: E \to \mathbb{R}$ as $(x_n) \mapsto x_1$. This is the sequence we're going to apply ergodic theorems to. (Recall that the shift map is the map $\theta: E \to E$ defined by $(x_1, x_2, x_3, \dots) \mapsto (x_2, x_3, x_4, \dots)$ — it just shifts the sequence one place to the right.)

Then as we said earlier, we can write $Y_n = (f \circ \theta^{n-1})(Y)$; this means we have $S_n(f)(Y) = Y_1 + \cdots + Y_n$. (That's the key point here.) This implies that

$$\frac{S_n(f)}{n}(Y) = \frac{Y_1 + \cdots + Y_n}{n} \to \overline{f}$$

as $n \to \infty$, for μ-almost every Y in E. This is by Birkhoff's theorem — so we have the *existence* of the limit (almost everywhere).

Now what remains is to *identify* $\overline{f}$. And for that, we apply von Neumann's ergodic theorem to say that here the convergence is also in L^1. Specifically, we also have that

$$\frac{S_n(f)}{n} \to \overline{f}$$

in $L^1(E)$, by von Neumann's ergodic theorem. So now we need to argue that $\overline{f}$ is just ν. To do this, we use ergodicity to say that $\overline{f}$ is constant.

Since $\overline{f}$ is θ-invariant (by Birkhoff's ergodic theorem) and θ is ergodic (we proved this last lecture), this implies that there exists a constant $c \in \mathbb{R}$ such that $\overline{f} = c$ μ-almost everywhere (this follows by the fact that our system is ergodic). And this implies that $c = \mu(\overline{f})$ (because $\overline{f} = c$ almost everywhere). But by L^1-convergence, we also have

$$\mu(\overline{f}) = \lim_{n \to \infty} \mu\left(\frac{S_n(f)}{n}\right).$$

But we always have $\mu(\frac{S_n(f)}{n}) = \nu$.

And that's exaclty what we wanted to prove. □

This proof is clean, in the sense that it's not very technical; it just uses the things ew've learned from ergodic theory. And once you have the right setup (with Bernoulli shifts), everything is straightforward.

That's the second main theorem of the course; intuitively, it says that the average of i.i.d. integrable random variables behaves roughly like the mean of those variables (in the long run).

The next main theorem is the central limit theorem, which we'll do next lecture.

Student Question. *How did we use von Neumann?*

Answer. To obtain L^1 convergence. We needed that because we're using it in the last line to say that $\mu(\overline{f}) = \lim_{n\to\infty} \mu(\frac{S_n(f)}{n})$. We know $\overline{f}$ is some constant, and we want to identify that constant. If we have L^1 convergence, then this constant has to be $\mathbb{E}[\overline{f}]$ and that has to be the limit of the expectation of the left-hand side. But we always have $\mathbb{E}[Y_j] = \nu$, so $\mathbb{E}[\sum Y_j] = n\nu$, which means $\mathbb{E}[\frac{\sum Y_j}{n}] = \nu$ — so this is the same for all n, which means the limit of that integral is ν. So we get $\mathbb{E}[\overline{f}] = \nu$, and because $\overline{f}$ is constant, we get that $\overline{f}$ itself is ν.

Student Question. *Do we have a measure of how fast the average converges to the mean?*

Answer. There are some estimates in certain cases — stronger versions that keep track of the rate of convergence. But of course the rate is dependent on the distribution.

Student Question. *Is the $\overline{f}$ we get from von Neumann the same as from Birkoff?*

Answer. Yes — because whenever you have L^1 convergence, you also have almost sure convergence along a subsequence. So along a subsequence, you have almost sure convergence to $\overline{f}$; but you also have convergence almost surely to the $\overline{f}$ coming from Birkhoff, so the two functions have to be the same (almost surely).

Next time we'll prove the central limit theorem and also move to the second part of the course, on conditional expectations and martingales and Brownian motion.

§15 October 29, 2024

Last class, we proved the strong law of large numbers under a fourth moment assumption, where we had independent but not necessarily identically distributed variables, and we knew their means were the same and we had a uniform bound on their fourth moments. We also proved the strong law of large numbers (in the most commonly used version), which only requires that you have an i.i.d. sequence in L^1. For that, we used ergodic theory. Prof. Kavvadias presented the ergodic theory proof for two reasons — it motivates us to learn more about ergodic theory, and the proof is clean. There are other proofs in the references, which are more elementary but more tedious.

§15.1 Central limit theorem

Today we'll state and prove the third main limit theorem, the central limit theorem. With this, we'll complete the first part of the course.

Theorem 15.1 (Central limit theorem)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$ (for each i). Then if we set $S_n = X_1 + \cdots + X_n$, then we have

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

We use $\xrightarrow{d}$ to denote convergence in distribution; this means for every fixed x, we have

$$\mathbb{P}\left[\frac{S_n}{\sqrt{n}} \le x\right] \to \int_{-\infty}^{x} \frac{e^{-y^2//2}}{\sqrt{2\pi}}\, dy$$

(the integral on the right is the distribution function of a standard Gaussian with mean 0 and variance 1).

The central limit theorem states that it doesn't matter what distribution X_n has (we just require that they're independent and have the same distribution, with mean 0 and variance 1); the long-range behavior of their sums behaves like a Gaussian.

Proof. We need to prove convergence in distribution, and there are several ways to do this. One is to directly compute the probability, but that's tedious, so we won't. Recall that a sequence of random variables converges in distribution if and only if the characteristic functions converge pointwise. So that's what we're

going to show — that the characteristic function of $n^{-1/2}S_n$ converges pointwise to the characteristic function of Z, a standard normal variable.

Let ϕ be the characteristic function of X_1 (for example) — i.e.,

$$\phi(u) = \mathbb{E}[e^{iuX_1}].$$

Since $\mathbb{E}[X_1^2] = 1 < \infty$, we can differentiate under the expectation twice (by applying the dominated convergence theorem) to obtain that

$$\phi'(u) = \mathbb{E}[iX_1 e^{iuX_1}] \quad \text{and} \quad \phi''(u) = \mathbb{E}[-X_1^2 e^{iuX_1}]$$

for all $u \in \mathbb{R}$. (We can differentiate inside the expectation because of the second moment assumption — in general, it's not true that we can differentiate inside the expectation, and here we're using the assumption to use the dominated convergence theorem — the dominating function the first time we differentiate is $|X_1|$, and the second time it's X_1^2). Taking $u = 0$, we get that

$$\phi(0) = 1, \quad \phi'(u) = 0, \quad \phi''(u) = -1.$$

Then we can apply Taylor expansion (because ϕ'' is twice differentiable,and its second derivative is continuous — again by the dominated convergence theorem). If we Taylor expand ϕ at 0, then we get

$$\phi(u) = 1 - \frac{u^2}{2} + o(u^2)$$

as $u \to 0$. (By $o(u^2)$ we mean that if we take this term and divide it by u^2 and let $u \to 0$, then this term tends to 0.)

Now we want to show that the characteristic function of $n^{-1/2}S_n$ converges pointwise to that of a standard normal variable. We'll let ϕ_n denote the characteristic function of $n^{-1/2}S_n$; this is by definition

$$\phi_n(u) = \mathbb{E}[e^{iuS_n/\sqrt{n}}].$$

Because S_n is a sum, this is a product; and for independent random variables, the expectation of a sum is the product of expectations. So

$$\phi_n(u) = \prod_{j=1}^{n} \mathbb{E}[e^{iuX_j/\sqrt{n}}] = \phi(u/\sqrt{n})^n.$$

And as $n \to \infty$, we'll have $u/\sqrt{n} \to 0$ (for any fixed u), so we can take advantage of the Taylor expansion — so we can write this as

$$\phi_n(u) = \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n.$$

And now you could take the limit directly, but a more convenient way is to take the logarithm of both sides and show $\log \phi_n(u)$ converges to $\log \phi_Z(u)$. We have

$$\log \phi_n(u) = n \cdot \log\left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right).$$

And using the behavior of log near 1, this means

$$\log \phi_n(u) = -\frac{u^2}{2} + o(1)$$

as $n \to \infty$ (we know $\frac{\log(1+x)}{x} \to 1$ as $x \to 0$, for example by L'Hopital's rule), which implies that as $n \to \infty$, we have

$$\log \phi_n(u) \to -\frac{u^2}{2}.$$

This implies that

$$\phi_n(u) \to e^{-u^2/2}$$

as $n \to \infty$. But $e^{-u^2/2}$ is precisely the characteristic function of a standard normal random variable (we've proved this earlier). So we have pointwise convergence of characteristic functions, which means we have convergence in distribution (since we showed earlier that these are equivalent). □

Student Question. *This proof kind of feels like magic, and it feels like there's something deep going on with the characteristic functions; is there any intuition for this?*

Answer. The intuition probably comes from the computation that

$$\phi_n(u) = \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n.$$

It's known that $(1 + \frac{x}{n})^n \to e^x$. And you know that the characteristic function has the form on the left as $n \to \infty$, so it should converge to some exponential. And what random variable has a characteristic function that looks like an exponential? The most natural answer is the standard normal Gaussian, which has a characteristic function of this form.

That's the first part of the course. We did all this preparation to prove these three limit theorems, which were the really important theorems of the course. Now the second part is a bit more advanced, but we've become more experienced (we've seen the basics and rigorous construction of probability, and we've been exposed to proofs).

§15.2 Conditioning on an event

The next thing we'll talk about is conditional expectation. We've probably seen this — what's the conditional probability of B given A? But we're going to generalize this — instead of just conditioning on an event A, we'll condition on an entire σ-algebra (and obtain a random variable).

As usual, we'll work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We'll briefly review what the conditional probability of one event with respect to another is.

Definition 15.2. Let $A, B \in \mathcal{F}$ be events such that $\mathbb{P}[B] > 0$. Then the conditional probability of A given B is defined as

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Somehow we want to generalize this to what happens if we replace B by a σ-algebra.

Student Question. *Can this be thought of as the restriction measure (where we're restricting the space to B)?*

Answer. Yes — this is exactly the restriction measure, normalized by $\mathbb{P}[B]$ to be a probability measure.

Similarly, we can define the conditional expectation of a random variable (replacing A by a random variable). For this, we'll need the assumption that $X \in L^1$.

Definition 15.3. Let X be a random variable with the property that $\mathbb{E}[|X|] < \infty$, and let A be an event with $\mathbb{P}[A] > 0$. Then the conditional expectation of X given A is defined by

$$\mathbb{E}[X \mid A] = \frac{\mathbb{E}[X \cdot \mathbf{1}_A]}{\mathbb{P}[A]}.$$

(So $\mathbb{E}[X \mid A]$ is a number; and since we're dividing by a real number $\mathbb{P}[A]$, we of course need this number to be nonzero.)

The goal of today's lecture is to make sense of the situation where we're starting with X being a L^1 random variable, but we have a σ-algebra instead of an event.

§15.3 Motivation — the discrete case

As usual, we'll start with a simpler case — the discrete case (for intuition). Assume that Ω can be partitioned into a countable collection of disjoint events — so $\Omega = \bigcap_{i \in I} B_i$ where I is countable and the B_i are disjoint (and $B_i \in \mathcal{F}$ for all i). For now we'll study spaces with this particular structure. Not all spaces have this structure (but for instsance, any finite or countable space does have this structure).

And we'll consider $\mathcal{G} = \{\bigcup_{j \in J} B_j \mid J \subseteq I\}$. It's not hard to show that this is a σ-algebra — specifically, it's $\mathcal{G} = \sigma(B_i \mid i \in I)$.

What we're going to compute now is the conditional expectation of x given this σ-algebra $\mathcal{G}$. What's the natural thing to do? We'll define X' as our candidate for the conditional expectation of X given $\mathcal{G}$, as

$$X' = \sum_{i \in I} \mathbb{E}[X \mid B_i] \cdot \mathbf{1}_{B_i}.$$

Here, the convention we use is that $\mathbb{E}[X \mid B_i] = 0$ if $\mathbb{P}[B_i] = 0$ (because it might be the case that some of the B_i's have 0 probability; but we only care about the events in the partition with positive probability). This is a very natural candidate for the conditional expectation.

This X' satisfies several properties:

- It's $\mathcal{G}$-measurable — that's because $\mathbf{1}_{B_i}$ is $\mathcal{G}$-measurable, and the sum of measurable functions is still measurable. (Note that this is stronger than saying it's measurable with respect to $\mathcal{F}$.)
- It's integrable — this is because

  $$\mathbb{E}[|X'|] \le \sum_{i \in I} \mathbb{E}[|X| \cdot \mathbf{1}_{B_i}] = \mathbb{E}[|X|] < \infty$$

  (here we're using the fact that the events are disjoint).
- If we take any $G \in \mathcal{G}$, then we have

  $$\mathbb{E}[X \cdot \mathbf{1}_G] = \mathbb{E}[X' \cdot \mathbf{1}_G].$$

  (This is not hard to see — it follows by the definition.) So the integrals of X' agree with those of X when restricted to events in this smaller σ-algebra.

Our goal is to take *any* probability space and sub-σ-algebra, and always construct some X' with these three properties. That's the main goal of today's lecture — to prove the existence of such an X' in a general setting.

In general, it's not easy to get an explicit form; but in this case, we do. The theorem we're going to prove is just one of existence (and we'll also prove uniqueness, up to a set of measure 0).

Student Question. *Where did we use discreteness?*

Answer. We used discreteness to say that $\sigma(B_i \mid i \in I)$ has the form $\{\bigcup_j B_j \mid J \subseteq I\}$ — for this, we use the fact that I is countable. We also write $\sum_{i \in I} \mathbb{E}[X \mid B_i]\mathbf{1}_{B_i}$ — if you want to sum over uncountable sets, you need a different definition of the sum.

§15.4 Conditional expectation

First we'll state the main theorem.

Theorem 15.4

Let X be an integrable random variable, and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. Then there exists a random variable Y such that the following properties hold:

(a) Y is $\mathcal{G}$-measurable.

(b) Y is integrable, and for every $G \in \mathcal{G}$, we have $\mathbb{E}[Y \cdot \mathbf{1}_G] = \mathbb{E}[X \cdot \mathbf{1}_G]$.

Furthermore, if Y' also satisfies (a) and (b), then $Y = Y'$ almost surely (i.e., with probability 1).

So we can always find such a random variable Y which satisfies the properties mentioned earlier, and it's unique — any other random variable with these properties agrees with Y with probability 1.

Definition 15.5. We call Y the conditional expectation of X given $\mathcal{G}$, denoted by $\mathbb{E}[X \mid \mathcal{G}]$.

In general, it's not easy to construct the conditional expectation explicitly (except in special cases, like the one we saw earlier).

We'll start with uniqueness, which is the easier part.

Proof of uniqueness. Suppose that both Y and Y' satisfy (a) and (b). Then if we define $A = \{Y > Y'\}$, we have $A \in \mathcal{G}$ (because Y and Y' both belong to $\mathcal{G}$ — this was probably in the first or second problem set). Then by (b), we have that

$$\mathbb{E}[(Y - Y') \cdot \mathbf{1}_A] = \mathbb{E}[Y \cdot \mathbf{1}_A] - \mathbb{E}[Y' \cdot \mathbf{1}_A] = \mathbb{E}[X \cdot \mathbf{1}_A] - \mathbb{E}[X \cdot \mathbf{1}_A] = 0.$$

And the random variable $(Y - Y') \cdot \mathbf{1}_A$ is nonnegative; and if the expectation of a nonnegative random variable is 0, then it has to be 0 almost surely. So this implies $(Y - Y')\mathbf{1}_A = 0$ almost surely, and therefore $\mathbb{P}[A] = 0$. This means $Y \leq Y'$ almost surely.

Similarly, we can show that $Y \geq Y'$ almost surely (by replacing A with the event that $Y < Y'$). These two inequalities together imply $Y = Y'$ almost surely, giving uniqueness. □

Now what's less easy is proving existence. We're going to prove this in three different steps. In the first step, we'll assume $X \in L^2$ (which is not necessarily the case in general). In the second, we'll prove the claim when X is nonnegative. Finally, we'll conclude the general case, using limiting arguments as usual.

§15.4.1 Existence when $X \in L^2$

We'll now prove existence starting with the first step, where we assume that $X \in L^2$. Recall that L^2 has the structure of a Hilbert space — specifically, the space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is equipped with the inner product

$$\langle U, V \rangle = \mathbb{E}[U \cdot V].$$

We've shown that it's a Hilbert space. We've also shown that if we consider $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ instead, then this is a closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. (We'll use $\mathcal{L}^2(\mathcal{F})$ and $\mathcal{L}^2(\mathcal{G})$ as shorthand.) And from the theory of Hilbert spaces, this means we can write

$$\mathcal{L}^2(\mathcal{F}) = \mathcal{L}^2(\mathcal{G}) + \mathcal{L}^2(\mathcal{G})^\perp,$$

i.e., we can decompose $\mathcal{L}^2(\mathcal{F})$ into $\mathcal{L}^2(\mathcal{G})$ and its orthogonal complement. This means we can write $X = Y + Z$ where $Y \in \mathcal{L}^2(\mathcal{G})$ and $Z \in \mathcal{L}^2(\mathcal{G})^\perp$. (This follows by the theory of Hilbert spaces that we briefly mentioned in a previous chapter.)

Now we have a candidate Y — we define $\mathbb{E}[X \mid \mathcal{G}]$ as Y. Clearly (a) is satisfied, because we chose $Y \in \mathcal{L}^2(\mathcal{G})$, which means it's $\mathcal{G}$-measurable. So now we want to show that (b) is also satisfied.

For this, fix any $A \in \mathcal{G}$. Then we have that $\mathbb{E}[X \cdot \mathbf{1}_A] = \langle X, \mathbf{1}_A \rangle$. But we can write $X = Y + Z$, so

$$\mathbb{E}[X \cdot \mathbf{1}_A] = \langle X, \mathbf{1}_A \rangle = \langle Y, \mathbf{1}_A \rangle + \langle Z, \mathbf{1}_A \rangle.$$

But recall that $A \in \mathcal{G}$, so $\mathbf{1}_A \in \mathcal{L}^2(\mathcal{G})$; and $Z \in \mathcal{L}^2(\mathcal{G})^\perp$, so their inner product has to be 0 by the definition of the orthogonal complement. This means $\langle Z, \mathbf{1}_A \rangle = 0$, while $\langle Y, \mathbf{1}_A \rangle = \mathbb{E}[Y \cdot \mathbf{1}_A]$. So (b) is satisfied.

> **Student Question.** *When did we show that Y is integrable?*
>
> **Answer.** It's in L^2, and for a finite measure space, any random variable in L^2 is also in L^1. (We're always assuming we're in a probability space, which has finite measure; but this proof shows the statement would hold in any space with finite measure.)

§15.4.2 Existence for nonnegative random variables

The next step is to prove the same statement for nonnegative random variables, using the fact that we can construct the conditional expectation in L^2.

Here, suppose that $X \geq 0$ (it's not necessarily in L^2, but of course it's still in L^1). We want to somehow apply a limit theorem, so we need to cut off X so that it doesn't get too large. For this, we consider $X_n = X \wedge n$ (so that $X_n \leq n$ is bounded and nonnegative for each n).

Since X_n is bounded, it's certainly in L^2; this means that from the first step, we can define

$$Y_n = \mathbb{E}[X_n \mid \mathcal{G}]$$

(and this is well-defined, because $X_n \in L^2$ and the conditional expectation is well-defined for L^2 random variables).

Now we need to somehow take a limit, and we need to justify why the limit exists. This is for the following reason — X_n is nondecreasing, so Y_n also has to be nondecreasing by the way it's constructed. Specifically, in the special case that $X \geq 0$ and $X \in L^2$, we also have that $Y \geq 0$ almost surely (so nonnegativity is preserved); this follows from the definition of Y.

So this implies Y_n is also nondecreasing, since X_n is nondecreasing. In particular, this means if we define $Y = \lim_{n\to\infty} Y_n$ to be the pointwise limit of the Y_n, this limit exists almost surely. And all the Y_n are $\mathcal{G}$-measurable, so the same is true for Y. (We're permitting Y to be infinite here; but we're going to show it's in L^1, so the measure of the set where it's infinite is 0. Whenever you have a nondecreasing sequence, the limit always exists — it might be infinite, but it exists. The good thing about the monotone convergence theorem is that it doesn't mind if the function is infinite.)

This is our candidate; what remains to show is that it satisfies (b). For this, we're going to apply the monotone convergence theorem. If we fix an event $A \in \mathcal{G}$, then we have $\mathbb{E}[Y_n \cdot \mathbf{1}_A] = \mathbb{E}[X_n \cdot \mathbf{1}_A]$ for all n, because of the first step. And now we can take limits — the left-hand side converges pointwise increasingly to $Y \cdot \mathbf{1}_A$. Similarly, $X_n \cdot \mathbf{1}_A$ converges pointwise increasingly to $X \cdot \mathbf{1}_A$. This means $\mathbb{E}[Y_n \cdot \mathbf{1}_A] \to \mathbb{E}[Y \cdot \mathbf{1}_A]$, and $\mathbb{E}[X_n \cdot \mathbf{1}_A] \to \mathbb{E}[X \cdot \mathbf{1}_A]$ (by the monotone convergence theorem). But since the terms of the sequences are equal, this means $\mathbb{E}[Y \cdot \mathbf{1}_A] = \mathbb{E}[X \cdot \mathbf{1}_A]$.

Finally, if we take A to be the whole space Ω (which is in $\mathcal{G}$ because it's a σ-algebra), then we have $\mathbb{E}[Y] = \mathbb{E}[|Y|] = \mathbb{E}[X] < 0$ (because Y is finite, and $X \in L^1$). So Y is indeed integrable. (In particular, this

means Y has to be finite almost surely; so if you want to get it to be finite everywhere, you can just take the parts where it's infinite and reset them to 0.)

This gives the conditional expectation in the case of nonnegative random variables.

Student Question. *How does it follow from the definition that the Y_n are nonnegative and increasing, and that the limit exists almost everywhere?*

Answer. The construction for L^2 means that if $X \geq 0$ then $Y \geq 0$ almost surely. And if we apply this statement to $X_{n+1} - X_n$ (so the random variable we get out is $Y_{n+1} - Y_n$), we get the result.

For how we get the first statement (that if $X \geq 0$, then $Y \geq 0$), one way to show this is to consider $\mathbb{E}[Y \cdot \mathbf{1}_{Y<0}]$. Then this has to be equal to $\mathbb{E}[X \cdot \mathbf{1}_{Y<0}]$ by definition. But since X is always nonnegative, then $X \cdot \mathbf{1}_{Y<0}$ is always nonnegative. On the other hand, $Y \cdot \mathbf{1}_{Y<0}$ is nonpositive; so we've shown the expectation of a nonpositive random variable is nonnegative. So it has to be 0 almost everywhere.

§15.5 Existence for the general case

Finally, the last part is to conclude the proof. For this, things are quite simple — as usual, we just write X as the difference of two nonnegative random variables, and consider the difference of their conditional expectations.

Here, X is a general random variable in $L^1(\mathcal{F})$. Then we can write $X = X^+ - X^-$ where $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$ (we've done this many times so far). Then we set Y (our candidate for $\mathbb{E}[X \mid \mathcal{G}]$) as

$$Y = \mathbb{E}[X^+ \mid \mathcal{G}] - \mathbb{E}[X^- \mid \mathcal{G}].$$

Then it's easy to see (by the linearity of the integral) that (a) and (b) are satisfied. This concludes the proof of the theorem.

§15.6 Properties of conditional expectations

To finish the lecture, we'll write down some basic properties of conditional expectations (they're immediate from the definition, so we won't prove them).

Proposition 15.6

Let $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, and fix a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Then we have the following properties:

(1) We have $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.

(2) If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X \mid \mathcal{G}] = X$ almost surely.

(3) If X is independent of $\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ almost surely.

(4) If $X \geq 0$ almost surely, then $\mathbb{E}[X \mid \mathcal{G}] \geq 0$ almost surely.

(5) Conditional expectation is linear — for any $\alpha, \beta \in \mathbb{R}$, we have

$$\mathbb{E}[\alpha X + \beta Y \mid \mathcal{G}] = \alpha \mathbb{E}[X \mid \mathcal{G}] + \beta \mathbb{E}[Y \mid \mathcal{G}]$$

(almost surely).

(6) We have $|\mathbb{E}[X \mid \mathcal{G}]| \leq \mathbb{E}[|X| \mid \mathcal{G}]$ almost surely.

(1) follows from applying property (b) where A is the entire space Ω. For (2), it's straightforward to see that if X is $\mathcal{G}$-measurable, then X itself satisfies (a) and (b), and then we can use uniqueness. (In the definition

of conditional expectation, the only thing preventing us from just taking $Y = X$ is that X may not be $\mathcal{G}$-measurable.) In (3), it's clear that $\mathbb{E}[X]$ (which is a constant) is in L^1; and (b) is satisfied because of the definition of independence. We proved (4) a few minutes ago. (5) is fairly straightforward. And (6) is true because if we integrate these random variables over any $A \in \mathcal{G}$, we'll see that the integral of the left is at most the integral of the right (by taking absolute values in $\mathbb{E}[Y \cdot \mathbf{1}_G] = \mathbb{E}[X \cdot \mathbf{1}_G]$, and putting the absolute values inside).

And all the stuff we've learned, like Fatou's lemma, pass into conditional expectations as well; we're going to talk about that next lecture.

§16 October 31, 2024

Last lecture, we completed the first part of the course, which was about proving the main limit theorems in probability. For that, we had to do a lot of preparation — we had to rigorously construct probability and Lebesgue integration and prove a lot of stuff about Hilbert spaces and ergodic theory and Fourier analysis. Last lecture, we also proved the existence and uniqueness of the conditional expectation of a random variable with respect to a fixed σ-algebra, which generalizes the conditional probability of an event given another.

At the end of the class, we reviewed some basic properties of the conditional expectation, which come directly from the definitions.

§16.1 Convergence theorems for conditional expectations

We've seen some convergence theorems (e.g., Fatou's lemma, dominated convergence, monotone convergence, and so on); these theorems have analogs in conditional expectation, which we'll prove today.

In all the following statements, we let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra.

Proposition 16.1 (Conditional monotone convergence theorem)

Suppose that (X_n) is an increasing sequence of (real-valued) random variables and $X_n \geq 0$ for all n, and $\lim_{n\to\infty} X_n = X$ almost surely. Then we have $\mathbb{E}[X_n \mid \mathcal{G}] \nearrow \mathbb{E}[X \mid \mathcal{G}]$ almost surely.

Note that $\mathbb{E}[X_n \mid \mathcal{G}]$ are random variables; and the conclusion says they converge increasingly to the random variable $\mathbb{E}[X \mid \mathcal{G}]$.

This is the same as the monotone convergence theorem, but with conditional expectations rather than expectations. Note that this implies the usual monotone convergence theorem, if we take $\mathcal{G}$ to be the trivial σ-algebra.

Proposition 16.2 (Conditional Fatou's lemma)

If $X_n \geq 0$ for all n, then $\mathbb{E}[\liminf_{n\to\infty} X_n \mid \mathcal{G}] \leq \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}]$ almost surely.

This is the same as Fatou's lemma; we've just replaced the expectations with conditional expectations.

Proposition 16.3 (Conditional dominated convergence theorem)

If $X_n \to X$ as $n \to \infty$ almost surely, and $|X_n| \leq Y$ (for all n) for some random variable $Y \in \mathcal{L}^1(\mathcal{F})$, then $\mathbb{E}[X_n \mid \mathcal{G}] \to \mathbb{E}[X \mid \mathcal{G}]$ almost surely.

(Saying $Y \in \mathcal{L}^1(\mathcal{F})$ is the same as saying Y is integrable.)

Finally, we also have an analog of Jensen's inequality.

Proposition 16.4 (Conditional Jensen's inequality)
If $X \in \mathcal{L}^1(\mathcal{F})$ and $\varphi: \mathbb{R} \to (-\infty, +\infty]$ is any convex function, such that either $\varphi(X) \in \mathcal{L}^1(\mathcal{F})$ or φ is nonnegative, then

$$\mathbb{E}[\phi(X) \mid \mathcal{G}] \geq \varphi(\mathbb{E}[X \mid \mathcal{G}])$$

almost surely.

Note that we allow φ to take values in $+\infty$, but not $-\infty$.

In particular, taking $\varphi(x) = x^p$ (this function is convex as long as $p \geq 1$), we get the following statement:

Corollary 16.5
For all $1 \leq p < \infty$, we have $\|\mathbb{E}[X \mid \mathcal{G}]\|_p \leq \|X\|_p$.

So these are the analogous statements to our usual convergence theorems, just with expectations replaced with conditional expectations. The proofs are similar in spirit to what we've did in the unconditional case.

Proof of 16.1. Let $Y_n = \mathbb{E}[X_n \mid \mathcal{G}]$. In the previous lecture, we showed that since the X_n are nonnegative and increase pointwise to X (almost surely), we have that the Y_n are also nonnegative and increasing (almost surely). We can then set $Y = \lim_{n\to\infty} Y_n$ (the pointwise limit — this exists because the Y_n are increasing, though it might take infinite values). To complete the proof, it suffices to show that $Y = \mathbb{E}[X \mid \mathcal{G}]$ (almost surely); and by uniqueness, to do this, it suffices to show that Y satisfies the two properties that characterize the conditional expectation.

First, Y is clearly $\mathcal{G}$-measurable, since each Y_n is $\mathcal{G}$-measurable and the pointwise limit of $\mathcal{G}$-measurable functions is also $\mathcal{G}$-measurable.

Also, if we fix any set $A \in \mathcal{G}$, then by the monotone convergence theorem we have that

$$\mathbb{E}[X \cdot \mathbf{1}_A] = \liminf_{n\to\infty} \mathbb{E}[X_n \cdot \mathbf{1}_A]$$

(since $X_n \cdot \mathbf{1}_A$ converges increasingly pointwise to $X \cdot \mathbf{1}_A$). But for all n we have $\mathbb{E}[X_n \cdot \mathbf{1}_A] = \mathbb{E}[Y_n \cdot \mathbf{1}_A]$, by the definition of the conditional expectation. And now we can apply the monotone convergence theorem again — $Y_n \cdot \mathbf{1}_A$ converges pointwise increasingly to $Y \cdot \mathbf{1}_A$, so we have

$$\mathbb{E}[X \cdot \mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[X_n \cdot \mathbf{1}_A] = \lim_{n\to\infty} \mathbb{E}[Y_n \cdot \mathbf{1}_A] = \mathbb{E}[Y \cdot \mathbf{1}_A].$$

So we've shown Y satisfies both the conditions in the definition of conditional expectation, which by uniqueness means $Y = \mathbb{E}[X \mid \mathcal{G}]$. □

(Here we're using the unconditional monotone convergence theorem in our proof.)

Proof of conditional Fatou. As in the proof of Fatou's lemma, we consider the sequence $\inf_{k>n} X_k$. This sequence is increasing in n (by definition), and

$$\lim_{n\to\infty} (\inf_{k>n} X_k) = \liminf_{n\to\infty} X_n$$

by definition. Now we can just apply Proposition 16.1 to this sequence; this gives

$$\lim_{n\to\infty} \mathbb{E}[\inf_{k>n} (X_k) \mid \mathcal{G}] = \mathbb{E}[\liminf_{n\to\infty} X_n \mid \mathcal{G}]$$

almost surely. Now note that $\mathbb{E}[\inf_{k>n} X_k \mid \mathcal{G}] \leq \inf_{k>n} \mathbb{E}[X_k \mid \mathcal{G}]$ almost surely. This is because if we fix any $k > n$, then $\inf_{k>n} X_k \leq X_k$, so the same is true if we take conditional expectations.

And now we take limits as $n \to \infty$; this implies that

$$\mathbb{E}[\liminf_{n\to\infty} X_n \mid \mathcal{G}] \le \lim_{n\to\infty} \inf_{k>n} \mathbb{E}[X_k \mid \mathcal{G}] = \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}]$$

(almost surely). And that's exactly what we wanted to prove. □

So this has two inputs — the first is the conditional monotone convergence theorem, and the second is that conditional expectations preserve ordering (if $X \le Y$, then $\mathbb{E}[X \mid \mathcal{G}] \le \mathbb{E}[Y \mid \mathcal{G}]$), which gives $\mathbb{E}[\inf X_k \mid \mathcal{G}] \le \inf \mathbb{E}[X_k \mid \mathcal{G}]$. (To see why again, fix any $k > n$. Then $\inf_{k>n} X_k \le X_k$ almost surely, so $\mathbb{E}[\inf_{k>n} X_k \mid \mathcal{G}] \le \mathbb{E}[X_k \mid \mathcal{G}]$. But this is true for every k, so if you take the infimum over k you will still have this inequality.)

Proof of conditional DCT. In the unconditional case, we proved the DCT by creating some nonnegative random variables and applying Fatou's lemma. That's exactly what we're going to do now, with conditional Fatou's lemma.

Consider the random variables $Y + X_n$ and $Y - X_n$. These are both nonnegative random variables (by the hypothesis $|X_n| \le Y$). So we have sequences of nonnegative random variables, which lets us apply conditional Fatou's; this gives that

$$\mathbb{E}[Y + X \mid \mathcal{G}] = \mathbb{E}[\liminf_{n\to\infty}(Y + X_n) \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathbb{E}[Y + X_n \mid \mathcal{G}].$$

And by linearity, we have

$$\mathbb{E}[Y + X \mid \mathcal{G}] = \liminf_{n\to\infty} \mathbb{E}[X_n + Y \mid \mathcal{G}] = \mathbb{E}[Y \mid \mathcal{G}] + \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}].$$

If we do the same for $Y - X$, we get

$$\mathbb{E}[Y - X \mid \mathcal{G}] = \mathbb{E}[\liminf_{n\to\infty}(Y - X_n) \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathbb{E}[Y - X_n \mid \mathcal{G}] = \mathbb{E}[Y \mid \mathcal{G}] - \limsup_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}]$$

(here we're using that $\liminf(-x_n) = -\limsup x_n$).

Now we just combine these — by linearity, we have $\mathbb{E}[Y + X \mid \mathcal{G}] = \mathbb{E}[Y \mid \mathcal{G}] + \mathbb{E}[X \mid \mathcal{G}]$, which means the $\mathbb{E}[Y \mid \mathcal{G}]$ term cancels, and we get

$$\mathbb{E}[X \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}].$$

Similarly, for the second equality we can write $\mathbb{E}[Y - X \mid \mathcal{G}] = \mathbb{E}[Y \mid \mathcal{G}] - \mathbb{E}[X \mid \mathcal{G}]$, and again the Y cancels and we get

$$\mathbb{E}[X \mid \mathcal{G}] \ge \limsup_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}].$$

Combining these, we get that

$$\mathbb{E}[X \mid \mathcal{G}] \le \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}] \le \limsup_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}] \le \mathbb{E}[X \mid \mathcal{G}]$$

(since we always have $\liminf \le \limsup$ by definition). This means all the terms in this chain have to be equal; in particular, since the $\liminf$ and $\limsup$ are equal, $\lim_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}]$ exists and equals $\mathbb{E}[X \mid \mathcal{G}]$. (This is the same thing as in our proof of unconditional dominated convergence theorem, just using the conditional instead of unconditional Fatou's lemma.) □

Finally, it remains to prove conditional Jensen. For this, we're going to invoke something which is nontrivial — given any convex function, you can express it as a supremum of countably many affine (linear) functions. (This is nontrivial but not extremely hard to see.)

Fact 16.6 — For any convex function φ, we can write $\varphi(x) = \sup_{i \in I}(a_i x + b_i)$ for all $x \in \mathbb{R}$ (for some set I).

We'll accept this, and use this nice representation to prove our claim.

Proof of conditional Jensen. Write φ in the above form. This means for all $i \in I$, we have

$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \leq \mathbb{E}[a_i X + b_i \mid \mathcal{G}] = a_i \mathbb{E}[X \mid \mathcal{G}] + b_i$$

(since conditional expectation satisfies linearity).

This is true for all i; now taking the supremum over i, we have

$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \geq \sup_{i \in I}(a_i \mathbb{E}[X \mid \mathcal{G}] + b_i).$$

But the right-hand side is just $\varphi(\mathbb{E}[X \mid \mathcal{G}])$ almost surely (by the same fact). □

This is another way to see the unconditional case of Jensen's inequality.

And that's the end of the proof — the final statement $\|\mathbb{E}[X \mid \mathcal{G}]\|_p \leq \|X\|_p$ is just the special case with $\varphi(x) = x^p$, which is convex for $p \geq 1$.

Student Question. *Where did we use that φ is in $\mathcal{L}^1$ or nonnegative?*

Answer. We need this so that we can make sense of $\mathbb{E}[\varphi(X) \mid \mathcal{G}]$. If φ is nonnegative, then you can always make sense of this, even if $\varphi(X)$ is not in $\mathcal{L}^1$ (and if it's in $\mathcal{L}^1$, then this makes sense by the construction of conditional expectation from last class).

So these are the analogs of the limit theorems we learned in the unconditional case. Apart from conditional Jensen, the proofs are the same as before.

Student Question. *What's the intuition behind conditioning on a σ-algebra — why are we doing this?*

Answer. We'll see this later when talking about martingales — you can imagine you perform an experiment in different steps, $\mathcal{F}$ is the possible space of all the outcomes of the experiment, and $\mathcal{G}$ represents all the information you've gained after the first n steps. You can imagine X as the next result, and you want to keep track of how useful the information you have so far is; that corresponds exactly to the conditional expectation. So this captures, given that we know the first n results, how much you know about the next result. In general, σ-algebras are information — they're a collection of events.

§16.2 More properties of conditional expectation

We'll now prove a few more properties of conditional expectation.

Proposition 16.7 (Tower property)
Let $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ be σ-algebras, and let $X \in \mathcal{L}^1(\mathcal{F})$. Then we have

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$$

almost surely.

In other words, if I condition on a smaller σ-algebra, it doesn't matter whether I start with X itself, or X conditioned on a bigger σ-algebra. In other words, this means if we know all the information encoded by $\mathcal{H}$, then this gives us the same information for the entire random variable X and for the conditional random variable $\mathbb{E}[X \mid \mathcal{G}]$. In other words, this says that if $\mathcal{H} \subseteq \mathcal{G}$, then you can just forget the conditioning on $\mathcal{G}$.

Proof. The proof is just by definition — clearly, we have that $\mathbb{E}[X \mid \mathcal{H}]$ is $\mathcal{H}$-measurable (by the definition of the conditional expectation), and for all $A \in \mathcal{H}$, we have

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot \mathbf{1}_A] = \mathbb{E}[X \cdot \mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot \mathbf{1}_A]$$

(the first equality is because any event in $\mathcal{H}$ is also in $\mathcal{G}$). And so by the definition of the conditional expectation, this means $\mathbb{E}[X \mid \mathcal{H}]$ satisfies the properties that the conditional expectation $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]$ is supposed to have. So this implies the left-hand side is the same as the right-hand side (by uniqueness). □

The next property, roughly speaking, says that we can always take out what is known.

Proposition 16.8

Let $X \in \mathcal{L}^1(\mathcal{F})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. If Y is bounded and $\mathcal{G}$-measurable, then

$$\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]$$

almost surely.

We need Y to be bounded in order to say that XY is in $\mathcal{L}^1$ (that's what we use it for). What this says is that we can always take out what is known — if you know $\mathcal{G}$ then you know Y, because Y is $\mathcal{G}$-measurable; and what this means is that you can always take it out of the expectation.

Proof. Again, we'll just prove that the random variable $Y \cdot \mathbb{E}[X \mid \mathcal{G}]$ satisfies the properies of the conditional expectation that $\mathbb{E}[XY \mid \mathcal{G}]$ is supposed to satisfy; then they have to be the same by uniqueness.

Suppose we fix any $A \in \mathcal{G}$. Then we have

$$\mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}] \cdot \mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot (\mathbf{1}_A \cdot Y)].$$

Now, if Y were an indicator function, then the expectations of these two random variables would be the same. In general, Y might not be an indicator function. But that's fine, because we can always approximate it by a linear combination of indicator functions. So we still do have that this is equal to $\mathbb{E}[X \cdot (\mathbf{1}_A Y)]$. (To repeat, this would be true by the definition of $\mathbb{E}[X \mid \mathcal{G}]$ if Y were an indicator function of an element of $\mathcal{G}$. But it's still true for any $Y \in \mathcal{L}^1$ — this property is satisfied when Y is an indicator function, and it's also satisfied by linear combinations; so by taking limits, it's also satisfied for all $Y \in \mathcal{L}^1(\mathcal{G})$.)

In particular, this is equal to $\mathbb{E}[(XY)\mathbf{1}_A]$, so we get

$$\mathbb{E}[(Y \cdot \mathbb{E}[X \mid \mathcal{G}])\mathbf{1}_A] = \mathbb{E}[(XY)\mathbf{1}_A].$$

By the uniqueness of conditional expectation, this means

$$Y \cdot \mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[XY \mid \mathcal{G}]$$

almost surely, which completes the proof. □

(When we approximate Y by indicator functions, we're implicitly using e.g. the dominating convergence theorem.)

We'll see one more property, and then we'll study conditional expectations of Gaussian random variables (in this special case, we can actually identify the conditional expectations).

Proposition 16.9

Let $X \in \mathcal{L}^1(\mathcal{F})$, and fix two σ-algebras $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$. If $\sigma(X, \mathcal{G})$ is independent of $\mathcal{H}$, then

$$\mathbb{E}[X \mid \sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \mid \mathcal{G}]$$

(almost surely).

Here we're conditioning on more information — a larger σ-algebra (the one generated by $\mathcal{G}$ and $\mathcal{H}$). But if the σ-algebra generated by X and $\mathcal{G}$ is independent of $\mathcal{H}$, then we can actually just forget $\mathcal{H}$. The intuition is that $\mathcal{H}$ doesn't give us any further information about X.

Remark 16.10. If we weaken the assumption of independence and just say that $\sigma(X)$ is independent of $\mathcal{H}$ and that $\mathcal{G}$ is independent of $\mathcal{H}$, then this is not true anymore; it is a nice exercise to construct a counterexample. Here the assumption is stronger — that $\sigma(X, \mathcal{G})$ is independent of $\mathcal{H}$ (which is not equivalent to just saying each of X and $\mathcal{G}$ individually is independent of $\mathcal{H}$ — the first implies the second, but not the other way).

Proof. Here, it suffices to prove this in the case where X is nonnegative, since for general $X \in \mathcal{L}^1$ we can always write X as a difference of nonnegative functions.

Suppose we fix some $A \in \mathcal{G}$ and $B \in \mathcal{H}$. Then we have that

$$\mathbb{E}[\mathbb{E}[X \mid \sigma(\mathcal{G}, \mathcal{H})] \cdot \mathbf{1}_{A \cap B}] = \mathbb{E}[X \cdot \mathbf{1}_{A \cap B}]$$

(by the definition of conditional expectation). But we also have $\mathbf{1}_{A \cap B} = \mathbf{1}_A \mathbf{1}_B$, so this is equal to $\mathbb{E}[X \mathbf{1}_A \mathbf{1}_B]$. And the random variable $X\mathbf{1}_A$ is measurable with respect to $\sigma(X, \mathcal{G})$, while $\mathbf{1}_B$ is measurable with respect to $\mathcal{H}$. And since $\mathcal{H}$ is independent of $\sigma(X, \mathcal{G})$, we can split this up as $\mathbb{E}[X\mathbf{1}_A] \cdot \mathbb{P}[B]$. Further, we have $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot \mathbf{1}_A]\mathbb{P}[B]$. And again, we can put B inside, because $\mathbb{E}[X \mid \mathcal{G}]\mathbf{1}_A$ is also measurable with respect to $\sigma(X, \mathcal{G})$. So we can also write this as

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\mathbf{1}_{A \cap B}].$$

So for all A and B, we have that

$$\mathbb{E}[\mathbb{E}[X \mid \sigma(\mathcal{H}, \mathcal{G})]\mathbf{1}_{A \cap B}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\mathbf{1}_{A \cap B}].$$

So our expectations agree on sets of the form $A \cap B$. And the idea from now is that $\sigma(\mathcal{H}, \mathcal{G})$ is generated by sets of this form $A \cap B$; so this equality has to hold for all $C \in \sigma(\mathcal{G}, \mathcal{H})$. But to make this more precise, we need to construct some measures and say two measures agree on a π-system generating the σ-algebra, so they agree everywhere. We'll do this precisely, since this is a bit delicate.

Set $Y = \mathbb{E}[X \mid \sigma(\mathcal{H}, \mathcal{H})]$, so that $Y \geq 0$ almost surely (since we assumed X is). Then we define the measures $\mu(F) = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot \mathbf{1}_F]$ and $\nu(F) = \mathbb{E}[Y\mathbf{1}_F]$ (for all $F \in \sigma(\mathcal{G}, \mathcal{H})$). If we know that these two measures are the same, then we're done — it'll mean this equality is true for all $F \in \sigma(\mathcal{G}, \mathcal{H})$, so by uniqueness we'd have to have this equality.

For this, the point is we've already shown μ and ν agree on the π-system $\{A \cap B \mid A \in \mathcal{G}, B \in \mathcal{H}\}$. And this π-system generates $\sigma(\mathcal{G}, \mathcal{H})$. Also, these two measures agree on the entire space — we have $\mu(\Omega) = \mathbb{E}[X]$, and similarly $\nu(\Omega) = \mathbb{E}[X]$. So we have two finite measures which agree on a generating π-system, which means they agree everywhere.

So we've shown that $\mu = \nu$ on $\sigma(\mathcal{H}, \mathcal{G})$, which implies by uniqueness that $\mathbb{E}[X \mid \sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \mid \mathcal{G}]$ almost surely. □

As we can see in these proofs, when we want to prove certain conditional expectations are the same, we often invoke uniqueness.

Student Question. *Why did we need nonnegativity here?*

Answer. The point is just so that μ and ν are actually measures (we need them to be nonnegative for this). For the general case, we can just write $X = \max\{X, 0\} - \max\{-X, 0\}$, and just apply the claim to these two functions.

§16.3 Conditional expectations of Gaussians

In general, it's not easy to compute conditional expectations — usually it's not easy to get an explicit form for $\mathbb{E}[X \mid \mathcal{G}]$, unless the space is nice (e.g., if it can be partitioned into countably many measurable sets). However, there are some cases where we can actually have an explicit form, and one is that of Gaussians — if you have a Gaussian vector (X, Y), then there is an explicit form for $\mathbb{E}[X \mid Y]$. This is one way in which Gaussians are very nice.

Let (X, Y) be a Gaussian vector on $\mathbb{R}^2$, and set $\mathcal{G} = \sigma(Y)$. What we want to do is compute $X' = \mathbb{E}[X \mid \mathcal{G}]$. If X and Y are independent, then knowing the information in $\mathcal{G}$ gives us nothing (they're independent, so this doesn't add any information); so in that case we'd have $X' = \mathbb{E}[X]$. In general they might not be independent, but it turns out we can still get an explicit form.

The first thing to note is the following (which is nontrivial but not extremely hard to see) — there exists a Borel $f : \mathbb{R} \to \mathbb{R}$ such that $X' = f(Y)$. So if you have a random variable which is measurable with respect to the σ-algebra generated by another, then this random variable has to be some (Borel) function of that other. (This is always true, not just for Gaussians — it'd be true for any random variables X and Y.) The idea now is to actually compute f. And the way we'll do this is by guessing — we'll guess what the right form of X' should be.

Recall Gaussians satisfy some nice linearity properties — linear combinations of Gaussians are still Gaussians. So we'd expect that X' is still a Gaussian. This means it's natural to try taking f to be linear; then what should its value be?

So as a guess, we'll try $X' = aY + b$ for some $a, b \in \mathbb{R}$ (because Gaussian random variables respect linearity). Suppose that f has this form. What should the values of a and b be? If X' has this form, then we would have

$$\mathbb{E}[X'] = \mathbb{E}[X] = a\mathbb{E}[Y] + b.$$

This means a and b would be determined in terms of each other, but we need another equality to be able to determine both. Another thing we'd have is that

$$\mathbb{E}[(X - X')Y] = 0,$$

by the definition of X'. This implies that $\operatorname{Cov}[X - X', Y] = 0$, which implies that

$$\operatorname{Cov}[X, Y] = \operatorname{Cov}[X', Y] = a \operatorname{Var}(Y).$$

So we would expect these two things to be true. In particular, if we set

$$a = \frac{\operatorname{Cov}[X, Y]}{\operatorname{Var}(Y)} \quad \text{and} \quad b = \mathbb{E}[X] - \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(Y)} \cdot \mathbb{E}[Y]$$

(so a and b depend only on X and Y), then we would have $\operatorname{Cov}(X - X', Y) = 0$. And since $(X - X', Y)$ is a Gaussian (if we assume X' is a Gaussian and has this form), then we get that $X - X'$ and Y are *independent* (for this choice of X').

Now what we do is we've guessed, if X' had this form, what a and b would be. And for this choice of a and b, we just set X' to that random variable. Then we have that $X - X'$ and Y are independent (because their covariance is 0). This implies that if Z is $\sigma(Y)$-measurable, then we also have $\mathbb{E}[(X - X')Z] = 0$ —

in other words, this means $\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[X'\mathbf{1}_A]$ for all $A \in \sigma(Y)$. So this means that X' indeed satisfies the properties of the conditional expectation.

So the conditional expectation has to be $X' = aY + b$. This means it has an explicit form (since a and b only depend on X and Y).

Again, the idea is that we try to guess f — we ask, what if f were a linear function? Then we figure out what the values of a and b would be, to end up with a particular choice — if the conditional expectation had a linear form, then a and b would have to be these values. Then we set X' to be the corresponding random variable, and we prove that X' does satisfy the necessary properties; so it has to actually be the conditional expectation.

> **Student Question.** *How do we know $\mathbb{E}[(X - X')Z] = 0$ for $\sigma(Y)$-measurable?*
>
> **Answer.** Because $X - X'$ and Y are independent, so $X - X'$ has to be independent of any function of Y.

The main idea is we set a and b such that $\mathrm{Cov}(X - X', Y) = 0$, which ensures $X - X'$ and Y are independent; and that means $\mathbb{E}[(X - X')Z] = 0$ for all Z measurable with respect to Y. And taking $Z = \mathbf{1}_A$ for $A \in \sigma(Y)$, we get that indeed X' satisfies the properties that $\mathbb{E}[X \mid \sigma(Y)]$ is supposed to.

§17 November 5, 2024

Last lecture, we finished by explicitly computing the conditional expectation of a Gaussian random variable given another Gaussian, provided that the vector (X, Y) is also Gaussian — it might be the case that X and Y are both one-dimensional Gaussians but (X, Y) is not, in which case we can't do this. But if (X, Y) is indeed Gaussian, then we can compute $\mathbb{E}[X \mid Y]$, and we did this in the previous lecture.

§17.1 Conditional density functions

Suppose X and Y are random variables in $\mathbb{R}$ with joint density function $f_{(X,Y)}(x, y)$ (for $(x, y) \in \mathbb{R}^2$). In general, a random variable doesn't necessarily have a density function; but here we're assuming the vector (X, Y) does. Given this, we're going to compute the conditional law of X given Y, if we know this density function.

More generally, let $h\colon \mathbb{R} \to \mathbb{R}$ be a Borel function (i.e., a Borel-measurable function) such that $h(X)$ is integrable. Our goal is to compute $\mathbb{E}[h(X) \mid Y]$. This is slightly more general than what we did last class — we computed this conditional expectation where (X, Y) is Gaussian, in which case the density function is known.

Recall that if we know (X, Y) has a probability density function, then we also know Y has a density function, and this is given by integrating the joint density with respect to X — i.e.,

$$f_Y(y) = \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dx.$$

In order to find the conditional expectation, we need to compute the expectation of $h(X)$ restricted to sets which are measurable with respect to Y. For that, suppose we fix $g\colon \mathbb{R} \to \mathbb{R}$ to be a bounded measurable function. Then we have

$$\mathbb{E}[h(X) \cdot g(Y)] = \int_{\mathbb{R}^2} h(x)g(y)f_{(X,Y)}(x, y)\, dx\, dy$$

(this is just by the definition of the joint density function). And the point is that by applying Fubini's theorem, we can rewrite this in a convenient way, by integrating with respect to x first; then we get

$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x) \cdot \frac{f_{(X,Y)}(x, y)}{f_Y(y)}\, dx \right) g(y) f_Y(y)\, dy.$$

Here what we've done is just divided and multiplied by $f_Y(y)$, and applied Fubini's theorem. And now we have our candidate for the conditional expectation — it's the thing inside the integral (at least, whenever it makes sense, which means the denominator is positive). So we consider the function

$$\phi(y) = \begin{cases} \int_{\mathbb{R}} h(x) \cdot \frac{f_{(X,Y)}(x,y)}{f_Y(y)}\,dx & \text{if } f_Y(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

With this definition of ϕ, we have

$$\mathbb{E}[h(X)g(Y)] = \int_{\mathbb{R}} \phi(y) \cdot g(y) f_Y(y)\,dy.$$

But this integral on the right, by the definition of the density function of Y, is just $\mathbb{E}[\phi(Y)g(y)]$. (And this is true for every bounded measurable function g.)

This implies that

$$\mathbb{E}[h(X)\mathbf{1}_A] = \mathbb{E}[\phi(Y)\mathbf{1}_A]$$

for all $A \in \sigma(Y)$. And by the uniqueness of the conditional expectation, this means

$$\mathbb{E}[h(X) \mid Y] = \phi(Y)$$

almost surely, where ϕ is the explicit function we defined.

What this tells us is that if you know the joint density function, then all information about conditional expectations is encoded in it — in that case we can compute conditional expectations in this way. Of course, $f_{(X,Y)}(x,y)$ does not necessarily exist, and in that case we don't necessarily have a closed form for the conditional expectation. But this lets us find conditional expectations for a large number of pairs of random variables. And what we did with Gaussians is a special case of this, because for Gaussians we do know the joint density (and if we plugged that in, we'd get the same thing as last lecture).

Student Question. *How do we know the set of points with $f_Y(y) = 0$ doesn't affect our integral — if f vanishes on a set that's not of measure 0, why do we still have $\mathbb{E}[h(x)g(y)] = \mathbb{E}[\phi(Y)g(Y)]$?*

Answer. By the definition of f_Y, if $f_Y(y) = 0$, then $\int f_{(X,Y)}(x,y)\,dx = 0$. So if we define

$$A = \{y \in \mathbb{R} \mid f_Y(y) = 0\},$$

then we have

$$\int_{\mathbb{R}\times A} f_{(X,Y)}(x,y)\,dx\,dy = 0.$$

This means we can neglect this set in our integral.

§17.2 Stochastic processes and filtrations

This completes the chapter on conditional expectations; now we'll move on to the next chapter, which is discrete-time martingales. This is the most important part for the remainder of the course.

Here the setup is as follows. As usual, we always have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. And we also have a measurable space $(E, \mathcal{E})$ where our random variables take values.

We're going to be interested in a sequence of random variables $X = (X_n)_{n\geq 0}$, defined in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (we say that they're *coupled* in the same probability space), which take values in E.

Definition 17.1. We call such a sequence $X = (X_n)$ a stochastic process on E.

So that's the setup — we have a bunch of random variables defined in the same probability space. And given that we have such a sequence, we'll be interested in *filtrations*.

Definition 17.2. A filtration $(\mathcal{F}_n)_{n\geq 0}$ is an increasing family of sub-σ-algebras of $\mathcal{F}$ (meaning that $\mathcal{F}_n \subseteq \mathcal{F}$ for all n, and $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all n).

Intuitively, we can interpret a filtration as information that we keep track of. Suppose that we perform an experiment at different times, and X_n is the result of the experiment during the nth time. (For instance, we can flip a coin, or observe the velocity of a particle at time n.) Intuitively, $\mathcal{F}_n$ corresponds to the information of the first n results of the experiment. That's the best way to understand intuitively what a filtration is — it's very abstract at first glance, but really it just corresponds to keeping track of information.

Of course, someone might ask whether such a filtration exists. It does exist. In fact, a stochastic process always induces a filtration.

Definition 17.3. For a stochastic process (X_n), we define its natural filtration, denoted $(\mathcal{F}_n^X)_{n\geq 0}$, by

$$\mathcal{F}_n^X = \sigma(X_k \mid 1 \leq k \leq n).$$

This corresponds to the above intuition (it's what the filtration should be if we think of it as tracking the information from the first n experiments).

As another natural question:

Question 17.4. If we start with a filtration and a stochastic process, how are they related?

We saw a stochastic process induces a filtration, but the converse is not necessarily true. But somehow these two concepts might be related.

Definition 17.5. We say a stochastic process X is adapted to $(\mathcal{F}_n)_{n\geq 0}$ if X_n is $\mathcal{F}_n$-measurable for all n.

In this case, clearly we have $\mathcal{F}_n^X \subseteq \mathcal{F}_n$ for all n (because by the definition of a filtration, if X is adapted with respect to $\mathcal{F}_n$, then $X_1, \ldots, X_n$ are all $\mathcal{F}_n$-measurable).

The point of martingale theory is to analyze such stochastic processes, and their relationships with such filtrations. They have applications, but are also important mathematically.

§17.3 Martingales, supermartingales, and submartingales

Now it's time to define martingales and some related objects. We'll assume the same setup as above, with $E = \mathbb{R}$ with the Borel σ-algebra.

Definition 17.6. Suppose that $X = (X_n)_{n\geq 0}$ is adapted with respect to some filtration $(\mathcal{F}_n)_{n\geq 0}$, such that $X_n \in \mathbb{R}$. Then:

(1) X is a martingale if $\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m$ for all $m \leq n$.

(2) X is a supermartingale if $\mathbb{E}[X_n \mid \mathcal{F}_m] \leq X_m$ almost surely for all $m \leq n$.

(3) X is a submartingale if $\mathbb{E}[X_n \mid \mathcal{F}_m] \geq X_m$ almost surely for all $m \leq n$.

This is the main concept of the chapter. We're going to prove lots of properties and theorems related to martingales, submartingales, and supermartingales.

As mentioned earlier, $\mathbb{E}[X_n \mid \mathcal{F}_m]$ can be interpreted as the expected profit someone can have, given that we know the results of the first m rounds. In particular, you can imagine that in every single round, there is some kind of a game, where you either lose or gain with a certain rule. And X_m will represent the total profit that you have made up until the mth step. Here, $\mathbb{E}[X_n \mid \mathcal{F}_m]$ represents the amount of money you expect to have accumulated after the next $n-m$ rounds, given that you know what has happened up until the mth round. If we have equality, as in a martingale, then the game is considered to be fair. If you have a supermartingale, then the game is not fair — you expect your total profit to drop from your profit at time m. If you have a submartingale, then the game is in your favor — the amount that you expect to gain after $n-m$ rounds is at least the amount of money that you have at the mth round. So this is another intuitive way to somehow understand these definitions — they come from real-life scenarios. We're studying this in an abstract way, but we'll see some examples to understand more intuitively what these concepts represent.

Fact 17.7 — Every process which is a martingale with respect to a given filtration $(\mathcal{F}_n)$ is also a martingale with respect to its natural filtration $(\mathcal{F}_n^X)$.

This follows by the tower property from the previous lecture (and the fact that $\mathcal{F}_n^X \subseteq \mathcal{F}_n$).

(The same holds for supermartingales and submartingales.)

Remark 17.8. It doesn't matter whether we start our indexing at 0 or 1. In experiments, you usually have a starting condition, and then the system evolves; that's why we usually start at 0. But it's just a matter of preference.

§17.3.1 Some examples of martingales

Now that we've introduced a new concept, we should check that it actually exists, by seeing some examples.

Example 17.9 (One-dimensional simple random walk)
Let $(\xi_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with $\mathbb{E}[\xi_i] = 0$. Then if we take $X_n = \sum_{i=1}^n \xi_i$, it's easy to check that X_n is a martingale.

(This is with respect to the natural filtration — we usually assume this when the filtration is not mentioned.)

Example 17.10
Let $(\xi_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence with $\mathbb{E}[\xi_i] = 1$. Then $X_n = \prod_{i=1}^n \xi_i$ is a martingale (again, with respect to the natural filtration).

Here we need expectation 1 because we're taking products.

So we now have two families of examples of martingales; throughout the course, we're going to see more examples.

§17.4 Stopping time

When you have a martingale, it's interesting to ask:

Question 17.11. What happens to the martingale if we stop it at a random time?

So N is a random variable taking values in $\mathbb{N}$. And we want to see how the martingale behaves when we take X_N, where N is random. This leads to the notion of stopping times.

Definition 17.12. A stopping time T is a random variable $T: \Omega \to \mathbb{N} \cup \{\infty\}$ such that $\{T \leq n\} \in \mathcal{F}_n$ for all n.

In other words, to decide whether the stopping time is at most n, it's enough to observe the first n terms of the experiment. So a stopping time is a random variable taking integer values (possibly ∞) such that if we want to know whether it's at most n, it suffices to observe just the first n terms of our sequence — we don't have to observe subsequent terms.

Equivalently, we have the following (this is an easy exercise).

Fact 17.13 — A random variable T is a stopping time if and only if $\{T = n\} \in \mathcal{F}_n$ for all n.

To see this, we can write

$$\{T_A = n\} = \{T_A \leq n\} \setminus \{T_A \leq n-1\}.$$

The first term on the right belongs to $\mathcal{F}_n$, and the second to $\mathcal{F}_{n-1} \subseteq \mathcal{F}_n$; so their difference is in $\mathcal{F}_n$.

The point of the chapter is to prove properties for martingales, submartingales, supermartingales, and stopping times — random variables of this form.

§17.4.1 Examples

Again, since we introduced a new concept, let's give some examples of such stopping times.

The first example is the trivial one.

Example 17.14

Any constant time is trivially a stopping time.

But we can also have more complicated stopping times.

Example 17.15

Fix any Borel $A \in \mathcal{B}(\mathbb{R})$. Then if we define

$$T_A = \inf\{n \mid X_n \in A\}$$

as the first time our sequence enters A, then T_A is a stopping time.

This might take infinite values — we use the convention that $\inf \emptyset = \infty$, so if the sequence never enters A, then T_A will take value ∞.

To see that T_A is a stopping time, we can just apply the definition — we can write

$$\{T_A \leq n\} = \bigcup_{k=1}^{n} \{X_k \in A\}.$$

But X_k is $\mathcal{F}_k$-measurable, so $\{X_k \in A\} \in \mathcal{F}_k$. And since we have a filtration, this means $\{X_k \in A\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$. So $\{T_A \leq n\} \in \mathcal{F}_n$.

Student Question. *What do you mean by 'observing'?*

Answer. This means if we want to decide whether the stopping time is at most n, this is $\mathcal{F}_n$-measurable. And intuitively, $\mathcal{F}_n$ can be understood as all the information you've gathered in the first n steps. So to decide whether the stopping time has occured by time n, we just need to have observed the first n

random variables (in the case $\mathcal{F}_n = \mathcal{F}_n^X$).

§17.5 Some properties of stopping times

We'll now mention some properties of stopping times. We won't prove them — they're an easy exercise — but it's nice to have them written down.

Proposition 17.16

Let S, T, and $(T_n)_{n\geq 0}$ be stopping times, on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$. Then $\min\{S,T\}$, $\max\{S,T\}$, $\inf T_n$, $\sup T_n$, $\liminf_{n\to\infty} T_n$, and $\limsup_{n\to\infty} T_n$ are all stopping times.

(By a *filtered probability space*, we just mean a probability space endowed with some filtration $(\mathcal{F}_n)$.)

This is analogous to how when we discussed random variables, we said that if we had two measurable random variables, then their minimum and maximum are measurable; and if we have a sequence, then their infimum, supremum, lim inf, and lim sup are all measurable. Next chapter we'll prove similar things for continuous-time martingales, but there you need additional conditions, and things are more complicated.

§17.6 Stopping times and stopped processes

We have a filtration and stopping time, and we can wonder if we can define a σ-algebra representing the information we've gained up to time T. What $\mathcal{F}_n$ means is that we perform an experiment for a deterministic number of times n, and record all the information we gained up to this time. But what if we run it for a random number of times? This leads to the following definition.

Definition 17.17. Let T be a stopping time on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$. Then we define

$$\mathcal{F}_T = \{A \in \mathcal{F} \mid A \cap \{T \leq n\} \in \mathcal{F}_n \text{ for all } n\}.$$

Intuitively, we can interpret $\mathcal{F}_T$ as all the information available up to time T. In other words, it's the collection of all possible outcomes such that if we restrict to the event that our stopping time is at most n, then A only depends on the first n terms of the experiment.

Student Question. *When we have a filtration, do we require $\bigcup_n \mathcal{F}_n = \mathcal{F}$?*

Answer. No, not necessarily — it just has to be an increasing collection of σ-algebras.

So this can be intuitively interpreted as the information up to time T — if we restrict ourselves to the event that the stopping time T has occurred by time n, then this outcome should only depend on the first n terms of the experiment.

We can also define a *stopped process*, as mentioned earlier.

Definition 17.18. We define the random variable X_T as

$$X_T(\omega) = X_{T(\omega)}(\omega)$$

for all $\omega \in \Omega$, whenever $T(\omega) < \infty$. We also define the stopped process X^T as the sequence $X_t^T = X_{T\wedge t}$ for all $t \geq 0$.

This means you take the random time T, and observe all outcomes of the experiment up to time T and then just stop it at that time.

(We write $T \wedge t = \min\{T, t\}$.)

§17.6.1 Some properties

Proposition 17.19

Let S and T be stopping times, and suppose we have an adapted process $X = (X_n)_{n\geq 0}$.

(i) If $S \leq T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

(ii) The random variable $X_T \cdot \mathbf{1}_{T<\infty}$ is $\mathcal{F}_T$-measurable.

(iii) The stopped process X^T is adapted.

(iv) If X is integrable (meaning that $X_n \in L^1$ for all n), then X^T is also integrable.

What (1) says is that if $S \leq T$, then all the information we get up to time S is contained in the information we get up to time T. (This is obvious from the definitions.)

For (2), we restrict to the event $T < \infty$ because for now it doesn't make sense to define X_T when $T = \infty$. Then this makes sense, because intuitively $\mathcal{F}_T$ is all the information that you get up to time T.

The proofs of these properties are not that complicated; they're almost immediate from the definitions. As mentioned earlier, (1) is straightforward (it immediately follows from the definition of $\mathcal{F}_T$).

Proof of (2). For this, we need to show measurability. So suppose that we fix some $A \in \mathcal{B}(\mathbb{R})$. Then we want to consider the event $\{X_T \mathbf{1}_{T<\infty} \in A\}$. We need to show

$$\{X_T \mathbf{1}_{T<\infty} \in A\} \cap \{T \leq n\} \in \mathcal{F}_n$$

for all n; then by the definition of $\mathcal{F}_T$, this means $\{X_T \mathbf{1}_{T<\infty} \in A\}$ is in $\mathcal{F}_T$, which means our random variable is measurable.

For this, we can write the above event as

$$\bigcup_{m=1}^{n} \{X_m \in A\} \cap \{T = m\}.$$

We know T is an integer random variable, so we just take the union over all its possible values.

But we know T is a stopping time, and that X is adapted. This means $\{T = m\} \in \mathcal{F}_m$ (for example, since $\{T = m\} = \{T \leq m\} \setminus \{T \leq m-1\}$, and these lie in $\mathcal{F}_m$ and $\mathcal{F}_{m-1}$). And we also know $\{X_m \in A\} \in \mathcal{F}_m$, because X is an adaptive process. So their intersection is also in $\mathcal{F}_m$, and therefore in $\mathcal{F}_n$ (because $(\mathcal{F}_n)$ is increasing). Therefore, the entire union also belongs to $\mathcal{F}_n$.

So the above event $\{X_T \mathbf{1}_{T<\infty} \in A\} \cap \{T \leq n\}$ belongs to $\mathcal{F}_n$ for all n.

So since this is true for all n, this implies $\{X_T \mathbf{1}_{T<\infty} \in A\}$ is $\mathcal{F}_T$-measurable for all $A \in \mathcal{B}(\mathbb{R})$. And by the definition of measurability, that means $X_T \mathbf{1}_{T<\infty}$ is indeed $\mathcal{F}_T$-measurable. □

Then (3) follows immediately from (2); let's see why this is true.

Proof of (3). For every n, the random variable $X_{T\wedge n}$ (X stopped at the minimum of T and n) can also be written as

$$X_{T\wedge n} \cdot \mathbf{1}_{T\wedge n<\infty}$$

(since $T \wedge n = \min\{T, n\}$ is always finite). And by (2), we've shown that this is $\mathcal{F}_{T\wedge n}$-measurable — here we're using the fact that $T \wedge n$ is a stopping time.

And we also have $\mathcal{F}_{T\wedge n} \subseteq \mathcal{F}_n$. So this means $X_{T\wedge n}$ is $\mathcal{F}_n$-measurable for all n, which shows (3), that the stopped process X^T is adapted (with respect to the same filtration — by the definition of an adapted process). □

Proof of (4). For (4), we want to show the stopped process is also an integrable process. For this, we have

$$\mathbb{E}[|X_{T\wedge n}|] = \mathbb{E}\left[\sum_{m=0}^{\infty} |X_{m\wedge n}\mathbf{1}_{T=m}|\right]$$

(we're just summing over all possible values of T). And this can be written as

$$\mathbb{E}\left[\sum_{m=0}^{n-1} |X_m|\,\mathbf{1}_{T=m}\right] + \mathbb{E}\left[\sum_{m\geq n} |X_n|\,\mathbf{1}_{T=m}\right].$$

(In the second term, this is because if $T \geq n$ then we always have $T \wedge n = n$; so we have X_n instead of n.) But the first term here is at most $\sum_{m=0}^{n-1}\mathbb{E}[|X_m|]$. And the secod term is at most $\mathbb{E}[|X_n|]$ (because it's the expectation of $|X_n|$ restricted to the event $T \geq n$, which is at most $\mathbb{E}[|X_n|]$). So in total, we get that this is $\sum_{m=0}^{n}\mathbb{E}[|X_m|]$, which is finite (since the sequence is integrable).

So this is finite, which means the stopped process is indeed integrable. □

§18 November 7, 2024

Last time we started our chapter on discrete-time martingales; this chapter contains the most important results of the course, together with the limit theorems. It'll also be useful if we take stochastic calculus later.

In the previous lecture, we gave rigorous definitions of martingales, filtrations, and all the important concepts we'll use. We also introduced stopping times, and gave a bit of intuition behind these definitions.

At the end, we proved some simple properties of stopping times and stopped martingale processes. This lecture, we'll start by stating and proving the optional stopping theorem, which is quite useful.

§18.1 The optional stopping theorem

Theorem 18.1 (Optional stopping)

Let $X = (X_n)_{n\geq 0}$ be a martingale. Then:

(1) If T is a stopping time, then the stopped process X^T is also a martingale. In particular,

$$\mathbb{E}[X_{T\wedge t}] = \mathbb{E}[X_0].$$

(2) If $S \leq T$ are bounded stopping times, then $\mathbb{E}[X_T \mid \mathcal{F}_S] = X_S$ almost surely.

(3) If $S \leq T$ are bounded stopping times, then $\mathbb{E}[X_T] = \mathbb{E}[X_S]$.

(4) If there exists an integrable random variable Y which dominates all the X_n (i.e., $|X_n| \leq Y$ for all n) and T is a stopping time which is finite almost surely (but not necessarily bounded), then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

(5) If X has bounded increments (i.e., there exists $M \in (0,\infty)$ such that $|X_{n+1} - X_n| \leq M$ almost surely for all n) and T is a stopping time such that $\mathbb{E}[T] < \infty$, then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

The second part of (1) follows immediately from the fact that X^T is a martingale.

For (2), recall that $\mathcal{F}_S$ intuitively is all the information that's encoded by our martingale up to the stopping time S (we defined it rigorously last lecture).

The point of the optional stopping theorem is that you have a martingale X and stop it at some random time T, and you want to compute the expectation of that random observable. The optional stopping theorem gives you conditions under which this is computable, and not just computable but actually equal to $\mathbb{E}[X_0]$. Condition (3) is the first such condition (where we take $S = 0$). Condition (4) is another sufficient condition to have this equality, which is more general. We don't assume anymore that T is bounded, only that it's finite almost surely — so boundedness has been replaced with a random variable Y that dominates everything (we're going to apply the dominated convergence theorem).

Note that $\mathbb{E}[T] < \infty$ is a stronger assumption than T being finite almost surely (as we have in (4)).

What someone should take from the OST is that it gives you conditions under which you have $\mathbb{E}[X_T] = \mathbb{E}[X_0]$. We have (4) and (5). In (4), we're assuming something more general on T, but we have to pay a price — that we have a uniform integrable bound on the X_n. In (5) we don't assume that, but we assume the increments are bounded; and we make the stronger assumption that $\mathbb{E}[T]$ is finite, which is stronger. (We'll see counterexamples without these conditions.)

We'll prove these one by one.

Proof of (1). Here we need to show the stopped process X^T is a martingale. By the definition of a martingale, we want to show that

$$\mathbb{E}[X_{T\wedge t} \mid \mathcal{F}_{t-1}] = X_{T\wedge(t-1)} \text{ almost surely}$$

for all $t \in \mathbb{N}$. (This is the definition of the martingale property.) Recall that last lecture, we showed that X^T is integrable, so these expectations makes sense.

Now we expand the left-hand side — we have

$$\mathbb{E}[X_{T\wedge t} \mid \mathcal{F}_{t-1}] = \mathbb{E}\left[\sum_{s=0}^{t-1} X_s \cdot \mathbf{1}_{\{T=s\}} \mid \mathcal{F}_{t-1}\right] + \mathbb{E}[X_T \cdot \mathbf{1}_{\{T>t-1\}} \mid \mathcal{F}_{t-1}].$$

What's the first term? The first term is simply given by

$$X_T \cdot \mathbf{1}_{\{T\le t-1\}},$$

because the event $\{T \le t-1\}$ is measurable with respect to $\mathcal{F}_{t-1}$ (because T is a stopping time), and all the random variables X_s we get in this case are measurable with respect to $\mathcal{F}_s \subseteq \mathcal{F}_{t-1}$.

For the second term, $\mathbf{1}_{T>t-1}$ is measurable with respect to $\mathcal{F}_{t-1}$, so we can move it outside the expectation; and X_t is a martingale, so its expectation given $\mathcal{F}_{t-1}$ is just X_{t-1}. So the second term is $X_{t-1}\mathbf{1}_{T>t-1}$. This gives

$$\mathbb{E}[X_{T\wedge t} \mid \mathcal{F}_{t-1}] = X_T\mathbf{1}_{T\le t-1} + X_{t-1}\mathbf{1}_{T>t-1} = X_{T\wedge(t-1)}.$$

And that's exactly what we wanted to prove. □

Student Question. *Why are we only proving this for $t-1$, and not previous indices?*

Answer. By the tower property, if this is true for $\mathcal{F}_{t-1}$, then it'll also be true for $\mathcal{F}_s$ for any $s \le t-1$, by the tower property.

We'll also prove (2) and (3); we'll leave (4) and (5) as exercises so that we can get familiar on our own with these techniques (we already have all the ingredients we need).

Proof of (2). Since S and T are bounded, let $n \in \mathbb{N}$ be some fixed integer such that $T \le n$ almost surely. Then we can write

$$X_T = (X_T - X_{T-1}) + (X_{T-1} - X_{T-2}) + \cdots + (X_{S+1} - X_S) + X_S,$$

because $S \leq T$. (So we're writing X_T as a sum of increments.) And we can rewrite this as

$$X_T = X_S + \sum_{k=0}^{n} (X_{k+1} - X_k) \mathbf{1}_{S \leq k < T}.$$

We need to prove (2), so we just need to show that X_S satisfies the properties of the conditional expectation $\mathbb{E}[X_T \mid \mathcal{F}_S]$, and then invoke uniqueness of the conditional expectation (as we've done before).

For this, fix $A \in \mathcal{F}_S$. Then we have

$$\mathbb{E}[X_T \cdot \mathbf{1}_A] = \mathbb{E}[X_S \cdot \mathbf{1}_A] + \sum_{k=0}^{n} \mathbb{E}[(X_{k+1} - X_k) \cdot \mathbf{1}_{S \leq k < T} \cdot \mathbf{1}_A],$$

using the above decomposition and linearity of expectation. To complete the proof, it suffices to show that each term inside the sum is 0, because then we'll have $\mathbb{E}[X_T \cdot \mathbf{1}_A] = \mathbb{E}[X_S \cdot \mathbf{1}_A]$, and therefore that X_S satisfies the properties of $\mathbb{E}[X_T \mid \mathcal{F}_S]$ (and by uniqueness, they have to be the same almost surely).

To show each term is 0, we use the martingale property and the definition of $\mathcal{F}_S$. We can rewrite this expectation as

$$\mathbb{E}[(X_{k+1} - X_k) \mathbf{1}_{\{S \leq k < T\} \cap A}] = \mathbb{E}[X_{k+1} \mathbf{1}_{A \cap \{S \leq k < T\}}] - \mathbb{E}[X_k \mathbf{1}_{A \cap \{S \leq k < T\}}].$$

And the point is that $A \cap \{S \leq k < T\}$ is in $\mathcal{F}_k$. Why? This is because S and T are stopping times, so $\{S \leq k < T\} = \{S \leq k\} \cap \{T > k\}$. And both of these events are in $\mathcal{F}_k$ (because S and T are stopping times). Then this means, by the definition of $\mathcal{F}_S$ — for $A \in \mathcal{F}_S$ means that if we take A and intersect it with any event in $\mathcal{F}_k$, then it'll still be in $\mathcal{F}_k$. So that means $A \cap \{S \leq k < T\}$ is in $\mathcal{F}_k$.

But we know X is a martingale, so we have

$$\mathbb{E}[X_{k+1} \mid \mathcal{F}_k] = X_k;$$

that implies the two terms are equal, so their difference is 0.

This means $\mathbb{E}[X_T \mid \mathcal{F}_S] = X_S$ almost surely. □

Then (3) is immediate, by taking expectations in (2) — because the expectation of the conditional expectation is equal to the expectation in the original σ-algebra. Explicitly, we have

$$\mathbb{E}[X_T] = \mathbb{E}[\mathbb{E}[X_T \mid \mathcal{F}_S]] = \mathbb{E}[X_S]$$

(by properties of conditional expectations, and the fact that $\mathbb{E}[X_T \mid \mathcal{F}_S] = X_S$ almost surely).

And (4) and (5) are a nice exercise, which will be left to us to do. The reasoning is the same. In (4), you use Y to use the dominated convergence theorem. (5) is something similar with increments, but you have to be slightly careful; but you do something similar as in the proof of (2) (where we write out differences).

This is a very useful theorem, and we'll use it a lot of times in the course (and in stochastic calculus).

As a remark, you can have a similar version of the optional stopping theorem when you start with a supermartingale or submartingale; but here you will have some inequalities (depending on whether you have a submartingale or supermartingale) — for example, you'll have $\mathbb{E}[X_T] \geq \mathbb{E}[X_0]$ or $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$. But the proofs are the same.

§18.1.1 Some counterexamples

Now we'll see that the conditions in (4) and (5) are quite subtle in general.

Example 18.2 (One-dimensional simple random walk)

Let $(\xi_k)_{k\geq 0}$ be a sequence of i.i.d. random variables such that $\mathbb{P}[\xi_k = -1] = \mathbb{P}[\xi_k = +1] = \frac{1}{2}$. Set $X_n = \xi_0 + \cdots + \xi_n$. Define the stopping time T as the first time the walk hits the value 1, i.e.,

$$T = \inf\{n \geq 0 \mid X_n = 1\}.$$

Then T is finite almost surely (this is not obvious, but can be proven in infinitely many ways). But

$$\mathbb{E}[X_T] = 1 \neq \mathbb{E}[X_0] = 0.$$

(Clearly $\mathbb{E}[X_T] = 1$, by the definition of T; and $\mathbb{E}[X_0] = 0$ by symmetry.)

What's going on is that T doesn't have finite expectation. We do have bounded increments — the increments are bounded by 1 — but we don't have $\mathbb{E}[T] < \infty$. So there are counterexamples — it's not easy to give very general conditions.

Student Question. *Why couldn't we apply (2) in this example, setting $S = 0$?*

Answer. They're not bounded — T is not bounded, and in fact has infinite expectation.

So this shows that when applying the optional stopping theorem, one has to be quite careful.

§18.1.2 A version for nonnegative supermartingales

Proposition 18.3

Suppose that X is a nonnegative supermartingale. Then for any stopping time T which is finite almost surely, we have $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$.

The proof is just Fatou's lemma. So if we have a nonnegative supermartingale and stop it at *any* time, then $\mathbb{E}[X_T]$ will always be finite (since $\mathbb{E}[X_0]$ is). We don't need any assumptions on T; the only assumption is that we're working with a nonnegative supermartingale and a stopping time which is finite almost surely.

Proof. We have $\mathbb{E}[X_T] = \mathbb{E}[\lim_{n\to\infty} X_{T\wedge n}]$. Now we apply Fatou's lemma to say that this is at most

$$\liminf_{n\to\infty} \mathbb{E}[X_{T\wedge n}].$$

But the latter term is at most $\mathbb{E}[X_0]$ by the optional stopping theorem (the version of (1) for supermartingales). □

Student Question. *Where are we using that T is finite almost surely?*

Answer. So that X_T is well-defined — otherwise X_T doesn't make sense, at least for the moment.

(Note that this result is only for supermartingales, not submartingales.)

§18.2 Random walks

It's remarkable how many things you can compute about random walks — they're remarkable mathematical objects, and you can say lots of things about them. We're going to compute certain probabilities, using the optional stopping theorem.

Suppose we have the same setup as before — we have some i.i.d. sequence $(\xi_i)_{i\geq 1}$ such that each takes the values ± 1 with probabilities $\frac{1}{2}$, and we consider the simple random walk $X_n = \sum_{i=1}^n \xi_i$ (where $X_0 = 0$).

For every integer $c \in \mathbb{Z}$, we can consider the first hitting time of c by the random walk, defined as

$$T_c = \inf\{n \geq 0 \mid X_n = c\}.$$

Here, the point is to calculate certain probabilities.

Question 18.4. What is $\mathbb{P}[T_{-a} < T_b]$, the probability that the random walk hits $-a$ before b?

(We assume $a, b > 0$.)

Here we can visualize this by drawing time on the x-axis, and drawing the random walk; it might go up and then down and then back up. And we have two levels (drawn as horizontal lines) — one level b, and one level $-a$. And we stop the walk at the first time it hits either the lower level or the upper level; we want to compute the probability it hits the lower level before the upper level. And we'll do this using the optional stopping theorem.

Here we'll use the stopping time $T = T_{-a} \wedge T_b$ (this is the first time we hit either $-a$ or b). There are certain conditions on the optional stopping theorem; the easiest candidate here is (5), because the X_n have bounded increments (in fact, the absolute value of the increments is 1, because each time we go either up or down by 1). To apply the OST, we'll show that $\mathbb{E}[T] < \infty$. Individually, each of T_{-a} and T_b doesn't have a finite expectation; but if we take their minimum, then it does. That's what we're going to show.

Claim 18.5 — We have $\mathbb{E}[T] < \infty$.

Proof. For this, we'll do a trick. If you take the random walk and run it for $a + b$ steps, and those $a + b$ steps are all $+1$, then clearly the random walk will exceed the bounds. So in particular, T is bounded above by the first time that there are $a + b$ consecutive $+1$'s in the walk.

So we want the first time such that $\xi_i = 1$ for $a + b$ consecutive steps; if this happens, then clearly the random walk is going to exceed the region. So T is at most this random variable, which we'll call S — so S is the first time that you have $a + b$ consecutive $+1$'s.

To be mathematically precise, we define

$$S = \inf\{n \mid \xi_i = 1 \text{ for all } n - (a+b) + 1 \leq i \leq n\}.$$

And we have $T \leq S$.

Now, what's the probability that the first $a + b$ terms are $+1$? This is computable — each ξ_i is 1 with probability $\frac{1}{2}$, so the probability the first $a + b$ are (by independence) is $2^{-(a+b)}$. In other words,

$$\mathbb{P}[\xi_i = 1 \text{ for all } 1 \leq i \leq a+b] = \prod_i \mathbb{P}[\xi_i = 1] = \frac{1}{2^{a+b}}.$$

Then we perform the same experiment for the next block of $a + b$ consecutive numbers. If this first event happens, then we're done — the stopping time T has already occurred before $a + b$. If this is not the case — there is some $1 \leq i \leq a + b$ with $\xi_i = 1$ — then we go to the next $a + b$ elements. We have

$$\mathbb{P}[\xi_i = 1 \text{ for all } a+b+1 \leq i \leq 2(a+b)] = \frac{1}{2^{a+b}}$$

for the same reason; and if this event happens, then T is at most $2(a+b)$. And we iterate — we keep doing exactly the same thing.

If we perform the same experiment n times, then for the nth block we again have

$$\mathbb{P}[\xi_i = 1 \text{ for all } (n-1)(a+b)+1 \le i \le n(a+b)] = \frac{1}{2^{a+b}}.$$

And we consider the first number n for which this happens. If we let this event be A_n, then we have

$$S \le (a+b) \cdot \inf\{n \in \mathbb{N} \mid A_n \text{ occurs}\}.$$

(So we're taking the first $a+b$ consecutive integers, and seeing if all their ξ_i's are 1. If they are, we stop; this means T has already occurred. If this doesn't happen, we move on to the next $a+b$, and so on. We keep doing this until the first n for which $a+b$ consecutive integers all have the value 1.)

Student Question. *Why are we using blocks of $a+b$ — if you begin with $a-1$, why don't we use the block starting with 2?*

Answer. We want the events to be independent, so that we can bound the expectation.

And we have $\mathbb{E}[T] \le \mathbb{E}[S]$, which is at most the expectation on the right.

And the thing in the infimum is a geometric random variable, so we know its expectation — if we call that random variable $I = \inf\{n \in \mathbb{N} \mid A_n \text{ occurs}\}$, then we have $\mathbb{E}[I] = 2^{a+b}$ (if you have a geometric random variable with success parameter p, the expectation should be $1/p$). In particular, this is finite. So

$$\mathbb{E}[T] \le \mathbb{E}[S] \le (a+b)\mathbb{E}[I] = (a+b)2^{a+b},$$

which is a finite number. □

So now T has finite expectation and the increments of the random walk are bounded, so we can apply the optional stopping theorem — by (5) of the optional stopping theorem, we have

$$\mathbb{E}[X_T] = \mathbb{E}[X_0] = 0$$

(because we start the random walk at 0). But X_T can only take two values, $-a$ and b. This means

$$0 = \mathbb{E}[X_T] = -a\mathbb{P}[T_{-a} < T_b] + b\mathbb{P}[T_{-a} > T_b].$$

And the two probabilities sum to 1, so we get

$$0 = -a\mathbb{P}[T_{-a} < T_b] + b - b\mathbb{P}[T_{-a} < T_b] = (-b-a)\mathbb{P}[T_{-a} < T_b] + b.$$

Now we rearrange, and we get

$$\boxed{\mathbb{P}[T_{-a} < T_b] = \frac{b}{a+b}}.$$

So we've explicitly computed this probability using the optional stopping theorem. Here the hard work was proving that $\mathbb{E}[T] < \infty$, so that we could apply the OST; then we just did some manipulations. There are other ways to compute this probability as well, but this is an elegant way that avoids a lot of computations and combinatorial things.

In general, we should try to understand the optional stopping theorem as well as possible — it's a very important result, and there's a huge range of problems one can generate using it (and it's useful in stochastic calculus, if we plan to study that).

Of course, random walks also make sense in higher dimensions, but things are more complicated.

§18.3 The martingale convergence theorem

Now we'll move to the next topic, which is the martingale convergence theorem. We'll state it and a quick consequence, and we'll try to prove it in the next lecture.

Theorem 18.6 (Almost sure martingale convergence theorem)
Let $X = (X_n)_{n\geq 0}$ be a supermartingale which is bounded in L^1 (meaning that $\sup_n \mathbb{E}[|X_n|] < \infty$). Then there is a random variable $X_\infty \in L^1(\mathcal{F}_\infty)$ such that $X_n \to X_\infty$ converges almost surely (where $\mathcal{F}_\infty = \sigma(\mathcal{F}_n \mid n \geq 0)$).

So this theorem says that if we have a sequence of supermartingales which is bounded in L^1, then you have convergence almost surely to *some* limit. We'll also see other kinds of convergence in the next lectures.

As a quick comment, this in particular means L^1-bounded martingales converge almost surely, because a martingale is a supermartingale.

Usually when we try to prove convergence in general, we do it by guessing what the limit should be, and after we've guessed the limit, we prove that it actually converges. That trick doesn't work here — here it's not easy to construct the limit explicitly. So we'll instead follow an indirect approach — a very beautiful trick that counts the number of up-crossings of every fixed interval (we'll explain what this means later).

Before we do this, we'll write down an easy corollary of this theorem.

Corollary 18.7
Let $X = (X_n)_{n\geq 0}$ be a nonnegative supermartingale. Then X_n converges almost surely as $n \to \infty$, to an almost surely finite limit.

Proof. To prove this (given the theorem), it suffices to show that X_n is bounded in L^1. That follows from the supermartingale property: since X_n is nonnegative, we have

$$\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leq \mathbb{E}[X_0] < \infty$$

(and this bound is uniform for all n). So (X_n) is bounded in L^1, and by the almost sure martingale convergence theorem, the X_n will converge almost surely to some limit (which is in L^1, so therefore finite almost surely). □

Student Question. *Are we assuming X_0 is integrable?*

Answer. When we defined martingales, we required the random variables to be integrable.

§18.3.1 Up-crossings

As mentioned earlier, the main goal is to prove the almost sure martingale convergence theorem. But it's not easy to explicitly construct the limit, and that's why we're going to employ a very nice trick, which counts the number of *up-crossings* of a sequence. In particular, we have the following definition.

Imagine we have two (positive) levels a and b (with $a < b$), and we want to count how many times the sequence X_n makes a crossing — how many times you start with some $X_n \leq a$, and end up with another term $X_m \geq b$ (with some intermediate terms in between). So we're interested in how many times the sequence crosses this strip.

The point is that a sequence converges if and only if for all finite a and b, the number of times the sequence crosses this strip (up and down) is finite. This is a fact from real analysis that we'll prove.

We'll consider our sequence (X_n), and we're going to bound the number of times it crosses this strip up and down, for all a and b. We'll prove that this expectation is finite, and from that we'll conclude almost sure convergence in an indirect way.

Now we'll use mathematical symbols to explain what we mean by this picture. For this, fix a sequence $(x_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$, and two real numbers $a < b$. We inductively define the following times:

- We define $T_0 = 0$, and $S_1 = \inf\{n \geq 1 \mid x_n \leq a\}$ as the first time the sequence goes below a. So S_1 is the first time our sequence goes below the level a.
- We define $T_2 = \inf\{n \geq S_1 \mid x_n \geq b\}$ as the first time after S_1 that the sequence goes above level b. Whenever this happens, we say the sequence makes a *crossing* of the strip.
- Inductively, we define
$$S_{k+1} = \inf\{n \geq T_k \mid x_n \leq a\}$$
to be the first time after T_k such that the sequence goes below level a, and
$$T_{k+1} = \inf\{n \geq S_{k+1} \mid x_n \geq b\}$$
as the first time after S_{k+1} that the sequence goes above b.

To more rigorously define the *number of crossings*, for each fixed n, we define

$$N_n([a,b];x) = \sup\{k \geq 0 \mid T_k \leq n\}.$$

(This is the number of crossings of the strip up to time n.) This is an increasing sequence in n, so it'll converge increasingly as $n \to \infty$ to some $N([a,b];x)$; and this number N is the *total* number of crossings of the strip (the number of times the sequence goes from the lower to the upper level).

We're mentioning this for the following lemma.

Lemma 18.8

A sequence x_n converges in $\mathbb{R} \cup \{\pm\infty\}$ if and only if $N([a,b];x)$ is finite for all $a < b$ in $\mathbb{Q}$.

That's the key point behind Doob's up-crossing proof of the almost sure martingale convergence theorem. So if for every fixed strip the number of times the sequence crosses from bottom to top is finite, then the sequence converges.

Student Question. *Is the statement equivalent if you allow $a, b \in \mathbb{R}$?*

Answer. Yes, by the density of $\mathbb{Q}$. If you fix $a, b \in \mathbb{R}$, then you can fix rational numbers a bit less than a and a bit larger than b; then the number of crossings between a and b is at most the number of crossings between these two rationals. So it suffices to just deal with rational numbers.

So the strategy to prove the almost sure convergence theorem is to show that for all fixed a and b, the expected total number of crossings is finite; then we can apply the lemma (we can take the intersection over all pairs of rationals, since the intersection of countably many events with probability 1 still has probability 1). And the bound we'll find is the Doob's up-crossing inequality.

§19 November 12, 2024

§19.1 The martingale convergence theorem

Last lecture, we stated the almost sure martingale convergence theorem — if we have a supermartingale bounded in L^1, then it converges almost surely. We said the main component of the proof was bounding

the number of up-crossings this sequence makes in a fixed interval. Imagine we draw the horizontal lines at a and b. Then we consider the number of times we have a picture where we go from $X_{S_k} \le a$ to $X_{T_k} \ge b$ (where T_k is the first time after S_k that the sequence goes above b). We said a sequence converges if and only if for every fixed interval, the number of up-crossings is finite. We want to prove an almost sure convergence theorem, so it suffices to show that for any such a and b, the number of up-crossings is finite almost surely. That's the content of Doob's upcrossing inequality, which we'll prove today.

§19.1.1 Doob's upcrossing inequality

Theorem 19.1 (Doob's upcrossing inequality)

Let X be a supermartingale and fix any $a < b$. Then we have

$$(b-a)\mathbb{E}[N_n(X,a,b)] \le \mathbb{E}[(X_n - a)^-].$$

(Here $\mathbb{E}[(X_n - a)^-]$ is the expectation of the negative part of $X_n - a$, i.e., $\min\{X_n - a, 0\}$.)

Recall that $N_n(X,a,b)$ is the number of up-crossings up to time n.

In particular, because X is a supermartingale, $\mathbb{E}[(X_n - a)^-]$ is finite, and we can show it's in fact bounded above by a constant that doesn't depend on n. So if we let $n \to \infty$, we get that the expected total number of up-crossings of the interval made by X is finite (since $\mathbb{E}[N_n(X,a,b)]$ will converge to the expected total number of up-crossings); and that means the number is finite almost surely (since its expectation is finite). And since this is true for all a and b, this completes the proof.

Unfortunately, this proof doesn't give an explicit construction of the limit, but it proves it exists.

Student Question. *How do you know $N_n(X,a,b)$ is measurable?*

Answer. This is because the S_k and T_k are stopping times, and we can express $N_n(X,a,b)$ in terms of these stopping times.

The proof is short, but it's the most important input.

Proof. Let $N = N_n(X,a,b)$ for convenience. By the definition of S_k and T_k, we have

$$X_{T_k} - X_{S_k} \ge b - a.$$

We also have that

$$\sum_{k=1}^{n}(X_{T_k \wedge n} - X_{S_k \wedge n}) = \sum_{k=1}^{N}(X_{T_k} - X_{S_k}) + (X_n - X_{S_{N+1}})\mathbf{1}_{S_{N+1} \le n}.$$

(this is just by the definition of N as the total number of up-crossings up to time n). This is because if $k \ge N+1$, then $S_k \wedge n = T_k \wedge n = n$, so we don't have any contributions from the second term.

Now we need an upper bound, so we need to somehow apply the optional stopping theorem for these times. In order to apply the optional stopping theorem, we need to say T_k and S_k are stopping times. This follows relatively easily by induction. When $k = 1$, this is easy (by their definitions) — T_1 is the first time the martingale enters a certain Borel set, and we've seen this is a stopping time. And then for general k, we just use induction — we assume T_k is a stopping time, and given that we prove that T_{k+1} is.

So T_k and S_k are all stopping times, by induction on k. This implies that

$$\mathbb{E}[X_{S_k \wedge n}] \ge \mathbb{E}[X_{T_k \wedge n}],$$

by the optional stopping theorem (because the minimum between two stopping times is always a stopping time, and since $S_k \wedge n \le T_k \wedge n$, we can apply the optional stopping theorem (the version for supermartingales)). In particular, this means

$$0 \ge \mathbb{E}\left[\sum_{k=1}^{n} X_{T_k \wedge n} - X_{S_k \wedge n}\right],$$

because each of the terms has nonpositive expectation. Note also that each difference is always at least $b-a$, as seen above (by the definitions of T_k and S_k). So we have

$$0 \ge \mathbb{E}\left[\sum_{k=1}^{n}(X_{T_k \wedge n} - X_{S_k \wedge n})\right] = \mathbb{E}\left[\sum_{k=1}^{N}(X_{T_k} - X_{S_k})\right] + \mathbb{E}\left[(X_n - X_{S_{N+1}})\mathbf{1}_{S_{N+1} \le n}\right].$$

The first term is at least $(b-a)\mathbb{E}[N]$. And the second is at least $-\mathbb{E}[(X_n - a)^-]$ — this is because

$$(X_n - X_{S_{N+1}})\mathbf{1}_{S_{N+1} \le n} \ge -(X_n - a)^-.$$

And if we rearrange, we get

$$(b-a)\mathbb{E}[N] \le \mathbb{E}[(X_n - a)^-]. \qquad \square$$

§19.1.2 Proof of the martingale convergence theorem

Now we're almost done, but not quite, because we need to be a bit careful. Fix $a, b \in \mathbb{Q}$ with $a < b$. Then by Doob's upcrossing inequality, we have that

$$\mathbb{E}[N_n(n, [a,b])] \le \frac{\mathbb{E}[(X_n - a)^-]}{b-a}.$$

But $(X_n - a)^- \le \mathbb{E}[|X_n|] + |a|$ by the triangle inequality, so

$$\mathbb{E}[N_n(X, [a,b])] \le (b-a)^{-1}(\mathbb{E}[|X_n|] + |a|).$$

Now we want to take $n \to \infty$, and get a uniform bound in n for the left-hand side.

First, we have that

$$N_n(X, [a,b]) \uparrow N(X, [a,b])$$

as $n \to \infty$ (the total number of up-crossings up to time n converges to the total number of up-crossings), by definition. So we can apply the monotone convergence theorem (because we have nonnegative increasing sequences) to say that

$$\mathbb{E}[N(X, [a,b])] \le (b-a)^{-1}(|a| + \sup_n \mathbb{E}[|X_n|]).$$

And this is finite, because we assumed that X is L^1-bounded.

So for any $a < b$, the otal number of up-crossings of this interval has finite expectation, which means it is finite almost surely. Now one might wonder why we just consider rationals; this is because eventually we'll take the intersection of these events over all a and b, and we need this to be countable to say that the probability of the intersection is still 1 (a countable intersection of probability-1 events is also probability 1). So we get that

$$\mathbb{P}\left[\bigcap_{a<b\in\mathbb{Q}} \{N_n(X, [a,b]) < \infty\}\right] = 1.$$

And we'll show that under this event, we have convergence. Let $\Omega_0 = \bigcap_{a<b\in\mathbb{Q}}\{N(X, [a,b]) < \infty\}$. Recall that this means for any real $a < b$, the total number of up-crossings of $[a,b]$ is finite (since for any a and b, you can find rationals arbitrarily close to a and b, so you can bound the number of up-crossings by one

where the interval is finite). So on Ω_0, X_n converges (i.e., the pointwise limit $\lim_{n\to\infty} X_n$ exists). Then we can just set

$$X_\infty(\omega) = \begin{cases} \lim X_n & \omega \in \Omega_0 \\ 0 & \text{otherwise} \end{cases}$$

(it doesn't matter how we define it outside Ω_0, because Ω_0 is a probability-1 event).

It's clear that X_∞ is a random variable, because pointwise limits of measurable functions are measurable functions (we showed this, on the second problem set).

So the remaining ting is to show that $X_\infty \in L^1$. But this follows by Fatou's lemma — we have

$$\mathbb{E}[|X_\infty|] = \mathbb{E}\left[\liminf |X_n|\right] \le \liminf \mathbb{E}[|X_n|].$$

But this $\liminf$ is finite, because we assumed X is bounded in L^1 (meaning there is a uniform bound on $|X_n|$ for all n). This implies $X_\infty \in L^1$, which was the last component of the proof.

Remark 19.2. In particular, if X is a martingale, then it is also a supermartingale; and therefore a L^1-bounded martingale converges almost surely (which is why we call it the martingale convergence theorem).

This of course constructs the limit in an abstract way; we don't have an explicit form for it from this approach.

Student Question. *How do you go from $\mathbb{E}[X_n - X_{S_{N+1}}]$ to $\mathbb{E}[X_n - a]$?*

Answer. This is because $X_{S_{N+1}} \le a$ by the definition of S_{N+1}. What it means for $S_{N+1} \le n$ is that X_n hasn't gone below a yet, so $X_n > a$.

§19.2 Doob's inequalities

Now we've completed the proof of the martingale convergence theorem, and the next topic is to mention Doob's inequalities, which will help us prove other notions of convergence (specifically, L^p convergence).

§19.2.1 Doob's maximal inequality

The first is Doob's maximal inequality, which bounds the probability that the martingale is very big.

Theorem 19.3 (Doob's maximal inequality)
Let $X = (X_n)$ be a nonnegative submartingale. If we set $X_n^* = \sup_{0\le k\le n} X_k$, then we have

$$\lambda \mathbb{P}[X_n^* \ge \lambda] \le \mathbb{E}[X_n \mathbf{1}_{X_n^* \ge \lambda}] \le \mathbb{E}[X_n].$$

So what this says is that if we have a nonnegative submartingale, then we have an upper bound that the maximum of the submartingale, over the first n units of time, is large. We'll see this is another consequence of the optional stopping theorem.

Proof. To apply the optional stopping theorem, we need to construct a stopping time. Here we have an obvious choice — we can just consider the first time the process goes above λ. So we let

$$T = \inf\{k \ge 0 \mid X_k \ge \lambda\}.$$

Then $T \wedge n$ is a bounded stopping time, so the optional stopping theorem implies that

$$\mathbb{E}[X_n] \geq \mathbb{E}[X_{T\wedge n}]$$

(here we have a submartingale, so the optional stopping theorem works in the reverse direction compared to before). And we can write

$$\mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_T \mathbf{1}_{T\leq n}] + \mathbb{E}[X_n \mathbf{1}_{T>n}]$$

(by the linearity of the integral). Now remember that if $T \leq n$, then $X_T \geq \lambda$ by the definition of T; this means the first term on the right-hand side is at least $\lambda\mathbb{P}[T \leq n]$. And now we can rearrange to get

$$\lambda \cdot \mathbb{P}[T \leq n] \leq \mathbb{E}[X_n] - \mathbb{E}[\mathbf{1}_{T>n}] = \mathbb{E}[X_n \mathbf{1}_{T\leq n}].$$

But $T \leq n$ if and only if there is some $0 \leq k \leq n$ with $X_k \geq \lambda$, which is the same as saying $X_n^* \geq \lambda$. Then we can just replace these two events (which are the same) to get

$$\lambda\mathbb{P}[X_n^* \geq \lambda] \leq \mathbb{E}[X_n \mathbf{1}_{X_n^*\geq\lambda}],$$

and since X_n is nonnegative, this is of course at most $\mathbb{E}[X_n]$. □

This will be useful because it'll let us upgrade the convergence in the martingale convergence theorem. Essentially what this says is it bounds the probability that the maximum value of the submartingale during the first n units of time is big. It's a Markov-type inequality, with the difference that in Markov's inequality, instead of X_n in $\mathbb{E}[X_n \mathbf{1}_{X_n^*\geq\lambda}]$, we would have X_n^*. Here we have X_n instead of X_n^*, so we have a stronger inequality.

We should consider this as an analog of Markov, in the context of submartingales; but the difference is that it's stronger.

§19.2.2 Doob's L^p inequality

One might wonder if we have similar bounds for the L^p norms. The answer is yes — we also have L^p type inequalities, also due to Doob. We'll prove these next, and it'll help us prove L^p convergence for martingales.

Theorem 19.4 (Doob's L^p inequality)

Let X be a martingale or nonnegative submartingale. Then for every $p > 1$, letting $X_n^* = \sup_{0\leq k\leq n} |X_n|$, we have

$$\|X_n^*\|_p \leq \frac{p}{p-1} \cdot \|X_n\|_p .$$

This means in other words that the pth norms of X_n and X_n^* are comparable, up to constants that depend only on p. (Clearly $\|X_n^*\|_p \geq \|X_n\|$ by the definition of X_n^*; but the point is that we also get a reverse inequality with a constant only depending on p.)

Proof. If we assume X is a martingale (the theorem states it can be either a martingale or nonnegative submartingale), then by Jensen's inequality, we have that $|X|$ is a nonnegative submartingale. (Here we're applying Jensen's inequality for conditional expectations on the absolute value function.) This theorem is about $|X|$, so if X is a martingale, we can directly apply the nonnegative submartingale case to $|X|$. In other words, for this theorem, it suffices to consider the case where X is a nonnegative submartingale.

The idea is that we're going to use Doob's maximal inequality to bound from above the probability that $|X|$ is very large during the first n units of time. Suppose that we fix some large $k < \infty$. Then we have

$$\mathbb{E}[(X_n^* \wedge k)^p] = \mathbb{E}\left[\int_0^k p x^{p-1} \mathbf{1}_{X_n^*\geq x}\, dx\right]$$

(this was also on the problem set, or something of a similar flavor). Now we can apply Fubini's theorem to bring the expectation inside, which gives

$$\int_0^k px^{p-1}\mathbb{P}[X_n^* \geq x]\,dx.$$

Now we can apply Doob's maximal inequality to bound this probability — this is at most

$$\int_0^k px^{p-2}\mathbb{E}[X_n \mathbf{1}_{X_n^* \geq x}]\,dx.$$

And for this, we can take the expectation outside again, so this is just

$$\mathbb{E}\left[X_n \int_0^k px^{p-2}\mathbf{1}_{X_n^* \geq x}\,dx\right] = \mathbb{E}[X_n \cdot (X_n^* \wedge k)^{p-1}] \cdot \frac{p}{p-1}.$$

And now we just apply Hölder's inequality to say that this is at most

$$\frac{p}{p-1} \cdot \|X_n\|_p \left\|X_n^* \wedge k\right\|_p^{p-1}.$$

And this is true for all k. Now if we rearrange, we obtain that

$$\left\|X_n^* \wedge k\right\|_p \leq \frac{p}{p-1}\|X_n\|_p.$$

And this is true for all k, so we can let $k \to \infty$, and this gives that $\|X_n^*\|_p \leq \frac{p}{p-1}\|X_n\|_p$. (Here we also used the monotone convergence theorem, since $X_n \wedge k$ is an increasing nonnegative sequence that converges almost surely to X_n.) □

So this is an L^p version of Doob's maximal inequality.

§19.3 L^p martingale convergence theorem

We've now seen two nice inequalities; one bounds the probability that the maximum of a nonnegative submartingale (during the first unit of times) is too big, and the second bounds the pth norm of that maximum.

The reason we mentioned these two is that Doob's L^p inequality is the main input to the L^p martingale convergence theorem — an analog of the martingale convergence theorem from before, but where now the convergence is in L^p.

Theorem 19.5

Let X be a martingale, and fix $p > 1$. Then the following statements are equivalent:

(1) X is bounded in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$, i.e., $\sup_{n\geq 0}\|X_n\|_p < \infty$.

(2) X_n converges almost surely and in $\mathcal{L}^p$ to a random variable X_∞.

(3) There exists a random variable $Z \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_n = \mathbb{E}[Z \mid \mathcal{F}_n]$ almost surely.

What this says roughly is that a martingale converges in L^p if and only if it's L^p-bounded. Of course, it's not obvious at all why this is the case, but we're going to prove it by combining everything mentioned so far. And not only that, but it also has a nice expression — it can be expressed as this conditional expectation, for some random variable Z. This characterizes the L^p convergence of martingales.

Proof (1) $\implies$ (2). We'll start by assuming (1) and proving (2). The almost sure limit follows from the almost sure martingale convergence theorem. Note that an L^p-bounded sequence is also L^1-bounded, by Jensen's inequality — Jensen's inequality gives that $\|X_n\|_1 \le \|X_n\|_p$ for all n, which means $\sup_n \|X_n\|_1 \le \sup_n \|X_n\|_p$, which is finite by our hypothesis in (1). So this implies X is L^1-bounded, and therefore X_n converges almost surely to some X_∞ as $n \to \infty$ by the martingale convergence theorem — so the martingale convergence theorem guarantees a limit actually exists.

Now we need to upgrade this to L^p convergence. For that, we're going to use Doob's L^p inequality.

First, in order to ask for L^p convergence, we need to show that $X_\infty \in L^p$. This follows by Fatou's lemma — if we apply Fatou's lemma, we have

$$\mathbb{E}[|X_\infty|^p] = \mathbb{E}[\liminf_{n\to\infty} |X_n|^p] \le \liminf_{n\to\infty} \mathbb{E}[|X_n|^p] \le \sup_n \|X_n\|^p,$$

which we assumed is finite. So X_∞ is indeed an L^p random variable.

Now we know the almost sure convergence, and we somehow need to apply the dominated convergence theorem to obtain L^p convergence. For this, we need to apply Doob's L^p inequality — because we know a bound on the *maximum* of the sequence up to time n. And somehow we need to say that if we take the supremum over *all* n, this is in L^p; for that we're going to use Doob's L^p inequality.

Doob's L^p inequality gives that $\|X_n^*\|_p \le \frac{p}{p-1} \|X_n\|_p$. And we have a uniform bound on the right-hand side — if we call $\sup_{n\ge 0} \|X_n\|_p = I$, then we get $\|X_n^*\|_p \le \frac{p}{p-1} I$. This is a uniform bound, and if we take $n \to \infty$, then the left-hand side converges to $\|X_\infty^*\|$ by the monotone convergence theorem, where $X_\infty^* = \sup_{n\ge 0} X_n$; so we get

$$\|X_\infty^*\| \le \frac{p}{p-1} \cdot I.$$

In particular, this implies that

$$|X_n - X_\infty|^p \le 2^p |X_\infty^*|$$

for all n; we know the left-hand side converges to 0 almost surely, and here we have a bound which is in L^1 (by the above). So we can apply the dominated convergence theorem to get that $\mathbb{E}[|X_n - X_\infty|^p] = \|X_n - X_\infty\|_p^p \to 0$ as $n \to \infty$. This means $X_n \to X_\infty$ in L^p (as $n \to \infty$), which is exactly what we wanted to show. □

Proof (2) $\implies$ (3). This is easier, because we assume almost sure and L^p convergence, and this gives us a candidate for Z — we just set $Z = X_\infty$ to be the almost sure limit. Then of course $Z \in \mathcal{L}^p$, because $X_\infty \in \mathcal{L}^p$ by (2). Here we want to show that

$$X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$$

almost surely; if we can show this, then we'll be done.

For this, it suffices to show that the pth norm of their difference is 0 (then the random variable has to be 0 almost surely). And for that, we'll use the martingale property of X_n. We have

$$\|X_n - \mathbb{E}[X_\infty \mid \mathcal{F}_n]\|_p = \|\mathbb{E}[X_m - X_\infty \mid \mathcal{F}_n]\|_p$$

for all $m \ge n$ — this is because $X_n = \mathbb{E}[X_m \mid \mathcal{F}_n]$ for all $m \ge n$, by the martingale property. Now we can apply the conditional Jensen's inequality to get that this is at most

$$\|X_m - X_\infty\|_p.$$

This is true for all $m \ge n$, so we can take $m \to \infty$. By (2), we know this term converges to 0 (that's what L^p convergence means). So that measn

$$\|X_n - \mathbb{E}[X_\infty \mid \mathcal{F}_n]\|_p = 0,$$

and therefore this random variable is 0 almost surely; so $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ almost surely. This proves (3). □

Proof of (3) $\implies$ (1). The easiest direction is to go from (3) to (1); for that, we can just apply the conditional Jensen's inequality. Recall that the conditional Jensen's inequality (applied to the function $x \mapsto x^p$) implies that

$$\|X_n\|_p = \|\mathbb{E}[Z \mid \mathcal{F}_n]\|_p \leq \|Z\|_p < \infty$$

(for all n) — so here we have a uniform bound on the pth norm of X_n over all n, which means $\sup_n \|X_n\|_p < \infty$. This completes the proof. □

That's the end of the proof; this gives equivalent characterizations for L^p convergence of martingales. A martingale converges in L^p and almost surely if and only if it's L^p-bounded, which is true if and only if it has this nice form. It's worth mentioning that for any L^p random variable Z and every filtration $\mathcal{F}_n$, the sequence $X_n = \mathbb{E}[Z \mid \mathcal{F}_n]$ is a martingale (by the tower property). And this says that a martingale converges in L^p if and only if it has this nice representation.

Next lecture we'll look at uniformly integrable martingales and backwards martingales and prove a few more theorems, before we see some applications of martingales.

§20 November 14, 2024

Last class we proved the almost sure martingale convergence theorem — if we have a supermartingale bounded in L^1, then it converges almost surely to some random variable. For that, we used Doob's upcrossing inequality. Then we also mentioned the L^p martingale convergence theorem, which gave necessary and sufficient conditions for a martingale to converge in L^p. We showed that L^p convergence is equivalent to the martingale being L^p-bounded (having a uniform bound on the L^p-norms of the sequence).

Today we'll give necessary and sufficient conditions for L^1-convergence. That condition will be uniform integrability.

§20.1 Uniform integrability

Definition 20.1. A collection $(X_i)_{i \in I}$ of random variables is called uniformly integrable (UI) if

$$\sup_{i \in I} \mathbb{E}[|X_i| \cdot \mathbf{1}_{|X_i| \geq a}] \to 0 \quad \text{as } a \to \infty.$$

Another equivalent way to describe uniform integrability involves ε and δ. That's the definition we gave earlier — that for every $\varepsilon > 0$, we can find $\delta > 0$ such that for every $A \in \mathcal{F}$ with $\mathbb{P}[A] < \delta$, we have

$$\sup_{i \in I} \mathbb{E}[|X_i| \, \mathbf{1}_A] < \varepsilon.$$

(This is something we did several lectures ago.)

One observation (which we also made when we talked about uniformly integrable families before) is that any uniformly integrable family is bounded in L^1. But the converse is not true, and we saw some counterexamples — you can have a collection which is bounded in L^1 but not uniformly integrable. However, if the family is bounded in L^p for $p > 1$, then you do have uniform integrability (we also showed this earlier).

We'll now state a theorem that will be useful.

Theorem 20.2

Let $X \in L^1$. Then the class of random variables

$$\{\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{G} \text{ is a sub-}\sigma\text{-algebra of } \mathcal{F}\}$$

is uniformly integrable.

Note that here our set I is potentially uncountable — it's the collection of all possible sub-σ-algebras that $\mathcal{F}$ has.

This will be useful because we'll later take $\mathcal{G} = \mathcal{F}_n$. Proving this is not very hard — it follows by the definition of conditional expectation and its basic properties.

Proof. First, the collection $\{X\}$ is uniformly integrable (any finite collection of L^1 random variables is uniformly integrable, and this is just a single-element collection); so for every $\varepsilon > 0$, we can find $\delta > 0$ such that for all $\mathbb{P}[A] \le \delta$, we have $\mathbb{E}[|X| \cdot \mathbf{1}_A] < \varepsilon$ (by the equivalent characterization of uniform integrability).

Now for this value of δ, we choose some very large (but finite) $\lambda > 0$ such that $\mathbb{E}[|X|] \le \lambda\delta$ (we can do this because $\mathbb{R}$ is unbounded, and $\mathbb{E}[|X|]$ is finite). This will help us find a uniform bound on the conditional expectations.

Now take any sub-σ-algebra $\mathcal{G}$. Then we have

$$\mathbb{E}[|\mathbb{E}[X \mid \mathcal{G}]|] \le \mathbb{E}[|X|] \le \lambda\delta$$

(by the conditional Jensen's inequality). This is a bound which is uniform in $\mathcal{G}$. And now if we set $Y = \mathbb{E}[X \mid \mathcal{G}]$, we can apply Markov's inequality to say that

$$\mathbb{P}[|Y| \ge \lambda] \le \frac{\mathbb{E}[|X|]}{\lambda} \le \delta.$$

(So this bound is uniform.)

Since we have $\{|Y| \ge \lambda\} \in \mathcal{G}$ (because Y is $\mathcal{G}$-measurable by construction), we have

$$\mathbb{E}[|Y| \cdot \mathbf{1}_{|Y|\ge\lambda}] = \mathbb{E}[|X| \cdot \mathbf{1}_{|Y|\ge\lambda}]$$

(by the property of the conditional expectation). But since $\{|Y| \ge \lambda\}$ has probability at most δ, the right-hand side is less than ε.

But that gives the definition of uniform integrability, which is what we wanted to prove. □

Now that we have this, we're ready to characterize the L^1-convergence of martingales. Recall a sequence of L^1 random variables converges to another in L^1 if and only if the sequence is uniformly integrable (given that you have almost sure convergence) — we stated (and probably proved) this earlier.

Fact 20.3 — Let $(X_n)_{n\ge1}$ be a sequence of random variables in L^1 such that $X_n \to X$ almost surely as $n \to \infty$. Then $X_n \to X$ in L^1 if and only if (X_n) is uniformly integrable.

In fact, here we replaced almost sure convergence by convergence in probability — we said that if you have convergence in probability, then you have L^1 convergence if and only if the sequence is uniformly integrable.

§20.2 L^1 convergence of martingales

Theorem 20.4

Let $X = (X_n)$ be a martingale. Then the following statements are equivalent:

(1) X is uniformly integrable.

(2) X_n converges almost surely and in L^1 to some limit X_∞.

(3) There exists a random variable $Z \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_n = \mathbb{E}[Z \mid \mathcal{F}_n]$ for all n.

The third characterization is very similar to what we had with L^p convergence. So for martingales, you have almost sure and L^1 convergence if and only if the martingale is uniformly integrable. But even for martingales, L^1-boundedness is not enough to guarantee L^1-convergence; we will see a counterexample.

Proof (1) $\implies$ (2). First, we need to extract an almost sure limit. But that we obtain from the martingale convergence theorem — the sequence is uniformly integrable, so it is L^1-bounded, and we've seen that L^1-bounded martingales converge almost surely to some limit. In other words, since X is uniformly integrable, it is L^1-bounded, and so the martingale convergence theorem implies that $X_n \to X_\infty$ almost surely, for some random variable $X_\infty \in L^1$ (the fact that $X_\infty \in L^1$ follows from Fatou's lemma).

And now we have a sequence which converges almost surely, and it's uniformly integrable, so it converges in L^1 by the previous fact — since X is uniformly integrable, we have $X_n \to X_\infty$ in L^1 as well, which gives (2). □

Proof (2) $\implies$ (3). For this proof, we'll use similar reasoning to as in the last lecture. We know a limit exists, so we just set Z to be that limit — i.e., $Z = X_\infty$ (just as in the previous lecture for L^p convergence). First, we clearly have $Z \in L^1$ (because $X_n \to X_\infty$ in L^1, so X_∞ is certainly in L^1).

As in last lecture, for every $m \geq n$, we have

$$\|X_n - \mathbb{E}[X_\infty \mid \mathcal{F}_n]\|_1 = \|\mathbb{E}[X_m - X_\infty \mid \mathcal{F}_n]\|_1$$

(by linearity of the conditional expectation, and the fact that $X_n = \mathbb{E}[X_m \mid \mathcal{F}_n]$ by the martingale property). Now we can apply Jensen's inequality for conditional expectations to say that this is at most $\|X_m - X_\infty\|_1$. But the right-hand side tends to 0 as $m \to \infty$. And $\|X_n - \mathbb{E}[X_\infty \mid \mathcal{F}_n]\|$ doesn't depend on m (only n, which is fixed); so it has to be 0. This implies $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ almost surely, which is exactly what we wanted to prove. □

Proof (3) $\implies$ (1). It is just an application of the tower property to show that this sequence (X_n) is a martingale — for $m \geq n$, we have

$$\mathbb{E}[\mathbb{E}[Z \mid \mathcal{F}_m] \mid \mathcal{F}_n] = \mathbb{E}[Z \mid \mathcal{F}_n]$$

by the tower property of conditional expectations. And uniform integrability follows by the previous theorem. □

So this completely characterizes the L^1 convergence of martingales.

Now we've shown almost sure convergence; and we've also characterized convergence in L^p for $p > 1$, as well as convergence in L^1. So we have a complete picture.

§20.3 Some counterexamples

We've said L^1 convergence is equivalent to uniform integrability — you need to assume something more than L^1 boundedness. We'll now see an example where L^1 boundedness isn't enough.

Example 20.5

Let (X_n) be an i.i.d. sequence of random variables such that

$$\mathbb{P}[X_n = 0] = \mathbb{P}[X_n = 2] = \frac{1}{2}.$$

Let $Y_n = X_1 \cdots X_n$. It's clear that (Y_n) is a martingale, because $\mathbb{E}[X_n] = 1$ for all n. And (Y_n) is L^1-bounded, since $\|Y_n\|_1 = 1$ for all n.

And $Y_n \to 0$ almost surely — this is because we have an independent sequence, so if we consider the events $X_n = 0$, we have a collection of independent events whose probabilities sum to ∞. So by the second Borel–Cantelli lemma, the event $X_n = 0$ has to occur infinitely many times; and whenever an event $X_n = 0$ occurs, $Y_n = 0$.

But we cannot have L^1 convergence to 0, because $\|Y_n\|_1 = 1$ for all n (which does not tend to 0).

What's wrong here is that Y_n is not uniformly integrable — you have L^1-boundedness but not uniform integrability.

§20.4 Stopping martingales at infinity

One advantage of uniformly integrable martingales is that we can stop the process even at ∞, because we know the limit exists.

Definition 20.6. If X is a uniformly integrable martingale, and T is *any* stopping time (possibly taking infinite values), then we define

$$X_T = \sum_{n=0}^{\infty} X_n \mathbf{1}_{T=n} + X_\infty \mathbf{1}_{T=\infty}.$$

So when the stopping time T is infinite, the process stopped at ∞ is just the almost sure limit, which we know exists; now there's no problem with this defintion (while for an arbitrary martingale there are some issues, which is why we needed some assumptions in the optional stopping theorem).

Now the advantage is that in the optional stopping theorem for uniformly integrable martingales, you can get rid of all the assumptions on stopping times being finite almost surely or bounded or so on.

Theorem 20.7 (Optional stopping for UI martingales)

Let X be a uniformly integrable martingale, and let S and T be stopping times (possibly taking infinite values) such that $S \le T$. Then

$$\mathbb{E}[X_T \mid \mathcal{F}_S] = X_S \quad \text{almost surely.}$$

Recall that in the OST for general martingales, we had a similar result, but we had to impose certain restrictions on the stopping time — we couldn't define the process at arbitrary stopping times. But now we can; the proof is quite smooth, and in a similar spirit to what we did last time when we proved the optional stopping theorem.

The main idea of the proof, which makes it much easier, is (3) from the previous theorem — that because X_n is uniformly integrable, we can write it in a certain nice way. So we'll be able to write $X_T = \mathbb{E}[X_\infty \mid \mathcal{F}_T]$, and then apply the tower property.

But first, (3) right now is only true for actual integers; we need to show it when n is replaced by an arbitrary stopping time.

Claim 20.8 — We have $\mathbb{E}[X_\infty \mid \mathcal{F}_T] = X_T$ almost surely, for any stopping time T.

Proof. There are two steps — first we need to show X_T defined in this way is in L^1, and then we'll show it satisfies the properties of the conditional expectation; then by uniqueness of the conditional expectation, these two things have to be hte same.

First we'll check that this is indeed a L^1 random variable. Recall that

$$|X_n| = |\mathbb{E}[X_\infty \mid \mathcal{F}_n]| \le \mathbb{E}[|X_\infty| \mid \mathcal{F}_n]$$

by the conditional Jensen's inequality. This implies that

$$\mathbb{E}[|X_T|] = \sum_{n=0}^{\infty} \mathbb{E}[|X_n| \mathbf{1}_{T=n}] + \mathbb{E}[|X_\infty| \mathbf{1}_{T=\infty}].$$

(This is just by the definition.) To bound these terms, we just use the above inequality — this is at most

$$\sum_{n=0}^{\infty} \mathbb{E}[\mathbb{E}[|X_\infty| \mid \mathcal{F}_n]\mathbf{1}_{T=n}] + \mathbb{E}[|X_\infty| \mathbf{1}_{T=\infty}].$$

And the event that $T = n$ is $\mathcal{F}_n$-measurable, so we can get rid of the conditioning (by properties of the conditional expectation) — we can write this as

$$\sum_{n=0}^{\infty} \mathbb{E}[|X_\infty| \mathbf{1}_{T=n}] + \mathbb{E}[|X_\infty| \mathbf{1}_{T=\infty}].$$

And this is just $\mathbb{E}[|X_\infty|]$, which is finite (because $X_\infty \in L^1$). This implies $X_T \in L^1$. So indeed we have an L^1 random variable.

Now, we want to show the equality $\mathbb{E}[X_\infty \mid \mathcal{F}_T] = X_T$. For this, we'll show that X_T satisfies the properties of the conditional expectation. To do this, fix $B \in \mathcal{F}_T$. Then we have

$$\mathbb{E}[X_T \mathbf{1}_B] = \sum_{n \in \mathbb{N} \cup \{0,\infty\}} \mathbb{E}[X_n \mathbf{1}_{T=n} \mathbf{1}_B].$$

But now, we have $B \cap \{T = n\} \in \mathcal{F}_n$ by the definition of $\mathcal{F}_T$. So we're integrating $|X_n|$ on an event which is $\mathcal{F}_n$-measurable. And X_n is the conditional expectation of X_∞ given $\mathcal{F}_n$, and we're integrating it on an event which is $\mathcal{F}_n$-measurable. This means we can replace it with X_∞ by the definition of the conditional expectation — so this is

$$\sum_{n \in \mathbb{N} \cup \{0,\infty\}} \mathbb{E}[X_\infty \cdot \mathbf{1}_{\{T=n\} \cap B}].$$

But this is just $\mathbb{E}[X_\infty \cdot \mathbf{1}_B]$, because we're summing over all possible values of T.

So we've shown that $\mathbb{E}[X_T \mathbf{1}_B] = \mathbb{E}[X_\infty \mathbf{1}_B]$. By the uniqueness of the conditional expectation, this means $\mathbb{E}[X_\infty \mid \mathcal{F}_T] = X_T$ almost surely. □

Proof of OST for UI martingales. Now we can just plug in the value of $X_T = \mathbb{E}[X_\infty \mid \mathcal{F}_T]$ and use the tower property — we have

$$\mathbb{E}[X_T \mid \mathcal{F}_S] = \mathbb{E}[\mathbb{E}[X_\infty \mid \mathcal{F}_t] \mid \mathcal{F}_s] = \mathbb{E}[X_\infty \mid \mathcal{F}_S] = X_S$$

(the second step is the tower property, and the third is the same fact). □

So now we have the OST in full generality for uniformly integrable martingales.

Student Question. *How did we replace X_n with X_∞ in $\mathbb{E}[X_n \mathbf{1}_{T=n} \mathbf{1}_B]$?*

Answer. We can write $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$ by the previous theorem, so this is

$$\mathbb{E}[\mathbb{E}[X_\infty \mid \mathcal{F}_n] \mathbf{1}_{B \cap \{T=n\}}].$$

But the event $B \cap \{T = n\}$ is in $\mathcal{F}_n$, by the definition of $\mathcal{F}_T$. So then we can get rid of the conditioning, and this is exactly $\mathbb{E}[X_\infty \mathbf{1}_{B\cap\{T=n\}}]$.

§20.5 Backwards martingales

So that's more or less the theory of forwards martingales — forward in the sense that time goes forward (to infinity). But one might wonder what happens if you have *backwards* martingales — what if you take negative times? That's the next thing we'll analyze.

Here we have negative time, which is counterintuitive; so we'll have to be careful making things precise.

We'll start with a sequence of σ-algebras indexed by the set of negative integers — so we have a sequence $\mathcal{G}_0 \supseteq \mathcal{G}_{-1} \supseteq \mathcal{G}_{-2} \supseteq \cdots$. (So now we have a decreasing sequence of σ-algebras.) This is still a filtration, because if we saw this in the forwards direction, it's an increasing sequence; we just don't have a starting point.

Notation 20.9. We use $\mathbb{Z}_-$ to denote the set of negative integers.

Definition 20.10. Given a filtration $\cdots \subseteq \mathcal{G}_{-2} \subseteq \mathcal{G}_{-1} \subseteq \mathcal{G}_0$, a sequence $(X_n)_{n\in\mathbb{Z}_-}$ is called a backwards martingale if it is adapted to the filtration (i.e., X_n is $\mathcal{G}_n$-measurable), $X_0 \in L^1$, and

$$\mathbb{E}[X_{n+1} \mid \mathcal{G}_n] = X_n$$

almost surely for all $n \in \mathbb{Z}_-$.

Here things are more or less almost the same, but not quite. The definitions are very natural. But the difference is we have reversed time — we're going towards the past rather than the future.

This has certain advantages that we'll see — it has nice behavior. One advantage is that you can express X_n in a very nice way. By the tower property of conditional expectations, we have

$$X_n = \mathbb{E}[X_0 \mid \mathcal{G}_n]$$

for all $n \in \mathbb{Z}_-$ (almost surely). So now X_n has this nice form; we didn't have this in the forwards case, where time ran forwards. We'll see that this somehow makes things easier.

The first observation is that (X_n) is uniformly integrable, because it has the form $\mathbb{E}[X_0 \mid \mathcal{G}_n]$ (by the previous theorem) — since $X_0 \in L^1$, we get that the family X is uniformly integrable. And using this, we can prove things. We will see that we'll have almost sure convergence and L^1 convergence. Actually what we'll show is that if we start with a L^p random variable for some $p \geq 1$, then X_n converges almost surely and in L^p (as $n \to -\infty$), because of this nice expression.

Intuitively, someone could see that step by step, we're losing a bit of information and going backwards.

§20.6 Convergence of backwards martingales

Theorem 20.11

Let X be a backwards martingale, and suppose that $X_0 \in L^p$ for some $p \in [1, \infty)$. Then X_n converges almost surely and in L^p as $n \to -\infty$, to the random variable given by

$$X_{-\infty} = \mathbb{E}[X_0 \mid \mathcal{G}_{-\infty}]$$

(where $\mathcal{G}_{-\infty} = \bigcap_{n \leq 0} \mathcal{G}_n$).

So here we have almost sure and L^p convergence, because X has this nice expression.

Student Question. *Why can't p be ∞?*

Answer. If you have $p = \infty$, then you can't really prove a reasonable notion of convergence — L^∞ is not a very rich space, so you can't hope for convergence there.

The spirit of the proof is similar to the proof of the almost sure martingale convergence theorem. We're again going to use Doob's upcrossing inequality to extract the limit. And for L^p convergence, we'll use uniform integrability; but we'll still need to go through Doob's upcrossing inequality.

§20.6.1 Adapting Doob's upcrossing inequality

First, we'll adapt Doob's upcrossing inequality — for any $a < b$, we set $N_{-n}([a, b], X)$ to be the number of up-crossings of $[a, b]$ by the process X between the times $-n$ and 0. Now, somehow in order not to repeat ourselves, we need to go back to the setup where time is forwards and use Doob's upcrossing inequality in the forwards case. For that, we'll reverse time — to go from 0 to n rather than $-n$ to 0. But in order to apply Doob's upcrossing inequality directly, we need to be a bit careful — we need to create a martingale and the appropriate filtration.

For this, we consider $\mathcal{F}_k = \mathcal{G}_{-n+k}$ (for $k = 0, \ldots, n$). Then we get a filtration (stopped at some time n). Then if we consider (X_{-n+k}) for $0 \leq k \leq n$, this is a martingale with respect to our filtration. So now we have created a martingale.

And now we can apply Doob's upcrossing inequality up to time n for this martingale. The crucial observation here is that the number of upcrossings doesn't change if we reverse time — it doesn't matter whether we're going in the forwards or the backwards direction. So this means Doob's upcrossing inequality implies that

$$(b - a) \cdot \mathbb{E}[N_{-n}([a, b], X)] \leq \mathbb{E}[(X_0 - a)^-].$$

(This follows by the forward Doob's upcrossing inequality — there we have the bound $\mathbb{E}[(Y_n - a)^-]$, and here we've defined $Y_k = X_{-n+k}$, so $Y_n = X_0$.)

Now we let $n \to \infty$. The bound on the right-hand side is finite, because we assumed $X_0 \in L^1$ (actually it's in L^p, which is even better), and uniform in n. So we can take $n \to \infty$, using the monotone convergence theorem — by the monotone convergence theorem, we get that

$$\mathbb{E}[N_{-n}([a, b], X)] \to \mathbb{E}[N_{-\infty}([a, b], X)]$$

(where $N_{-\infty}$ denotes the total number of upcrossings). This implies the expected total number of up-crossings is at most $\mathbb{E}[(X_0 - a)^-]$, which is finite. In particular, the total number of up-crossings of the interval is finite almost surely. In particular, it'll be finite over all rationals a and b and so on, and we can use exactly the same thing as we did in the almost sure martingale convergence theorem, which implies X_m converges to some random variable $X_{-\infty}$ almost surely. So now we have our limit, where $X_{-\infty}$ is some random variable (though not explicit); but it will be $\mathcal{G}_{-\infty}$-measurable.

So we have the almsot sure convergence. Now what we want to do is to prove the L^p convergence. For that, we're going to use uniform integrability.

First, as we mentioned earlier, by the conditional Jensen's inequality we have $\|X_n\|_p \le \|X_0\|_p < \infty$ for all n. So all of them are in L^p, and the bound is uniform. In particular, this implies that $X_{-\infty}$ is also in L^p — this follows by Fatou's lemma, because

$$\mathbb{E}[|X_{-\infty}|^p] = \mathbb{E}[\liminf_{n\to-\infty} |X_n|^p] \le \liminf_{n\to\infty} \mathbb{E}[|X_n|^p] \le \mathbb{E}[|X_0|^p] < \infty.$$

So $X_{-\infty}$ is indeed a L^p random variable.

Now we want to bound the pth power of the distance between X_n and $X_{-\infty}$. And because X_n has this very nice expression, we're going to use conditional Jensen's inequality — by the conditional Jensen's inequality, we have that

$$|X_n - X_{-\infty}|^p = |\mathbb{E}[X_0 - X_{-\infty} \mid \mathcal{G}_n]|^p .$$

And now we can bring $|\bullet|^p$ inside by Jensen's — so this si at most

$$\mathbb{E}[|X_0 - X_{-\infty}|^p \mid \mathcal{G}_n].$$

And $X_0 - X_{-\infty}$ is a fixed random variable, so this sequence is uniformly integrable (by the first theorem); and that implies $|X_n - X_{-\infty}|^p$ is also uniformly integrable (as a collection over n). And it converges to 0 almost surely, so that implies this sequence will also converge in L^1 — it's uniformly integrable and converges almost surely to 0, so it will converge to 0 in L^1. In particular, this means $\mathbb{E}[|X_n - X_{-\infty}|^p] \to 0$ as $n \to -\infty$, which is exactly L^p convergence. So indeed, we also have L^p convergence.

So we've proved the almost sure existence of the limit, and also L^p convergence. The last thing it remains to show is that $X_{-\infty} = \mathbb{E}[X_0 \mid \mathcal{G}_{-\infty}]$. For that, as usual we'll use the properties of the conditional expectation.

If we fix any set $A \in \mathcal{G}_{-\infty}$, then $A \in \mathcal{G}_n$ for all n (because $\mathcal{G}_{-\infty}$ is the intersection of all these σ-algebras), then we have

$$\mathbb{E}[X_0 \mathbf{1}_A] = \mathbb{E}[X_n \mathbf{1}_A]$$

for all $n \le 0$, by the fact that $X_n = \mathbb{E}[X_0 \mid \mathcal{G}_n]$. And now we know that $X_n \to X_{-\infty}$ in L^p, which means $X_n \mathbf{1}_A \to X_{-\infty}\mathbf{1}_A$ in L^p. And L^p convergence implies L^1 convergence, so $\mathbb{E}[X_0\mathbf{1}_A] \to \mathbb{E}[X_{-\infty}\mathbf{1}_A]$. And A was arbitrary, so this completes the proof. So this implies $X_{-\infty} = \mathbb{E}[X_0 \mid \mathcal{G}_{-\infty}]$, by the uniqueness of the conditional expectation.

So the point is when we reverse time, we have some nice behavior.

Next lecture we'll prove some nice properties like Kolgomorov's 0-1 law, the strong law of large numbers, and so on using martingale theory. And this will be the end of the chapter on discrete-time martingales; then we'll move to continuous-time martingales, where n is a nonnegative real number instead of an integer.

§21 November 19, 2024

Last lecture, we completely characterized the L^1 convergence of martingales — a martingale converges almost surely in L^1 if and only if it's uniformly integrable. This completes the characterization of L^p and almost sure convergence of martingales. We also introduced martingales where we reverse time (where we look to the past instead of future); we also proved that in that case we have L^p convergence if and only if the sequence is bounded in L^p (for $p > 1$).

We've now exhausted most of the theory of discrete-time martingales, and now we'll see some applications. One is to Kolgomorov's 0-1 law (which we've already proved, but we'll prove again using martingales), and another is the strong law of large numbers. We'll also prove the Radon–Nikodym theorem (the existence of the Radon–Nikodym derivative), and Caputani's martingale product theorem.

§21.1 Kolgomorov's 0–1 law

We'll start with the first application, an alternative proof of Kolgomorov's 0–1 law.

Theorem 21.1

Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. random variables, and let $\mathcal{F}_n = \sigma(X_k \mid k \geq n)$ and $\mathcal{F}_\infty = \bigcap_{n\geq 1} \mathcal{F}_n$. Then for all $A \in \mathcal{F}_\infty$, we have $\mathbb{P}[A] \in \{0, 1\}$.

Recall that $\mathcal{F}_\infty$ is called the *tail* σ-algebra. And this theorem states that it's trivial, in the sense that every event in it has probability either 0 or 1.

Proof. In order to construct a martingale, we first need to construct a filtration — an increasing sequence of σ-algebras with respect to which our sequence (X_i) is measurable. The obvious choice is to consider $\mathcal{G}_n = \sigma(X_k \mid k \leq n)$ (the σ-algebra generated by the first n terms).

Now fix some $A \in \mathcal{F}_\infty$; we want to show $\mathbb{P}[A]$ is either 0 or 1. The martingale we'll consider is

$$Y_n = \mathbb{E}[\mathbf{1}_A \mid \mathcal{G}_n].$$

It's easy to see that Y_n is a martingale (we've already encountered this kind of martingale).

But the point is that this martingale is constant. Why? We have that $\mathcal{G}_n$ and $\mathcal{F}_{n+1}$ are independent (because $\mathcal{G}_n$ encodes information about the first n terms of the sequence, and $\mathcal{F}_{n+1}$ encodes information about the rest; and since we have *independent* random variables, these two σ-algebras are independent). And we also have that $\mathcal{F}_\infty \subseteq \mathcal{F}_{n+1}$, so $A \in \mathcal{F}_{n+1}$. So Y_n has a simple form — it's just $\mathbb{P}[A]$ (by properties of the conditional expectation). So we obtain that

$$Y_n = \mathbb{P}[A] \quad \text{almost surely}$$

for all n. So here we have a *constant* sequence (almost surely).

And also, we know that (Y_n) converges, by the almost sure martingale convergence theorem — so

$$Y_n \to \mathbb{E}[\mathbf{1}_A \mid \mathcal{G}_\infty]$$

almost surely (this follows by the almost sure martingale convergence theorem).

But $\mathcal{F}_\infty \subseteq \mathcal{G}_\infty$, which implies that this conditional expectation is just

$$\mathbb{E}[\mathbf{1}_A \mid \mathcal{G}_\infty] = \mathbf{1}_A \quad \text{almost surely}$$

(because $\mathbf{1}_A$ is measurable with respect to $\mathcal{G}_\infty$). So this means

$$Y_n = \mathbb{P}[A] \to \mathbf{1}_A \quad \text{almost surely,}$$

and this implies $\mathbb{P}[A] = \mathbf{1}_A$ almost surely. But $\mathbf{1}_A$ takes values either 0 or 1; so this implies $\mathbb{P}[A]$ (which is just a fixed number) has to be either 0 or 1. And that's the end of the proof. □

Student Question. *Which result are you using when you say $Y_n = \mathbb{P}[A]$?*

Answer. If you have a random variable and condition on an independent σ-algebra, then the conditional expectation is just the expectation, and $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A]$.

This is a very elegant and simple proof, given that you know martingales. The idea is to construct the right martingale and the right filtration, and then just apply the almost sure martingale convergence theorem. The nontrivial thing is to come up with the right martingale.

So far, we've seen two different ways to prove Kolgomorov's 0–1 law. Of course there are infinitely many ways to do this; we've only seen two because there's a limit on the time. But Prof. Kavvadias prefers the second, because it's much more elegant and illustrates all the theory we've learned.

§21.2 The strong law of large numbers

Our next result will be the strong law of large numbers. We've already proved this. (We proved two versions — in the other version we assumed that the variables were independent, but instead of assuming they had the same law, we just assumed they have the same expectations and a uniform bound on the 4th moments. Here we'll consider the version where they're i.i.d.)

Theorem 21.2 (Strong law of large numbers)

Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. random variables in L^1 with mean $\mu = \mathbb{E}[X_i]$. Let $S_n = X_1 + \cdots + X_n$ (with $S_0 = 0$). Then $S_n/n \to \mu$ almost surely and in L^1 as $n \to \infty$.

Proof. We're going to prove this now using martingale theory. Here it's not clear directly what kind of martingale to consider — something analogous to before doesn't work because we're dividing by n, and n changes. So it's not clear how to cook up a martingale. For that reason, we'll reverse time — we'll consider a *backwards* martingale.

But first, we need to consider a decreasing sequence of σ-algebras. For this, we let

$$\mathcal{G}_n = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots).$$

This is the same as $\sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$ (because if you know S_n and the remaining averages then you can recover X_{n+1}, X_{n+2}, …, and the reverse as well).

Claim 21.3 — If we define $M_n = S_{-n}/(-n)$ (for $n < 0$), then M_n is a backwards martingale with respect to the filtration $\mathcal{F}_n = \mathcal{G}_{-n}$.

So we're going to show that M_n is a backwards martingale, and then we're going to invoke the almost sure and L^1 convergence of those martingales to deduce the strong law of large numbers (the statement of the strong law of convergence is equivalent to the almost sure and L^1 convergence of M_n).

Proof. Consider $\mathbb{E}[M_{m+1} \mid \mathcal{F}_m]$. By definition, this is

$$\mathbb{E}[M_{m+1} \mid \mathcal{F}_m] = \mathbb{E}\left[\frac{S_{-m-1}}{-m-1} \mid \mathcal{G}_{-m}\right]$$

(for $m < 0$). Let's analyze the term on the right-hand side; because what we want to show is that the left-hand side is M_m.

Let $n = -m$. Then because X_n is independent of X_{n+1}, X_{n+2}, X_{n+3}, …, we obtain that

$$\mathbb{E}\left[\frac{S_{n-1}}{n-1} \,\middle|\, \mathcal{G}_n\right] = \mathbb{E}\left[\frac{S_n - X_n}{n-1} \,\middle|\, \mathcal{G}_n\right].$$

And $S_n/(n-1)$ is $\mathcal{G}_n$-measurable, so we can take it out of the expectation, and this is

$$\frac{S_n}{n-1} - \mathbb{E}\left[\frac{X_n}{n-1} \,\middle|\, \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)\right].$$

And there's a property we proved of conditional expectation — here $\sigma(X_{n+1}, X_{n+2}, \ldots)$ is independent of $\sigma(X_n, S_n)$, and this means we can forget it. So this reduces to just

$$\frac{S_n}{n-1} - \mathbb{E}\left[\frac{X_n}{n-1} \,\middle|\, S_n\right].$$

And now we actually need to compute this last conditional expectation. The crucial thing is the observation that

$$S_n = \mathbb{E}[S_n \mid S_n] = \mathbb{E}[X_1 \mid S_n] + \mathbb{E}[X_2 \mid S_n] + \cdots + \mathbb{E}[X_n \mid S_n]$$

(by the linearity of the conditional expectation). And now, the point is that all the $\mathbb{E}[X_k \mid S_n]$ are the same (because the X_k have the same law and are independent). This implies that the right-hand side is just $n \cdot \mathbb{E}[X_n]$. So dividing by n, we obtain

$$\mathbb{E}[X_n \mid S_n] = \frac{S_n}{n}.$$

And then we just plug it into our earlier equation, and obtain that

$$\mathbb{E}\left[\frac{S_{n-1}}{n-1} \;\middle|\; \mathcal{G}_n\right] = \frac{S_n}{n-1} - \frac{S_n}{n(n-1)} = \frac{S_n}{n}$$

(almost surely). And this converts to the claim we want — recall that $n = -m$, so this implies that

$$\mathbb{E}[M_{m+1} \mid \mathcal{F}_m] = M_m$$

(for all $m \leq -1$) almost surely. So this shows that M is indeed a backwards martingale. □

And now we have the theory, and we can apply the backwards martingale convergence theorem. We've already done most of the hard work — we've shown almost sure and L^1 convergence of backwards martingales. So by the backwards martingale convergence theorem, we obtain that (S_n/n) converges almost surely and in L^1 as $n \to \infty$.

So we've shown (S_n/n) converges to *something*. We're not exactly done, but we're almost done — we need to show that the limit is constant almost surely (because then it has to be μ, because we have L^1 convergence — this is the same argument as in the ergodic theory proof). So set $Y = \lim_{n\to\infty}(S_n/n)$ (this is an almost sure and L^1 limit). The idea is to prove that Y is measurable with respect to the tail σ-algebra generated by the X_n; if that's the case, then it has to be constant almost surely, because the tail σ-algebra is trivial (by the Kolgomorov's 0–1 law). (We've seen in a previous lecture that if a random variable is measurable with respect to a trivial σ-algebra, then it has to be constant almost surely.)

To see this, observe that if we shift time by any constant integer, the limit is still the same — so we have

$$Y = \lim_{n\to\infty} \frac{X_{k+1} + X_{k+2} + \cdots + X_{k+n}}{n} \quad \text{almost surely for every } k \in \mathbb{N}.$$

This implies Y is measurable with respect to $\sigma(X_{k+1}, X_{k+2}, \ldots)$, for all k. So therefore it's measurable with respect to their intersection — if we let $\mathcal{T}_k = \sigma(X_{k+1}, X_{k+2}, \ldots)$, then Y is measurable with respect to $\bigcap_k \mathcal{T}_k$. But the intersection is just the tail σ-algebra, and by Kolgomorov's 0–1 law, we obtain that there exists some $c \in \mathbb{R}$ such that $Y = c$ almost surely.

And in particular, this implies $S_n/n \to c$ as $n \to \infty$ almost surely and in L^1. In particular, we have $\mathbb{E}[S_n/n] = \mu \to c$ as $n \to \infty$, which means $\mu = c$; and that completes the proof (showing that the average converges to μ almost surely and in L^1). □

So that's the proof of the strong law of large numbers using martingale theory. We still used Kolgomorov's 0–1 law, but you can use martingale theory to prove that as well.

This was a bit tricky — we had to reverse time, because we had to divide by n and it wasn't clear what martingale to consider if we considered forwards time. But if you reverse time, you magically obtain a backwards martingale.

§21.3 Kakutani's product martingale convergence theorem

We've already seen two applications of martingales. Another application is Kakutani's product martingale theorem. This is related to the convergence of products of random variables (instead of sums of random variables) — it's some kind of a version of the strong law of large numbers, but instead of summation, you have multiplication.

Theorem 21.4 (Kakutani's product martingale convergence theorem)

Let $(X_n)_{n\geq 0}$ be a sequence of independent nonnegative random variables with $\mathbb{E}[X_n] = 1$ for all n, and consider the *product martingale* M where $M_0 = 1$ and

$$M_n = X_1 X_2 \cdots X_n.$$

Then (M_n) is a nonnegative martingale, and converges almost surely to some random variable M_∞.

Furthermore, let $a_n = \mathbb{E}[\sqrt{X_n}]$.

(1) If $\prod_n a_n > 0$, then the convergence $M_n \to M_\infty$ is also in L^1, and $\mathbb{E}[M_\infty] = 1$.

(2) If $\prod_n a_n = 0$, then $M_\infty = 0$ almost surely.

Note that the X_n don't necessarily have the same law.

This is a result which is very similar in flavor to martingale convergence theorems and the strong law of large numbers, but instead of sums we have products (so far we've dealt only with sums, not with products).

This was originally proved with a different method, but now we're going to use martingale theory.

Proof. The first thing we want to do is extract an almost sure limit. This follows by the almost sure martingale convergence theorem (so we've already done the hard work for that) — by the almost sure martingale convergence theorem, we have that $M_n \to M_\infty$ almost surely for some random variable M_∞. (Recall that $\mathbb{E}[M_n]$ is always 1, which is bounded uniformly in n; and if you have a martingale which is bounded in L^1, then it converges almost surely (but not necessarily in L^1) to *some* random variable.)

For (1) and (2), the idea is to somehow apply martingale theory for an appropriate martingale. The first observation is that $a_n = \mathbb{E}[\sqrt{X_n}] \leq \sqrt{\mathbb{E}[X_n]} = 1$ by Cauchy–Schwarz or Hölder's inequality. Set

$$N_n = \frac{\sqrt{X_1 \cdots X_n}}{a_1 \cdots a_n}.$$

This is again a martingale whose expectation is $\mathbb{E}[N_n] = 1$ for all n, which means we can again apply the almost sure martingale convergence theorem to deduce that it converges almost surely to some random variable N_∞. So again by the almost sure martingale convergence theorem, we have that $N_n \to N_\infty$ almost surely (as $n \to \infty$) for some random variable N_∞.

Now we need to deduce certain properties of N using (1) and (2). First, we'll assume (1), which states that the infinite product $\prod_{n\geq 1} a_n$ is positive. This will be useful because we want to somehow apply the dominated convergence theorem to deduce L^1 convergence of M_n, but *a priori* it's not clear what the dominant function is. Somehow we need to take a supremum over n and find a bound that's uniform in n. For that we'll use Doob's maximum inequality combined with condition (1), which will guarantee that the expectation of the supremum of the first n terms of the sequence is bounded uniformly, and then take a supremum over all n and get a uniform bound.

So assuming (1), we have that

$$\sup_{n\geq 0} \mathbb{E}[N_n^2] = \sup_{n\geq 0} \frac{1}{\prod_{i=1}^n a_i^2} < \infty,$$

because we assumed the denominator is nonzero. And also, since $M_n = N_n^2 \cdot \prod_{i=1}^n a_i^2 \le N_n^2$, we get that

$$\mathbb{E}\left[\sup_{k\le n} M_k\right] \le \mathbb{E}\left[\sup_{1\le k\le n} N_k^2\right]$$

(again by Holder's inequality). And by Doob's maximal inequality, this is at most $4 \cdot \mathbb{E}[N_n^2]$. And this expectation is bounded uniformly — we have

$$4 \cdot \mathbb{E}[N_n^2] < \frac{4}{\prod_{n\ge 0} a_n^2} < \infty.$$

So taking $n \to \infty$, we get that

$$\mathbb{E}\left[\sup_{n\ge 0} M_n\right] < \frac{4}{\prod_n a_i^2} < \infty.$$

And now we have a dominant function — the dominant function is $\sup_n M_n$. This random variable dominates all the M_n, and we know the M_n converge almost surely to M_∞, so by the dominated convergence theorem we have that the convergence is in L^1 as well.

So by the dominated convergence theorem, we obtain that $M_n \to M_\infty$ in L^1. And clearly then $\mathbb{E}[M_\infty] = 1$, because $\mathbb{E}[M_n] = 1$ for all n. So we have shown (1). Here we used the martingale convergence theorem to deduce that this indeed converges almost surely, and then we crucially used Doob's maximal inequality to bound the expectation of the supremum of the first n terms.

Now we need to prove the second case (2), where we want to show $M_n \to 0$ almost surely (if $\prod_n a_n = 0$). For this, remember that

$$M_n = N_n^2 \prod_{i=1}^n a_i^2.$$

If the infinite product is 0, then the product on the right converges to 0, and N_n converges to some (finite) limit. So their product has to converge to 0. So we get that

$$M_n = N_n^2 \prod_{i=1}^n a_i^2 \to N_\infty^2 \prod_{i\ge 0} a_i^2.$$

And N_∞^2 is some finite random variable (almost surely), while the infinite product is 0 almost surely. So this means $M_n \to 0$ almost surely.

Here we used the martingale convergence theorem to deduce that the martingale N converges somewhere (so that's a nontrivial input here). □

This can of course be shown using different ways, but Prof. Kavvadias prefers this way because we somehow convince ourselves that all this theory can be very useful.

Student Question. *Why does $\mathbb{E}[\sup M_n]$ being bounded imply that $M_n \to M_\infty$ in L^1?*

Answer. This is the dominated convergence theorem — the dominated convergence theorem gives L^1 convergence. You have almost sure convergence, and that all your functions are dominated from above by some function with finite integral. Under these assumptions, you have L^1 convergence as well. Here the function that dominates everything is $\sup M_n$, and we showed here that it has finite expectation.

Alternatively, this bound on $\mathbb{E}[\sup M_n]$ shows that (M_n) is uniformly integrable, and we know uniformly integrable martingales converge in L^1 (we showed this in the previous lecture).

§21.4 The Radon–Nikodym theorem

This is the third application of martingale theory. We'll see one more, the last for today (next lecture we'll see continuous-time martingales), the Radon–Nikodym theorem — the existence of the Radon–Nikodym derivative between two measures which are absolutely continuous.

Theorem 21.5 (Radon–Nikodym)

Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Assume that $\mathcal{F}$ is countably generated, i.e., that there exists a (countable) collection of sets $(F_n)_{n\in\mathbb{N}}$ such that $\mathcal{F}_n = \sigma(F_n \mid n \in \mathbb{N})$. Then the following statements are equivalent:

(a) $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ — i.e., if $\mathbb{P}[A] = 0$, then $\mathbb{Q}[A] = 0$ (for all $A \in \mathcal{F}$).

(b) For every $\varepsilon > 0$, there is $\delta < 0$ such that for every $\mathbb{P}[A] \le \delta$, we have $\mathbb{Q}[A] \le \varepsilon$.

(c) There exists a nonnegative random variable X such that $\mathbb{Q}[A] = \mathbb{E}[X \cdot \mathbf{1}_A]$ for all $A \in \mathcal{F}$.

This can be generalized to σ-finite measures, but we'll focus on probability measures. The assumption that $\mathcal{F}$ is countably generated is not needed in general, but here it makes the proof easier.

Note that (a) doesn't require the converse — every $\mathbb{P}$-0 event is also $\mathbb{Q}$-0, but the converse isn't required. And (b) is a rephrase of (a); this is not obvious, but we're going to prove it. And (c) is very important. Then X is called a version of the *Radon–Nikodym derivative* of $\mathbb{Q}$ with respect to $\mathbb{P}$. If the measures are absolutely continuous in this sense (as in (a) or (b)), then this theorem says the Radon–Nikodym derivative exists.

Here we've assumed the σ-algebra is countably easier just to make the proof easier, but this is also true in an arbitrary measurable space (with no restriction). And the same is also true, with the same proof, for any finite measures (instead of probability measures), or even σ-finite ones.

We're going to prove this using martingale theory. (If you don't use martingale theory, you can still do it, but the proof is more technical and complicated, which is why people don't usually teach it.)

Proof $(a) \implies (b)$. Here we assume absolute continuity, and want to prove an ε-δ version. We'll prove this by contradiction — suppose that (b) does *not* hold, so there exists some $\varepsilon > 0$ such that for all $n \in \mathbb{N}$, there is a set $A_n \in \mathcal{F}$ with the property that

$$\mathbb{P}[A_n] \le \frac{1}{n^2} \quad \text{and} \quad \mathbb{Q}[A_n] \ge \varepsilon$$

(this is the negation of (b), replacing δ with $\frac{1}{n^2}$). Then by the first Borel–Cantelli lemma, we have that $\mathbb{P}[A_n \text{ i.o.}] = 0$ (i.e., the probability that A_n happens for infinitely many values of n is 0 — i.o. stands for 'infinitely often').

So then (a) gives that $\mathbb{Q}[A_n \text{ i.o.}] = 0$ as well. But on the other hand, by definition we have

$$\mathbb{Q}[A_n \text{ i.o.}] = \mathbb{Q}\left[\bigcap_n \bigcup_{k\ge n} A_k\right].$$

And $\bigcup_{k\ge n} A_k$ is a decreasing sequence, so this is just

$$\lim_{n\to\infty} \mathbb{Q}\left[\bigcup_{k\ge n} A_k\right].$$

And this term is at least ε (because $\mathbb{Q}[\bigcup_{k\ge n} A_k] \ge \mathbb{Q}[A_n]$, which is at least ε by assumption). So $\mathbb{Q}[A_n \text{ i.o.}] \ge \varepsilon > 0$, which is a contradiction. □

Proof $(b) \implies (c)$. From (b) to (c), we need to construct an appropriate martingale; and we'll show that this martingale converges almost surely to the Radon–Nikodym derivative. But for that, we need a filtration (as usual). We'll consider the natural filtration $\mathcal{F}_n = \sigma(F_1, F_2, \dots, F_n)$ — the σ-algebra generated by the first n of our events F_i.

How do we describe this σ-algebra? It's not too hard to describe — we can write it as $\sigma(\mathcal{A}_n)$ where

$$\mathcal{A}_n = \{H_1 \cap \cdots \cap H_n \mid H_i = F_i \text{ or } F_i^{\mathsf{c}} \quad \text{for all } 1 \le i \le n\}.$$

So here we're taking all the possible intersections of events which are either F_i or its complement, for all $i = 1, \dots, n$. And $\mathcal{F}_n$ is the σ-algebra generated by events of this form (this is not hard to see — clearly $\sigma(\mathcal{A}_n)$ contains all the F_j's, and $\mathcal{F}_n$ also contains $\mathcal{A}_n$, so this shows both directions).

Now we also need a martingale. For this, we define $X_n \colon \Omega \to \mathbb{R}_{\ge 0}$ by

$$X_n(\omega) = \sum_{A \in \mathcal{A}_n} \frac{\mathbb{Q}[A]}{\mathbb{P}[A]} \cdot \mathbf{1}_{\omega \in A}$$

(note that there are finitely many sets in $\mathcal{A}_n$). This might remind you of the Riemann or Lebesgue integral. There we discretize space, and that's exactly what we've done here, using $\mathcal{A}_n$ — $\mathcal{F}$ has a nice form, so we just discretize it. And this thing we're summing is just the value of the integral of the supposed Radon–Nikodym derivative on A.

Then for this choice of X_n, we have

$$\mathbb{Q}[A] = \mathbb{E}[X_n \mathbf{1}_A]$$

(all expectations are with respect to $\mathbb{P}$) for all $A \in \mathcal{F}_n$ — this follows by the definition of X_n. We also have that $(X_n)_{n \ge 0}$ is a nonnegative martingale with respect to the filtration $(\mathcal{F}_n)$.

So now we have a filtration and a martingale, and it's also bounded in L^1 — we have

$$\mathbb{E}[X_n] = \mathbb{Q}[A] = 1 \quad \text{for all } n,$$

so X_n is an L^1-bounded martingale. This means by the almost sure martingale convergence theorem that we have $X_n \to X_\infty$ almost surely, for some random variable X_∞. And this X_∞ is our candidate — this is what'll be the Radon–Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$.

But we're not exactly done — we had the nice equality $\mathbb{Q}[A] = \mathbb{E}[X_n \cdot \mathbf{1}_A]$, but somehow we need to take limits inside. To do this, we'd like to know that $X_n \to X_\infty$ in L^1. For that, we need the L^1 martingale convergence theorem, and for this we need to show uniform integrability — a uniformly integrable martingale converges almost surely and in L^1. So if we can show X_n is uniformly integrable, then we're done (because we can put the limit inside to get that $\mathbb{Q}$ satisfies (c)).

For this, fix some $\delta > 0$, and set $\lambda = 1/\delta$. Then by Markov's inequality, we have

$$\mathbb{P}[X_n \ge \lambda] \le \frac{\mathbb{E}[X_n]}{\lambda} = \frac{1}{\lambda} \le \delta.$$

This implies that

$$\mathbb{E}[X_n \cdot \mathbf{1}_{X_n \ge \lambda}] = \mathbb{Q}[X_n \ge \lambda] \le \varepsilon$$

by (b) (since whenever the probability of an event under A is less than δ, its probability under $\mathbb{Q}$ is less than ε by our assumption (b)). This is true for all n, so therefore (X_n) is uniformly integrable.

And so we also obtain that $X_n \to X_\infty$ in L^1 (not just almost surely, but also in L^1). And we had $\mathbb{Q}[A] = \mathbb{E}[X_n \cdot \mathbf{1}_A]$, so this means

$$\mathbb{Q}[A] = \mathbb{E}[X_\infty \mathbf{1}_A]$$

for all $A \in \mathcal{F}_n$ (for all n). And since all the $\mathcal{F}_n$ generates the entire space, we obtain that

$$\mathbb{Q}[A] = \mathbb{E}[X_\infty \mathbf{1}_A] \quad \text{for all } A \in \mathcal{F}.$$

So X_∞ indeed satisfies the properties of the Radon–Nikodym derivative, which proves (c). □

Proof $(c) \implies (a)$. This direction is obvious — if we assume $\mathbb{P}[A] = 0$, and we integrate a nonnegative random variable over an event of measure 0, then we also get something of measure 0 (by properties of the integral that we have shown). So $\mathbb{E}[X \cdot \mathbf{1}_A] = 0$. But this integral is just $\mathbb{Q}[A]$, so $\mathbb{Q}[A]$ is 0 as well. □

The same proof works in an arbitrary space (not just one that can be discretized), but you have to be more careful. We chose to present the proof here because here the definition of X_n intuitively makes sense — we're just discretizing $\mathcal{F}$ using the fact that it's generated by countably many elements.

Next lecture we'll study martingales in continuous time. We'll have to be a bit careful, but the essence is the same as in the discrete case — we'll repeat some of the proofs with some care.

§22 November 21, 2024

Last class, we completed the chapter on discrete-time martingale theory, where we proved many results on martingales and showed why this theory is important — we proved some very nice results (Kolgomorov's 0-1 law, the strong law of large numbers, and so on) using them. But one might wonder what happens in continuous time — if we have a sequence indexed by $\mathbb{R}_{\geq 0}$ instead of $\mathbb{Z}_{\geq 0}$. We can still make sense of the notion of a martingale in that case, and we can still prove results. (You'll see a lot more of this in 18.676 if you take that next semester.) Most results transfer with the exact same statements; but things become more delicate in the continuous case, so we have to be careful.

§22.1 Setup for continuous-time martingales

We always work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We'll always consider a filtration: a collection of sub-σ-algebras $(\mathcal{F}_t)_{t\geq 0}$ of $\mathcal{F}$ (indexed by nonnegative real numbers) such that $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$ for all $0 \leq t_1 \leq t_2$. We'll also have a process $X = (X_t)_{t\geq 0}$ — a collection of random variables indexed by nonnegative real numbers t (living in the same probability space). It'll be adapted to our filtration $(\mathcal{F}_t)_{t\geq 0}$, in the sense that X_t is $\mathcal{F}_t$-measurable for all $t \geq 0$. (This will always be the case.)

Here we'll consider processes where $X_t \in \mathbb{R}^d$ for all $t \geq 0$ (we won't consider more weird processes).

So we can still make sense of filtrations and adapted processes with respect to filtrations. And we can actually still have a notion of stopping times.

Definition 22.1. A stopping time T is a random variable taking values in $[0, \infty]$ such that the event $\{T \leq t\}$ is $\mathcal{F}_t$-measurable for all $t \geq 0$.

This is just like what we had in the discrete-time case — the definition is completely analogous.

So this is going to be our setup for this chapter. We have a notion of a filtration, a notion of an adapted process, and a notion of a stopping time.

§22.2 Differences to discrete case

One way to think about this is that you have a mapping $\Omega \times [0, \infty) \to \mathbb{R}^d$ where you take the pair $(\omega, t) \mapsto X_t(\omega)$. So this process can also be considered a random variable — $[0, \infty)$ is endowed with a Borel σ-algebra and Ω with $\mathcal{F}$, so their product is endowed by the product of $\mathcal{F}$ with the Borel σ-algebra. So another way to view this process is as a random variable on this product space.

In the discrete-time setup, we had a similar setup — our process was a random variable $\Omega \times \mathbb{N} \to \mathbb{R}^d$ given by $(\omega, n) \mapsto X_n(\omega)$. If we endow $\mathbb{N}$ with e.g. the power set σ-algebra (the collection of all possible subsets), then this is indeed a random variable, so it is a measurable function. But in the continuous-time case, this

is not necessarily the case — we have to be a bit careful. This function $X_t(\omega)$ may not always be a random variable — you need some regularity for the function. For example, if you have continuity, then this will indeed be a random variable.

Student Question. *How can you make it so that the measure of the product space is* 1*?*

Answer. Here measurability doesn't have to do with measures. (The definition of a random variable is just another name for a measurable function. This space doesn't have measure 1, but it doesn't matter — all we care about is measurability.)

In general, $(\omega, t) \mapsto X_t(\omega)$ is not necessarily a random variable; that's the first difference to the discrete case.

Another difference is that if you fix $A \subseteq \mathbb{R}$ and consider $T_A = \inf\{t \geq 0 \mid X_t \in A\}$ to be the first time that the process hits A, then this is *not* a stopping time in general. Intuitively, we can write

$$\{T_A \leq t\} = \bigcup_{s \leq t} \{T = s\}.$$

But this is a union over an uncountable set, so this set is not in $\mathcal{F}_t$ in general. So that's one intuitive reason why this is not necessarily a stopping time in our case. This means we need to impose regularity conditions on X, as well as further conditions on the set A (as we will see soon).

§22.3 Measurability

A natural requirement to impose on X so that $X_t(\omega)$ becomes a measurable function is continuity — if we know that for every ω, the function $t \mapsto X_t(\omega)$ is continuous, then this is indeed a measurable function. This is something we're going to prove. More generally, if we assume it's *right-continuous* and has left-limits, then we do have measurability.

Proposition 22.2

Suppose that (for every ω) the function $t \mapsto X_t(\omega)$ is Caedlag, meaning that it is right-continuous and has left limits. Then the function $\Omega \times [0, \infty) \to \mathbb{R}^d$ defined by $(\omega, t) \mapsto X_t(\omega)$ is measurable with respect to the product σ-algebra $\mathcal{F} \otimes \mathcal{B}([0, \infty))$.

By having a left limit, we mean for every fixed t, if we approximate from below, then we get a limit — it might not necessarily be $X_t(\omega)$, but it will be some number.

So under this additional assumption, we have measurability.

Proof. For each ω, we can express $X_t(\omega)$ as

$$X_t(\omega) = \lim_{n \to \infty} \sum_{k=0}^{2^n - 1} \mathbf{1}_{t \in (k2^{-n}, (k+1)2^{-n}]} \cdot X_{k \cdot 2^{-n}}(\omega),$$

for all $t \in [0, 1]$ (we'll restrict ourselves to this interval; the same argument works for $t > 1$). This is by right continuity.

The idea is that for fixed n, we have finitely many linear combinations here, so this function is indeed measurable with respect to our σ-algebra, because we assumed that X is adapted. Explicitly, for fixed n, the function $\Omega \times [0, \infty) \to \mathbb{R}^d$ mapping

$$(\omega, t) \mapsto \sum_{k=0}^{2^n - 1} \mathbf{1}_{t \in (k2^{-n}, (k+1)2^{-n}]} \cdot X_{k \cdot 2^{-n}}(\omega)$$

is $(\mathcal{F} \otimes \mathcal{B}([0,\infty)))$-measurable. So we have a sequence of functions which are measurable, and they converge pointwise to the function that we want. Therefore this function $(\omega, t) \mapsto X_\omega(t)$ also has to be measurable (as it's the pointwise limit of measurable functions).

(With the finite sum, we're using the fact that X is adapted; in the sum we only have to worry about finitely many terms.) □

There are some subtle things going on in continuous time, so we have to be careful; but nothing extremely unusual is going on, and we'll see that exactly the same statements hold. But we have some delicate situations, like here; so we have to be a bit careful while proving things.

From now on, we'll always work with Caedlag processes — ones that are right-continuous and have left limits, so that we know $X_t(\omega)$ is measurable with respect to the product σ-algebra.

Just for a bit of notation:

Notation 22.3. We use $\mathcal{C}(\mathbb{R}_+, E)$ and $\mathcal{D}(\mathbb{R}_+, E)$ denote the spaces of continuous and Caedlag (respectively) functions from $\mathbb{R}_+$ to E. We endow these spaces by the σ-algebras generated by the projections $\pi_t \colon X \mapsto X(t)$.

In our case, E will be $\mathbb{R}^d$ or $\mathbb{R}$. So π_t is a function defined either on the space of continuous or Caedlag functions; it takes in a function X, and maps it to $X(t)$. And we endow the spaces with the smallest σ-algebra that makes these measurable. If we assume our process is Caedlag, then it will be measurable with respect to this σ-algebra.

§22.4 Stopping times and stopped processes

In discrete time, for any stopping time T, we considered all the information of the process stopped at that time. We have a similar notion here — we can still stop the process at a random time, and look at the information up until that time.

Definition 22.4. For a stopping time T, we define

$$\mathcal{F}_T = \{A \in \mathcal{F} \mid A \cap \{T \le t\} \in \mathcal{F}_t \quad \text{for all } t \ge 0\}.$$

Definition 22.5. We define $X_T = X_{T(\omega)}(\omega)$, and we define the stopped process $X^T = (X_t^T)$ given by $X_t^T = X_{T \wedge t}(\omega)$.

These definitions are exactly the same as the discrete case.

That's our setup — the setup we're going to work with in this brief chapter (it will be brief because we've already done most of the work).

Now that we have the setup, let's try to start proving some things.

Proposition 22.6

Let S and T be two stopping times, and let X be a Caedlag adapted process. Then we have the following:

1. $S \wedge T$ is a stopping time.
2. If $S \le T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.
3. We have that $X_T \cdot \mathbf{1}_{T<\infty}$ is $\mathcal{F}_T$-measurable.
4. X^T is adapted.

(1) is immediate, and the proof is the same as the discrete case. (2) is also exactly the same as in the discrete case. But (3) is not immediate — we have to be a bit careful. (Of course, we've shown the analogous property for discrete time.) Given that we have (3), (4) is immediate (we just apply (3) when T is replaced by $\min\{T, t\}$). So the whole point is to prove (3) — that's the nontrivial thing.

(We're not going to prove (1) and (2), because they're exactly the same thing as the discrete case and quite easy.)

Proof of (3). How do we approach this? The main idea is the following observation:

> **Claim 22.7 —** A general random variable Z is $\mathcal{F}_T$-measurable if and only if the random variable $Z \cdot \mathbf{1}_{T \le t}$ is $\mathcal{F}_t$-measurable for all t.

If we know this, then we just need to show that $X_T \mathbf{1}_{T<\infty}$ (which is going to be our Z) has this property; that'll be the second step of the argument. But the first step is to prove this claim.

Proof. If we assume Z is $\mathcal{F}_T$-measurable, then $Z \cdot \mathbf{1}_{T \le t}$ is always $\mathcal{F}_t$-measurable, by the definition of $\mathcal{F}_T$ (and the definition of a stopping time). So one direction is immediate. One way to see this, for instance, is that if you take the event $A = \{Z\mathbf{1}_{T \le t} \le x\}$ (we want to show this is in $\mathcal{F}_t$ for all x, by an equivalent definition of real-valued measurable functions), this is just

$$\{Z \le x\} \cap \{T \le t\}.$$

But by our assumption, $\{Z \le x\} \in \mathcal{F}_T$, and $\{T \le t\} \in \mathcal{F}_t$ (because T is a stopping time). So their intersection is still in $\mathcal{F}_t$ (by the definition of $\mathcal{F}_T$). So this is always in $\mathcal{F}_t$, for every x. This means $Z\mathbf{1}_{T \le t}$ is $\mathcal{F}_t$-measurable.

So one direction is not that hard, and essentially follows from the definitions. What's less trivial is the other direction — we assume $Z \cdot \mathbf{1}_{T \le t}$ is measurable for all t, and want to prove that Z is $\mathcal{F}_T$-measurable.

For this direction, we're going to apply the monotone class theorem, an old theorem from the first lectures.

First, if we assume that $Z = c \cdot \mathbf{1}_A$ is a constant multiple of an indicator function where $A \in \mathcal{F}_T$, then clearly Z is $\mathcal{F}_T$-measurable. (This is obvious because A is just in $\mathcal{F}_T$.) Similarly, if Z is a linear combination of indicator functions in $\mathcal{F}_T$, i.e.,

$$Z = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i} \quad A_i \in \mathcal{F}_T,$$

then Z is still $\mathcal{F}_T$-measurable. Now in the general case, we approximate Z by functions of this form. If you consider

$$Z_n = 2^{-n} \lfloor 2^n Z \rfloor \wedge n,$$

then we have that $Z_n \to Z$ as $n \to \infty$ almost surely. (We've already used this kind of thing.) And in fact, it converges increasingly.

If we assume that $Z \cdot \mathbf{1}_{\{T \le t\}}$ is $\mathcal{F}_t$-measurable for all t, then we have that $Z_n \cdot \mathbf{1}_{\{T \le t\}}$ is also $\mathcal{F}_t$-measurable for all t. And this random variable has the earlier form, so Z_n is $\mathcal{F}_T$-measurable. And then we're done — because Z is the pointwise limit of the Z_n's, so it has to be measurable with respect to the stopped σ-algebra.

(Here it's not clear a priori that $A \in \mathcal{F}_T$. The point is that if $Z = c \cdot \mathbf{1}_A$ for some A and $Z \cdot \mathbf{1}_{T \le t}$ is in $\mathcal{F}_t$ for all t, then we automatically get that $A \in \mathcal{F}_T$.) □

So now it suffices to just prove that $Z = X_T \mathbf{1}_{T \le t}$ is $\mathcal{F}_t$-measurable for all t. So we'll show that $X_T \mathbf{1}_{T \le t}$ is $\mathcal{F}_t$-measurable for all t; if we manage to show this, this will complete the proof.

To show this, we can write

$$X_T \mathbf{1}_{T \le t} = X_T \mathbf{1}_{T<t} + X_t \mathbf{1}_{T=t}$$

(because if we restrict to the event $T = t$, then $X_T = X_t$). We want to show this sum is $\mathcal{F}_t$-measurable. But X is adapted, so X_t is $\mathcal{F}_t$-measurable. And T is a stopping time, so $\mathbf{1}_{\{T=t\}}$ is also $\mathcal{F}_t$-measurable. So it suffices to show that the first term $X_T \mathbf{1}_{T<t}$ is $\mathcal{F}_t$-measurable.

Now let's see how to prove this. The idea to prove such things is to discretize T, so that it takes discrete values — values on a set which is countable. In general we're in the continuous case, so T doesn't necessarily take values on a countable set. The way we discretize it is by considering

$$T_n = 2^{-n} \lceil 2^n T \rceil$$

for all $n \in \mathbb{N}$ — this is a discretized version (it takes countably many values).

Then we have that T_n is a stopping time for all n, taking values in the discrete set $\mathcal{D}_n = \{k \cdot 2^{-n} \mid k \in \mathbb{N}\}$. Why is it a stopping time? If you consider the event $\{T_n \le t\}$, this is the same as the event $\{\lfloor 2^n T \rfloor \le 2^n t\}$, and by the definition of $\lceil \bullet \rceil$, this is the same as saying $\{T \le 2^{-n} \lfloor 2^n t \rfloor\}$. And the latter is measurable with respect to $\mathcal{F}_{2^{-n}\lfloor 2^n t \rfloor} \subseteq \mathcal{F}_t$ (because T is a stopping time). So T_n is indeed a stopping time.

And now the idea is that

$$X_T \mathbf{1}_{T<t} = \lim_{n \to \infty} X_{T_n \wedge t} \mathbf{1}_{T<t}.$$

So if we manage to show that $X_{T_n \wedge t} \mathbf{1}_{T<t}$ is measurable, then we'll be done.

But this random variable is a *finite* sum, because $T_n \wedge t$ takes only finitely many values. So it's the sum of finitely many terms, where each is $\mathcal{F}_t$-measurable; and that means it's $\mathcal{F}_t$-measurable.

To write this more explicitly, for every fixed $n \in \mathbb{N}$, we consider

$$X_{T_n \wedge t} \mathbf{1}_{T \le t} = \sum_{d \in \mathcal{D}_n, d \le t} X_d \mathbf{1}_{T_n = d} + X_t \mathbf{1}_{T_n > t} \mathbf{1}_{T<t}.$$

And now we have a finite sum, where each is $\mathcal{F}_t$-measurable — the event $T_n = d$ is $\mathcal{F}_d$-measurable because T_n is a stopping time, X_d is $\mathcal{F}_d$-measurable and therefore $\mathcal{F}_t$-measurable as well (because $d \le t$), and of course X_t is $\mathcal{F}_t$-measurable.

This completes the proof, because pointwise limits of measurable functions are still measurable functions. So this is indeed $\mathcal{F}_t$-measurable.

(Here we are using the fact that $t \mapsto X_t$ is right-continuous, and T_n approximates T from the right. So here we crucially used the right-continuity of the process.) □

As we can see, it's a bit delicate because T can take values in an uncountable set, which is why we had to introduce this approximation. So things are more delicate, but the statements don't change. If we like intuition, it's just the same as the discrete case; there's just a few more technicalities. In stochastic calculus, unfortunately you'll work with continuous time martingales, so it's nice to have a rough idea about them.

So here the main idea is the approximation with T_n where we can reduce everything to the discrete case, where everything works well (and crucially using right continuity to say that $X_T \mathbf{1}_{T<t} = \lim(X_{T_n \wedge t} \mathbf{1}_{T \le t})$).

§22.5 Hitting times and stopping times

We'll mention two more things before the end of the lecture. One is we mentioned earlier that if we fix a set A and take the first time the process hits that set, this may not necessarily be a stopping time. We gave an intuitive reason, but not a mathematical one; we'll now give a counterexample (of a set and process where the first time the process hits the set is not a stopping time).

Example 22.8

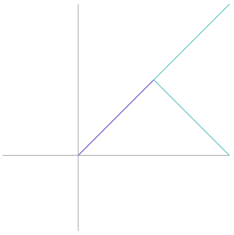
Let J be a random variable with the property that

$$\mathbb{P}[J = 1] = \mathbb{P}[J = -1] = \frac{1}{2}.$$

Consider the process

$$X_t = \begin{cases} t & 0 \le t \le 1 \\ 1 + J(t-1) & t > 1. \end{cases}$$

So we're considering a process that's just t up to 1, but suddenly changes. Graphically, X represents the following picture:



So up until 1 you go up on the diagonal (this is fixed); but after that, you either continue in the same direction or start dropping down, randomly.

Now let $A = (1, 2)$. Then if we consider the event $\{T_A \le 1\}$, this is not in $\mathcal{F}_1$ — because whether the process hits A at the point 1 or not actually depends on J. If $J = -1$, then this event doesn't happen — the process starts dropping, so it never goes above 1. So the event $\{T_1 \le 1\}$ depends on J. But J is not measurable with respect to $\mathcal{F}_1$ — $\mathcal{F}_1$ only depends on the process up to time 1, and does not depend on J. So information about this process related to J is not contained in $\mathcal{F}_1$; so this is not in $\mathcal{F}_1$, which means it's not a stopping time.

So in continuous time, we have to be a bit careful — in order to decide whether the process has hit A or not by time 1, we need to know J, and J is independent of $\mathcal{F}_1$.

Student Question. *Why does this argument not work in the discrete case — why can't we construct something such that at time* 10*, it depends on some random variable J?*

Answer. Because the information $\{T_A \le 1\}$ is encoded in what happens in an infinitestimal neighborhood around 1. But if you had discrete values, this wouldn't make sense — your next time would be at 2, not close to 1. Now the event $T_A \le 1$ depends on the values of the process in an infinitestimal neighborhood around 1, and this doesn't make sense in discrete time. That's why you need to be careful in continuous time — you have to worry about controlling the process in infinitestimal neighborhoods of time, which isn't a thing in the discrete case (where different times are isolated from each other).

Then one might wonder, is it possible for the first hitting time of A to be a stopping time? The answer is yes, when X is continuous (not just right-continuous) and A is closed — in that case it is a stopping time. That's the next thing we're going to do.

Here we'll invoke a compactness argument (though it's not extremely difficult).

Proposition 22.9

Suppose that A is a closed set and X is a continuous adapted process. Then

$$T_A = \inf\{t \geq 0 \mid X_t \in A\}$$

is a stopping time.

So if A is closed and X is a continuous process, then the first hitting time of A is a stopping time in that case.

Proof. The idea is again somehow to restrict things to discrete cases. So how can we do that? For this, we use continuity.

First, we claim that it suffices to show that

$$\{T_A \leq t\} = \left\{ \inf_{s \in \mathbb{Q}, s \leq t} \operatorname{dist}(X_s, A) = 0 \right\}.$$

For this, consider any $s \leq t$, and we consider the distance from X_s to A.

If we fix $s \leq t$, then the event $\operatorname{dist}(X_s, A) = 0$ is $\mathcal{F}_s$-measurable, because X is an adapted process. But this will also be $\mathcal{F}_t$-measurable, because $s \leq t$. So another way to write this is second event is that

$$\limsup_{n \to \infty} \left\{ \bigcup_{s \in \mathbb{Q}, s \leq t} \operatorname{dist}(X_s, A) \leq \frac{1}{n} \right\}.$$

If there are infinitely many n such that we can find rational $s \leq t$ with distance at most $1/n$, then we have this. And if we can show $\{T_A \leq t\}$ is this, then we're done — for every fixed n, this thing is a countable union of events (because $\mathbb{Q}$ is countable). So for every fixed s and n, the event $\{\operatorname{dist}(X_s, A) \leq 1/n\}$ is in $\mathcal{F}_s \subseteq \mathcal{F}_t$. ANd we're taking a countable union of events in $\mathcal{F}_t$, so it'll be in $\mathcal{F}_t$. ANd the limsup of events in $\mathcal{F}_t$ is also in $\mathcal{F}_t$. So if we can show $\{T_A \leq t\}$ is equal to this other event, then we're done.

And in order to show this, we will use continuity, and the fact that A is closed. So now our goal is to show

$$\{T_A \leq t\} = \limsup_{n \to \infty} \bigcup_{s \in \mathbb{Q}, s \leq t} \{\operatorname{dist}(X_s, A) < 1/n\}.$$

First we'll show the forwards direction — that if $T_A \leq t$, then the other thing holds. Suppose that $T_A = s \leq t$. Then by the definition of the infimum, there exists a sequence (s_n) of times such that $s_n \downarrow s$ as $n \to \infty$ (by the definition of the infimum) and $X_{s_n} \in A$ for all n. This follows by the definition of the infimum; we haven't used any kind of continuity so far.

Now we use continuity — by the continuity of X, and since A is closed, we have that $X_{s_n} \to X_s$ as $n \to \infty$ (because $s_n \to s$) and $X_s \in A$ (because if you have a closed set and a sequence inside the set which converges somewhere, then the point it converges to has to be in the set). So in particular, this implies $X_{T_A} \in A$.

And you can also take any sequence of rationals (q_n) with $q_n \uparrow T_A$ (we can always do this, because the rationals are dense in $\mathbb{R}$). But since X is continuous, this means $\operatorname{dist}(X_{q_n}, A) \to \operatorname{dist}(X_{T_A}, A)$ as $n \to \infty$. And the latter is 0, because we've shown that $X_{T_A} \in A$.

So what we've done is found a sequence of rational numbers $q_n \leq t$ such that $\operatorname{dist}(X_{q_n}, A) \to 0$. That means $\inf_{s \in \mathbb{Q}, s \leq t} \operatorname{dist}(X_s, A) = 0$, which is precisely what we wanted.

So we proved that if $T_A \leq t$, then this holds. Now to complete the proof, we'll assume that

$$\inf_{s \in \mathbb{Q}, s \leq t} \operatorname{dist}(X_s, A) = 0,$$

and prove that $T_A \le t$.

By the definition of the infimum, there exists a sequence $s_n \in \mathbb{Q}$ with $s_n \le t$ such that $\operatorname{dist}(X_{s_n}, A) \to 0$ as $n \to \infty$. There's a theorem (Bolzano–Weirstrass) that if you have a sequence of real numbers in a bounded interval, then they have a convergent subsequence. Here we have a sequence s_n which is bounded by t, so it'll have a subsequence which converges. So there exists a subsequence s_{k_n} and some time $s \le t$ such that $s_{k_n} \to s$ as $n \to \infty$. By the continuity of X, this means $\operatorname{dist}(X_{s_{k_n}}, A) \to \operatorname{dist}(X_s, A)$. And the left-hand side converges to 0, so $\operatorname{dist}(X_s, A) = 0$. And this means $X_s \in A$ (because A is closed). So $X_s \in A$, which means $T_A \le s \le t$. This gives the other direction. So we have our equality, and by the argument discussed earlier, this is $\mathcal{F}_t$-measurable. □

§23 November 26, 2024

§23.1 Hitting times and an enlarged filtration

Last lecture, we started the next chapter on continuous-time martingales. Now our process is indexed by $\mathbb{R}_{\ge 0}$, instead of the set of positive integers. We mentioned that more or less, most of the concepts and properties are almost identical. But to prove exactly the same statements, we have to be a bit careful — we saw a few counterexamples last lecture. For instance, if you take an open set and consider the first time the process enters it, this is not necessarily a stopping time. But if you have a closed set and a continuous process, then the first time it hits the set is indeed a stopping time.

We also saw we need to use some sort of regularization (approximating by the discrete case) to prove things like optional stopping.

Sometimes we'll also have to work with a slightly better σ-algebra — sometimes we'll need regularity on the filtration.

> **Definition 23.1.** Suppose we have a filtration $(\mathcal{F}_t)_{t\ge 0}$. Then for each $t \in \mathbb{R}_{\ge 0}$, we define $\mathcal{F}_{t^+} = \bigcap_{s>t} \mathcal{F}_s$. If $\mathcal{F}_{t^+} = \mathcal{F}_t$ for all t, then we call $(\mathcal{F}_t)$ right-continuous.

In general, $\mathcal{F}_s$ is a new σ-algebra which always contains $\mathcal{F}_t$, but they're not necessarily equal. But if they are equal (always), we say the filtration is right-continuous.

So we've slightly enlarged the filtration by going from $(\mathcal{F}_t)$ to $(\mathcal{F}_{t^+})$; as we'll see, this will be quite useful. For instance, under this filtration, if you consider the first time a continuous process hits an open set, this is in fact a stopping time.

> **Proposition 23.2**
> Let $A \subseteq \mathbb{R}$ be open, and let X be a continuous process. If we define
> $$T_A = \inf\{t \ge 0 \mid X_t \in A\},$$
> then T_A is a stopping time with respect to $(\mathcal{F}_{t^+})$.

Here X is a continuous and adapted process with respect to the original filtration $(\mathcal{F}_t)$; and if we slightly enlarge the filtration (allowing a bit more 'information'), then this becomes a stopping time.

Proof. The first thing to show is that $\{T_A < t\} \in \mathcal{F}_t$ for all t. This is what we'll show, and then we'll deduce our result.

For this, we'll use the continuity of X and the fact that A is open. By combining these, we have that $\{T_A < t\} = \bigcup_{q\in\mathbb{Q}, q\le t}\{X_q \in A\}$ (this follows by continuity, because if at some time you're in A, then in an

interval around that time you'll still be in A because A is open). But $\{X_q \in A\}$ is always in $\mathcal{F}_q$, because X is an adapted process. And since this is a countable union, we obtain that the entire union is in $\mathcal{F}_t$ (because $\mathcal{F}_q \subseteq \mathcal{F}_t$ for all $q \le t$). So the event that the first hitting time is *strictly* less than t is always in $\mathcal{F}_t$.

But now we can write $\{T_A \le t\} = \bigcup_{n\in\mathbb{N}} \{T_A < t + \frac{1}{n}\}$. And we have $\{T_A < t + \frac{1}{n}\} \in \mathcal{F}_{t+1/n}$ for all n, and the intersection over all n of these σ-algebras is just $\mathcal{F}_{t^+}$. So this belongs to $\mathcal{F}_{t^+}$, which is what we wanted. □

This says if we slightly enlarge our filtration, then the first hitting time of an open set is a stopping time as well — we just have to add a bit of information in order to keep track. (If A is a closed set, we don't need this extra information — we don't need to enlarge the filtration, as shown last lecture.)

This enlarged filtration will be useful (as we'll see later).

Student Question. *Where did we use that A is open?*

Answer. When saying that $\{T_A \le t\} = \bigcup_q \{X_q \in A\}$. For the forwards direction, suppose $T_A \le t$. This implies we can find some $s < t$ such that $X_s \in A$. And since A is open, this means we can find $\varepsilon > 0$ such that the ball centered at X_s with radius ε is in A. But now X is continuous, so we can find $\delta > 0$ such that $X_u \in \mathbb{B}(X_s, \varepsilon) \subseteq A$ for all $u \in (s-\delta, s+\delta)$. And since $\mathbb{Q}$ is dense, we can always find a rational q in this interval (and by taking δ sufficiently small, we can ensure that $q \in (0, t)$).

That's the harder direction; the other direction is obvious, because if $X_q \in A$ for some $q < t$, then $T_A \in \mathcal{F}_q \subseteq \mathcal{F}_t$.

Student Question. *Are there assumptions under which we can generalize this to any Borel set?*

Answer. No — continuity isn't enough, and there aren't really stronger assumptions that make sense.

That's more or less the introduction to continuous-time martingales. They're slightly different, but the main ideas are the same as the discrete case. What's left to prove is some results — for example, we have analogous results to the martingale convergence theorem, Doob's maximal inequality, L^p convergence, and so on. But because of continuous time we have some slight technicalities.

§23.2 Almost sure martingale convergence theorem

Recall that to prove the martingale convergence theorem in the discrete case, we had to prove Doob's upcrossing inequality. Here it's not *a priori* obvious how to define upcrossings, so that's what we'll do now.

We'll have the following lemma, which is the analogous result about convergence for functions defined on $\mathbb{R}_{\ge 0}$ rather than $\mathbb{N}$ (of the statement that a sequence with finitely many upcrossings of every interval converges).

Lemma 23.3

Let $f : \mathbb{Q}_+ \to \mathbb{R}$ be a function. Suppose that for all $a < b$ (for $a, b \in \mathbb{Q}_+$) and all bounded subsets $I \subseteq \mathbb{Q}_+$, the function f is bounded on I and $N([a,b], I, f)$ is finite. Then for all $t \ge 0$, we have that $\lim_{s\downarrow t} f(s)$ and $\lim_{s\uparrow t} f(s)$ exist and are finite.

As before, we use $N([a,b], I, f)$ to denote the number of upcrossings of $[a,b]$ by f on the time-interval I. To be precise, we define

$$N([a,b], I, f) = \sup\{n \mid \text{exist } s_1 < t_1 < s_2 < \cdots < s_n \text{ in } I \text{ with } f(s_i) < a,\ f(t_i) > b\}.$$

So the assumptions here are quite similar to the ones we had in the discrete case.

We use $\lim_{s\downarrow t}$ and $\lim_{s\uparrow t}$ to denote the left and right limits at t, respectively. And we have this result, which is somehow a version of the result from the previous lecture. Here we have a function defined on the set

of positive rationals, which is bounded on bounded subsets, such that the number of upcrossings on any interval (when f is restricted to any bounded subset of the rationals). Then the left and right limits exist, and are finite; but they're not necessarily the same. Here we have to be careful — we're *not* assuming that the *total* number of upcrossings of any interval is finite. We're only assuming this when resticted to any bounded interval. So we're assuming something weaker.

Of course, we'll use this lemma where f is our martingale, restricted to the positive rationals.

We won't prove this; the proof is elementary analysis (Prof. Kavvadias suggests we try to do it ourselves, and we can discuss it in office hours). So for the moment we'll take this for granted; and we'll use it to prove the almost sure martingale convergence theorem, or rather, its continuous version — the main result for today's lecture.

Theorem 23.4 (Almost sure martingale convergence)

Let X be a cadlag martingale, and assume that X is bounded in L^1. Then $X_t \to X_\infty$ almost surely as $t \to \infty$, for some random variable $X_\infty \in L^1$ which is $\mathcal{F}_\infty$-measurable (where $\mathcal{F}_\infty = \sigma(\bigcup_t \mathcal{F}_t)$).

So we have exactly the same result as in the discrete case; but a bit of care is needed, because we need to prove convergence as $t \to \infty$. So it has to converge simultaneously over *every* subsequence tending to ∞ (whereas with discrete time, that sequence was fixed — it was just the integers). That makes things slightly more technical here.

Proof. Set $I_M = \mathbb{Q}_+ \cap [0, M]$. Here the first thing to do is to show that the number of upcrossings of X restricted to I_M of any interval is finite (as in the setup of the above lemma).

The idea is to somehow invoke the discrete case — we need to construct a discrete martingale in order to apply Doob's inequality. Let's see how to do this carefully. Suppose we fix any $a < b$. Then we have that $N([a, b], I_M, f)$ (the total number of upcrossings of that interval by f restricted to I_M) is given by

$$\sup_{J \subseteq I_M \text{ finite}} (N([a, b], J, X))$$

(this makes sense because we can write I_M as an increasing union of finite sets).

Now suppose we fix such a set J, and let its elements be $a_1 < a_2 < \cdots < a_n$. Then we have that $(X_{a_i})_{i \in [n]}$ is a discrete-time martingale. And for this discrete-time martingale, we can apply Doob's upcrossing inequality. (That's what we meant by saying the results are the same, but some care is needed to deal with the continuous time.)

So Doob's upcrossing inequality implies that

$$(b - a)\mathbb{E}[N([a, b], J, X)] \leq \mathbb{E}[(X_{a_n} - a)^-] \leq \mathbb{E}[(X_k - a)^-]$$

where k is such that $I_M \subseteq [0, k]$. (The first step is Doob's upcrossing inequality.) Now we have a bound that's uniform in J, and $N([a, b], I_M, X)$ is the supremum over all J, so we get

$$\mathbb{E}[N([a, b], I_M, X)] \leq (b - a)^{-1}\mathbb{E}[(X_k - a)^-].$$

And this expectation is finite, because we assumed we have a L^1-bounded martingale. (We used L^1-boundedness in exactly the same way in the discrete case.)

In fact, we can actually get something stronger — we can get a bound of $(b-a)^{-1} \sup_{t \geq 0} \mathbb{E}[|X_t|]$. (We want something that's uniform in M as well.) (We probably get an extra term of $|a|$.)

Now we take $M \to \infty$. Then we'll obtain an upper bound on the expected total number of upcrossings of X when restricted to $\mathbb{Q}_+$ (because as $N \to \infty$, I_M increases to $\mathbb{Q}_+$). This means

$$\mathbb{E}[N([a, b], \mathbb{Q}_+, X)] < \infty$$

for all a and b. (For example, this is by the monotone convergence theorem.) In particular, this implies that if we define $\Omega_0 = \bigcap_{a<b\in\mathbb{Q}}\{N([a,b],\mathbb{Q}_+,X)<\infty\}$, then $\mathbb{P}[\Omega_0]=1$.

As with discrete-time martingales, we'll prove that when Ω_0 occurs, X_t indeed converges. So far, what we've obtained is we have the required convergence of X_t, but when t is restricted to $\mathbb{Q}^+$ — we don't have convergence over arbitrary sequences. We have convergence when t is restricted to $\mathbb{Q}^+$, but that's not enough; we somehow need to upgrade this.

So far, on Ω_0 we have that X_q converges as $q\to\infty$, where $q\in\mathbb{Q}_+$; let's suppose it converges to X_∞. We'd like to upgarde this convergence so that q is allowed to be any positive real number. So we want to prove $X_t\to X_\infty$ as $t\to\infty$ along any subsequence.

For this, we just apply the definition — we fix $\varepsilon>0$. Since $X_q\to X_\infty$ along the positive rationals, we can find a positive rational q_0 such that $|X_q-X_\infty|<\frac{\varepsilon}{2}$ for every $q\in\mathbb{Q}_+$ with $q>q_0$. (This is just the definition of convergence of sequences.) We want to say a similar bound is true for all $t>q_0$ (without necessarily having $t\in\mathbb{Q}_+$).

This is where we'll use the right continuity of X. If we take any $t>q_0$, then by the right continuity of X, there exists $q>t$ such that $|X_t-X_q|<\frac{\varepsilon}{2}$ (X is right-continuous, so if you approach t from the right by rational numbers, then the limit exists and equals X_t). By the triangle inequality, this means

$$|X_t-X_\infty|\le|X_t-X_q|+|X_q-X_\infty|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}.$$

And this is true for every $t>q_0$. So what we've done is fixed ε, and we've found a positive number q_0 such that $|X_t-X_\infty|<\varepsilon$ for all $t>q_0$. But this is just the definition of convergence. So this means that indeed $X_t\to X_\infty$ almost surely,. And of course, we also want $X_\infty\in L^1$, but this is clear by Fatou's lemma (we have $\mathbb{E}[|X_\infty|]\le\sup\mathbb{E}[|X_n|]$, and we assumed our martingale is L^1-bounded, so $\mathbb{E}[|X_n|]$ is bounded uniformly). □

The main idea is the same as the discrete case, but we have some technicalities, and we have to be a bit careful. To justify convergence as $q\to\infty$ for q a positive rational, we've used that lemma. (The lemma might actually not be needed — you can directly prove convergence arguing in the same way as in the discrete case.)

Student Question. *How do we know X_∞ is $\mathcal{F}_\infty$-measurable?*

Answer. X_q is $\mathcal{F}_q$-measurable and therefore $\mathcal{F}_\infty$-measurable for all q, so the limit is measurable as well.

This might be slightly boring because we have the same result, but there are some subtleties here that are not that obvious.

§23.3 Doob's maximal inequality

The next thing one could ask is whether we have the nice inequalities that we had in the discrete time case — e.g. Doob's maximal inequalities, which were quite useful in proving L^p convergence results. We'll now prove the same results in our setting, starting with a version of Doob's maximal inequality for continuous time martingales.

Theorem 23.5 (Doob's maximal inequality)
Let $X=(X_t)_{t\ge0}$ be a cadlag martingale, and let $X_t^*=\sup_{0\le s\le t}|X_s|$. Then

$$\lambda\mathbb{P}[X_t^*\ge\lambda]\le\mathbb{E}[|X_t|]\quad\text{for all } t,\lambda>0.$$

This is exactly the same thing as we had in the discrete case. Recall Doob's maximal inequality is stronger than just applying Markov's inequality — if you applied Markov's inequality then X_t would be replaced with X_t^*. But in general X_t^* is larger than X_t; so the point is that we get a better bound in the case of martingales.

Proof. Again, there are some subtleties and we can't go directly from the discrete case; we have to do something similar to what we did for the martingale convergence theorem.

Here, the main idea is to write

$$\sup_{0\le s\le t} |X_s| = \sup_{s=t \text{ or } s\in\mathbb{Q}_+\cap[0,t)} |X_s|.$$

So we can always write the supremum over a discrete set. (This is because X is cadlag — so here we've crucially used the cadlag property of X.) This means it suffices to show the same inequality where X_t^* is replaced by a supremum over this discrete set. And the main idea is writing this discrete set as an increasing union of finite sets and applying the discrete version of Doob's maximal inequality to each of those.

More carefully, we consider a sequence of *finite* sets $(J_n) \subseteq \{t\}\cup([0,t)\cap\mathbb{Q}_+)$ such that $J_n \uparrow \{t\}\cup([0,t)\cap\mathbb{Q}_+)$ (meaning the set on the right-hand side is a countable union of J_n over all n). Now if we restrict X onto the set J_n, it is a martingale — if we let $J_n = \{i_1^{(n)},\dots,i_{k_n}^{(n)}\}$, the point is that if we restrict X to this finite set J_n, then this is a discrete-time martingale, so we can apply Doob's maximal inequality to this discrete-time martingale (and then take $n\to\infty$).

Here we have that

$$\mathbb{P}[\sup_{s\in\{t\}\cup[0,t)\cap\mathbb{Q}_+} |X_s|\ge\lambda] = \lim_{n\to\infty}\mathbb{P}[\sup_{s\in J_n}|X_t|\ge\lambda]$$

(because J_n increases to the set on the left; this equality may also make use of right continuity). What remains to prove is that the desired bound is true for this probability on the right, for all fixed n (if we can prove this, then we'll be done by taking $n\to\infty$).

Recall that $(X_t)_{t\in J_n}$ is a discrete-time martingale. So we can apply Doob's maximal inequality to this martingale, to deduce that

$$\lambda\mathbb{P}[\sup_{s\in J_n}|X_s|\ge\lambda]\le\mathbb{E}[|X_t|].$$

(When we define J_n, we should require that $t\in J_n$ for this to hold.) And this is true for all n, so we can just let $n\to\infty$, and then we obtain the desired result. □

Again, the main idea is that you want to prove a property for a process in continuous time, so you want to discretize time and consider the process restricted to this discrete set. We've proved many things for such a process. And then we take that discrete set larger and larger. But in order to make the step where we go between discrete and continuous time, we need the regularity condition that X is cadlag; so this is what we need those further properties for.

§23.4 Some more results

Now we'll write down some more results; we won't prove them, because hopefully we've gotten a feel for how to prove things.

Recall that in the discrete case, we had Doob's L^p inequality; we can get a continuous time version by arguing in the exact same way as in the maximal inequality (except applying the L^p inequality instead of the maximal inequality).

Theorem 23.6 (Doob's L^p inequality)

Let $X = (X_t)_{t \geq 0}$ be a cadlag martingale. Then we have

$$\|X_t^*\|_p \leq \frac{p}{p-1} \|X_t\|_p$$

for all $t \geq 0$ and $p > 1$.

So this is the L^p Doob's inequality, just like in the discrete setting; and the proof is the same as with the maximal one.

Again, just as with discrete time, we also have the L^p martingale convergence theorem, which is translated almost identically; we'll state it without proving it, because the reasoning is the exact same as with the previous two theorems. (You can try to do it as an exercise.)

Theorem 23.7 (L^p martingale convergence theorem)

Let $X = (X_t)$ be a cadlag martingale, and let $p > 1$. Then the following statements are equivalent:

(1) X is bounded in L^p, i.e., $\sup_t \|X_t\|_p < \infty$.

(2) X converges almost surely and in L^p to some random variable X_∞.

(3) There exists a random variable $Z \in L^p$ such that $X_t = \mathbb{E}[Z \mid \mathcal{F}_t]$ almost surely (for all $t \geq 0$).

Again, this is exactly the same thing; as in the discrete case, this characterizes the L^p convergence of continuous time martingales. (The proof is again discretizing time and applying the discrete version of the theorem, as with the previous proofs.)

We also characterized L^1 convergence for discrete time martingales; we said a martingale converges almost surely and in L^1 if and only if it's uniformly integrable, and exactly the same thing holds here (we just combine the discrete time version of the theorem with the method applied in the proof of the almost sure martingale convergence theorem, discretizing time in an appropriate way).

Theorem 23.8 (Uniformly integrable martingale convergence theorem)

Let $X = (X_t)$ be a cadlag martingale. Then X is uniformly integrable if and only if X converges almost surely and in L^1 to some $X_\infty \in L^1$.

So this is a characterization of L^1 convergence of continuous time martinales; the proof again combines the discrete case with the technique introduced in the proof of the almost sure martingale convergence theorem.

Student Question. *In the discrete version, we also had a third statement that there exists Z with $\mathbb{E}[Z \mid \mathcal{F}_t] = X_t$. Is that also true in this case?*

Answer. Here it is more delicate — there we used backwards martingales, and here it's not clear how to define backwards martingales.

§23.5 Optional stopping theorem for uniformly integrable martingales

We also proved one last thing in discrete time — the optional stopping theorem for uniformly integrable discrete time martingales. Here we have the same statement; but we'll prove it, because it requires approximating stopping times.

Theorem 23.9 (Optional stopping theorem)

Let X be a cadlag and uniformly integrable martingale. Then for all stopping times S and T (possibly infinite) with $S \leq T$, we have

$$\mathbb{E}[X_T \mid \mathcal{F}_S] = X_S \quad \text{almost surely.}$$

Here we don't have to assume T is finite almost surely — this can hold for infinite stopping times (where we set $X_T = X_\infty$, where X_∞ is the almost sure limit that we know exists).

The proof is similar to the discrete case, but we need to discretize time. In the discrete case, all stopping times took values in $\mathbb{N}$; but here our stopping times might take any positive real values. So we have to be careful when discretizing the stopping times.

Proof. Let $A \in \mathcal{F}_S$. By the uniqueness of conditional expectation, what we need to show is that

$$\mathbb{E}[X_T \cdot \mathbf{1}_A] = \mathbb{E}[X_S \cdot \mathbf{1}_A].$$

(If we manage to show this, this means X_S satisfies the properties of the conditional expectation $\mathbb{E}[X_T \mid \mathcal{F}_S]$, and then by the uniqueness of the conditional expectation, we will have the required equality.) For this, we want to discretize time and apply the corresponding property for the discrete case.

To discretize these times, we define

$$T_n = 2^{-n} \lceil 2^n T \rceil \quad \text{and} \quad S_n = 2^{-n} \lceil 2^n S \rceil .$$

The advantage of introducing these new stopping times T_n and S_n is that we have $T_n \downarrow T$ (i.e., T_n decreases to T) and $S_n \downarrow S$ as $n \to \infty$. So we can apply the right continuity of X to get that $X_{S_n} \to X_S$ almost surely as $n \to \infty$, and $X_{T_n} \to X_T$ almost surely.

And now we can apply the discrete time optional stopping theorem for X_{T_n} and X_{S_n}. By the discrete time optional stopping theorem, we have that

$$X_{T_n} = \mathbb{E}[X_\infty \mid \mathcal{F}_{T_n}]$$

almost surely (for all n) by the discrete-time optional stopping theorem — we're applying the same theorem where T is replaced by ∞ and S by T_n. And since X_{T_n} has this form, this implies X_{T_n} is uniformly integrable; so it's indeed a uniformly integrable martingale. And so it converges almost surely and in L^1. We know it converges almost surely to X_T, so this means it also converges in L^1 to T. So we get $X_{T_n} \to X_T$ in L^1.

Also, we have that $\mathbb{E}[X_{T_n} \mid \mathcal{F}_{S_n}] = X_{S_n}$ almost surely — this is by the discrete-time optional stopping theorem. This implies in particular that

$$\mathbb{E}[X_{T_n} \cdot \mathbf{1}_A] = \mathbb{E}[X_{S_n} \cdot \mathbf{1}_A].$$

But $X_{S_n} \to X_S$ in L^1 (by the same argument as above) and $X_{T_n} \to X_T$ in L^1. So we can take limits as $n \to \infty$; the first term converges to $\mathbb{E}[X_T \mathbf{1}_A]$ and the second to $\mathbb{E}[X_S \mathbf{1}_A]$, and since the sequences are the same, their limits are too.

So we have the desired equality, and by uniqueness of conditional expectations we get the theorem. (The main idea was again to discretize appropriately and then apply the discrete-time results. But discretizing appropriately is not trivial; here we used the discretization with T_n and S_n so that we have convergence from the right, crucially using the cadlag property.) □

Those were our results about continuous time martingales. The important thing is to understand discrete time; then the ideas are the same, but with some technicalities.

Student Question. *How do we know $X_{T_n} = \mathbb{E}[X_\infty \mid \mathcal{F}_{T_n}]$?*

Answer. This is the discrete-time optional stopping theorem — we know $X_{T_n} \to X_\infty$ by the almost sure martingale convergence theorem. (We started with X being a uniformly integrable martingale, so X_∞ exists and is in L^1.)

Next lecture we'll start with Kolgomorov's continuity criteria, which is quite important. And then we'll have an introduction to large deviation theory, though we won't have time to do too many things. And Prof. Kavvadias will also upload some problems about martingales that he hasn't yet; we can consider this as a 6th problem set, but it won't be graded, just for our own practice. There's also an announcement about the exam, which is 12/16 at 1:30–4:30 in 32-155. The material covered in lectures is on the exam; we have to know statements and proofs of the theorems (the proofs may be asked for), and we should be able to solve the problems on the problem sets. There are some logistics as usual; there is a deadline for sending a deadline for accommodations.

§24 December 3, 2024

Last lecture, we completed the chapter on martingales in general. We showed that the same results hold when instead of a discrete-time martingale, we have one indexed by nonnegative real numbers. The main ideas are the same, but there are some nontrivial technicalities.

§24.1 Kolgomorov's continuity criterion

This lecture, we'll first state and prove Kolgomorov's continuity criterion. This will be useful in the last lecture (or next one), where we'll rigorously construct Brownian motion. The main idea of Kolgomorov's continuity criterion is we have two steps in order to construct a process indexed by $[0,1]$. The idea is to first construct the process at a countable set of points, and then use Kolgomorov's continuity criterion to extend it to all of $[0,1]$.

Theorem 24.1 (Kolgomorov's continuity criterion)

Let $\mathcal{D}_n = \{k \cdot 2^{-n} \mid 0 \le k < 2^n\}$ be the set of dyadic numbers of order n, and let $\mathcal{D} = \bigcup_n \mathcal{D}_n$ be the set of dyadic numbers in $[0,1]$. Let $X = (X_t)_{t\in\mathcal{D}}$ be a stochastic process taking real values.

Suppose that there exist $p, \varepsilon > 0$ such that $\mathbb{E}[|X_t - X_s|^p] \le c \cdot |t-s|^{1+\varepsilon}$ for all $s,t \in \mathcal{D}$ (for some finite constant c). Then for any $a \in (0, \frac{\varepsilon}{p})$, we have that $(X_t)_{t\in\mathcal{D}}$ is a-Hölder continuous, i.e.,

$$\sup_{s\neq t\in\mathcal{D}} \frac{|X_t - X_s|}{|s-t|^a} < \infty \quad \text{almost surely.}$$

So we start with a process X indexed by a discrete set. The idea is to impose certain conditions on this process so that we can define it continuously over the entire interval. And this says that under the given condition, when we restrict to dyadic rationals, the process is Hölder continuous. This means it extends to the entire interval — if we fix any $t \in [0,1]$, then we can approximate t by a sequence $(t_n) \subseteq \mathcal{D}$ (because $\mathcal{D}$ is dense). And if we consider (X_{t_n}), then letting the above supremum be K_a, we have

$$|X_{t_n} - X_{t_m}| \le K_a \, |t_n - t_m|^a .$$

This tends to 0 as $n, m \to \infty$, so X_{t_n} is Cauchy, which means it has to converge; and then we can just set $X_t = \lim_{n\to\infty} X_{t_n}$. So this means Kolgomorov's continuity criterion is a very nice way to define a process on the entire unit interval, or the entirety of $\mathbb{R}$ as well. There's two steps — to define it on the set of dyadic

rationals, and then once we have the condition on $\mathbb{E}[|X_t - X_s|^p]$, this guarantees Hölder continuity, which lets us extend it to the entire unit interval.

The proof is not extremely difficult, but it is slightly technical.

Proof. The first step of the proof is to bound these distances when s and t are in $\mathcal{D}_n$ — that's the main idea — and for that, we'll use our hypothesis combined with Markov's inequality. By Markov's inequality, we have that

$$\mathbb{P}[\left|X_k \cdot 2^{-n} - X_{(k+1)\cdot 2^{-n}}\right| \geq 2^{-na}] \leq C \cdot 2^{nap} \cdot 2^{-n-n\varepsilon}$$

(this is Markov's inequality on the pth powers, using the bound on the expectation). Now we can apply a union bound over all $0 \leq k < 2^n$ — by a union bound over k, we get that

$$\mathbb{P}\left[\max_{0\leq k<2^n} \left|X_{k\cdot 2^{-n}} - X_{(k+1)\cdot 2^{-n}}\right| \geq 2^{-na}\right] \leq C \cdot 2^{-n(\varepsilon - pa)}$$

(this is a union bound — we just sum over all k, because the probability the maximum is at least 2^{-na} is at most the sum over all k that the probability the difference for that k is at least 2^{-na}). And so this bounds the distances between dyadic rationals which are consecutive and of order n.

By summing over all n (applying Borel–Cantelli), we obtain that this is true for all sufficiently large n, almost surely — by the first Borel–Cantelli lemma, we have that

$$\max_{0\leq k<2^n} \left|X_{k\cdot 2^{-n}} - X_{(k+1)\cdot 2^{-n}}\right| \leq 2^{-na} \quad \text{for all sufficiently large } n, \text{ almost surely.}$$

(This is because the exponent of 2 is negative, so when we sum over all n, we get something finite; this means almost surely, this event only happens finitely many times.)

What does this mean? It means there exists a finite random variable M with the property that

$$\sup_{n\in\mathbb{N}} \max_{0\leq k<2^n} \frac{\left|X_{k\cdot 2^{-n}} - X_{(k+1)\cdot 2^{-n}}\right|}{2^{-na}} \leq M$$

(because our inequality was true for all sufficiently large N, and for small N, this quantity is always finite — there are finitely many terms, and they're all finite random variables). Now we've bounded the distance for consecutive dyadic rationals of the same order. (If we just define M as the left-hand side, this is a random variable, and this is finite almost surely — we showed it's finite if you take the supremum over all $n \geq n_0$, and the supremum for $1 \leq n \leq n_0$ has to be finite because all terms are finite random variables, and there are finitely many.)

So far, we've proved the Hölder continuity condition for consecutive dyadic rationals of the same order; we now need to generalize it to arbitrary dyadic rationals. For this, we'll use the dyadic representation of numbers — fix any $t, s \in \mathcal{D}$ (assume without loss of generality that $s < t$) and let $r \in \mathbb{N}$ be the unique integer with the property that $2^{-(r+1)} < t - s \leq 2^{-r}$. In other words, this means there exists some integer $k \in \mathbb{N}$ with the property that $s < k \cdot 2^{-(r+1)} \leq t$ — so between s and t, we can find a dyadic rational of order r (because if you fix r and let k run between 0 and 2^{r+1}, this partitions the unit interval, and s and t can't be in the same part; so for some k, we will have an inequality like this). Set $x = k \cdot 2^{-(r+1)}$ to that number. Then we have that $t - x \in (0, 2^{-r})$ (by the upper bound on $t - s$). This means in the dyadic representation of $t - x$, all the dyadic numbers of order at most r don't appear — because it's strictly less than 2^{-r}. In particular, we have that

$$t - x = \sum_{n\geq r+1} \frac{x_n}{2^n}$$

for some $x_n \in \{0, 1\}$ (any number in $[0, 1]$ can be written as an infinite sum of the form $\sum \frac{x_n}{2^n}$ — this is the dyadic representation of a number, which we discussed during the first lectures; and by the choice of r this number is strictly less than 2^{-r}, which means n can only be larger than $r + 1$). Similarly, we can write

$$x - s = \sum_{j\geq r+1} \frac{y_j}{2^j}$$

for some $y_j \in \{0,1\}$. So here we have these nice representations. The point is, what does this imply? This implies that if we take the interval $[s,t)$, we can write it as a disjoint union of dyadic intervals of length 2^{-n} for $n \geq r+1$ — because we can first break it as

$$[s,t) = [s,x) \sqcup [x,t)$$

as the union of two disjoint intervals, and then since $t-x$ and $x-s$ have the form mentioned, if you partition them into dyadic intervals, all the intervals that appear will have length at most $2^{-(r+1)}$. So we can express this interval as a disjoint union of such dyadic intervals.

This means if we apply the triangle inequality, we can bound

$$|X_t - X_s| \leq \sum_{d,n} |X_d - X_{d+2^{-n}}|$$

where d and n simultaneously range over the endpoints of the dyadic intervals and their length. (Here we've just written $[s,t) = [s,x_1) \cup [x_1,x_2) \cup \cdots$, and w're just saying $|X_t - X_s| \leq |X_{x_1} - X_s| + |X_{x_2} - X_{x_1}| + \cdots$. Here the x_is are consecutive dyadic rationals of some order $n \geq r+1$.)

And we can bound each of these terms using our nice bound from before — we can sum over *all* $n \geq r+1$ (instead of just the relevant ones), and we get at most

$$2 \cdot \sum_{n \geq r+1} M \cdot 2^{-na}$$

(where M is the random variable defined from earlier). But the quantity on the right-hand side is just a geometric series, so it has an explicit form; and that explicit form is

$$2M \cdot \frac{2^{-(r+1)a}}{1-2^{-a}}.$$

But recall that by our choice of r, we have $|s-t| \geq 2^{-(r+1)}$, which means this is at most

$$\frac{2M}{1-2^{-a}} \cdot |t-s|^a.$$

And this is exactly what we wanted to prove — $|X_t - X_s|^a$ is at most a (random) constant times $|t-s|^a$, which is exactly what we wanted, with

$$K_a = \frac{2M}{1-2^{-a}}.$$ □

The importance of this criterion is in constructing processes which are defined in $[0,1]$ and are continuous — the first step is to define them on a discrete set, prove that they satisfy the moment condition on $\mathbb{E}[|X_s - X_t|^p]$, and deduce the Hölder continuity when restricted to the dyadic rationals. And then we can extend the process on the entire interval. We'll do this when defining Brownian motion — we'll first construct it on dyadic rationals, and then extend it continuously by proving it satisfies our condition.

Remark 24.2. You don't have to know the details of this proof; what's important is how to use this theorem. Usually we use it to extend processes continuously in an uncountable set, like $[0,1]$ or $\mathbb{R}$.

Student Question. *The process that's induced by this theorem is continuous, right?*

Answer. Yes — because the process we constructed earlier will satisfy Hölder continuity for all $s,t \in [0,1]$, and a Hölder continuous function is also continuous. (The property doesn't apply just to dyadic indices; it applies to real indices as well once we extend it.)

Now there is some flexibility on what to cover for this lecture and the next two. Prof. Kavvadias decided to mention some things about large deviation theory and introduce Brownian motion. We don't have time to go into its details; you will probably do this in Stochastic Calculus if you attend this next semester. But we'll construct it and prove some properties.

§24.2 Large deviation theory

First let's motivate why people are interested in large deviations, and what this is.

Suppose we have a sequence $(X_i)_{i\in\mathbb{N}}$ of i.i.d. random variables with finite mean $\mathbb{E}[X_i] = \overline{x}$, and we set $S_n = \sum_{i=1}^n X_i$. We studied this sum and its behavior, and proved limit theorems about it. One of those was the central limit theorem:

Theorem 24.3 (CLT)
Assuming that $\mathrm{Var}[X_i] = \sigma^2$ is finite, we have that

$$\mathbb{P}\left[S_n \geq n\cdot\overline{x} + a\sigma\sqrt{n}\right] \to \mathbb{P}[Z \geq a]$$

for all (fixed) $a \geq 0$, where $Z \sim \mathcal{N}(0,1)$.

So this is the central limit theorem — in other words, the central limit theorem just computes the asymptotic behavior of the probability that the average of the X_n's is concentrated around the mean $\overline{x}$. And it says that this behaves like a normal Gaussian.

So far, all our results were focused on examining the asymptotic behavior of averages of large sums near their mean. Large deviation theory asks the following natural question:

Question 24.4. What are the asymptotics of

$$\mathbb{P}[S_n \geq an] \quad \text{as } n\to\infty,$$

where $a > \overline{x}$?

This probability will tend to 0 by the central limit theorem (which says the average S_n/n has to be concentrated around the mean). But it doesn't talk about the rate of convergence — is it polynomial in n? Exponential in n?

That's the point of large deviations — to actually compute how fast this tends to 0 as $n\to\infty$, when n is a fixed number larger than the mean. In this brief chapter, we'll characterize this convergence, for certain types of random variables.

Let's first see an example where we can actually compute the asymptotic behavior of this probability easily.

Example 24.5
Let X_i are i.i.d. $\mathcal{N}(0,1)$. Then $\mathbb{P}[S_n \geq an]$ can be explicitly computed, because the sum of n independet Gaussians is still a Gaussian, and we can compute all the relevant information about it.

In that case, we have

$$\mathbb{P}[S_n \geq an] = \mathbb{P}[X_1 \geq a\sqrt{n}] \sim \frac{1}{a\cdot\sqrt{2\pi n}} e^{-a^2 n/2}.$$

So in that case, we have an explicit description of this probability. In other words,

$$-\frac{1}{n}\log\mathbb{P}[S_n \geq an] \to \frac{a^2}{2}$$

as $n\to\infty$. This is just a function of a, which is explicit.

We're going to prove a similar result, generalizing the function $a/2$ (it'll depend on the law of X_i), when X_i is not necessarily a Gaussian (but we still have a sequence of i.i.d. random variables). We'll still prove

that $-\frac{1}{n}\log\mathbb{P}[S_n \geq an]$ converges to some function, and we'll identify this function. So in other words, the probability the average is far from the mean decays *exponentially* to 0. This means i.i.d. random variables concentrate around the mean — the probability they're far from the mean decays exponentially.

(In the general case, it won't be easy to explicitly give the form of the function replacing $\frac{a}{2}$.)

So the main idea of large deviations is to understand the asymptotics of $\mathbb{P}[S_n \geq an]$. We know this tends to 0, but we want to identify the rate of convergence, and this turns out to be exponential in n; this means it converges to 0 very strongly.

§24.3 Existence of the limit

Before identifying the limit here, the first thing is to actually prove that the limit exists. For that, we're going to invoke a fact from real analysis for subadditive sequences, which says that if you take a sequence b_n which is subadditive, then b_n/n converges.

(This has a name; we're not going to prove it, but it's not hard and is a nice exercise, and we can try to do it on our own.)

Fact 24.6 — Let $(b_n)_{n\geq 1} \subseteq \mathbb{R}$ be subadditive, meaning that $b_{n+m} \leq b_n + b_m$ for all $n, m \in \mathbb{N}$. Then $\lim_{n\to\infty} \frac{b_n}{n}$ exists.

The limit might be $\pm\infty$, but in any case it exists.

The main observation here is that by independence (and the fact that the X_i have the same law), if we set $b_n = -\log\mathbb{P}[S_n \geq an]$, then this will be subadditive; and therefore our sequence will converge. Next lecture, we'll actually identify the limit.

Claim 24.7 — The sequence $b_n = -\log\mathbb{P}[S_n \geq an]$ is subadditive.

Proof. We have $\mathbb{P}[S_{n+m} \geq a(n+m)] \geq \mathbb{P}[S_n \geq an,\ S_{m+n} - S_n \geq am]$ (because if $S_n \geq an$ and $S_{m+n} - S_n \geq am$, then their sum will be at least $a(n+m)$). Now these two events are independent — the first event only depends on the first n terms, and the second event only depends on the next m terms (X_{m+1}, X_{m+2}, …, X_{m+n}). So this probability is the product of the probabilities. And since all of the X_i have the same law, the law of $S_{n+m} - S_n$ is the same as the law of S_m. So this is

$$\mathbb{P}[S_{n+m} \geq a(n+m)] \geq \mathbb{P}[S_n \geq an]\mathbb{P}[S_m \geq am].$$

And now we just take logarithms (because the logarithm of a product is the sum of logarithms). This gives

$$\log\mathbb{P}[S_{n+m} \geq a(n+m)] \geq \log\mathbb{P}[S_n \geq an] + \log\mathbb{P}[S_m \geq am],$$

and now by negating everything, we obtain that $b_{n+m} \leq b_n + b_m$ (for all $n, m \in \mathbb{N}$). □

So this means $\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}[S_n \geq an]$ exists, by this subadditive argument. Our job now, for the rest of this lecture and half of the next, is to actually identify the limit.

One observation is that this limit might be infinite as well, because the result from real analysis doesn't necessarily say whether the limit is finite or not; it just says the limit exists.

§24.4 The limit

Now one might wonder, what's the best way to approach such a problem? Here, what we're going to use is moment generating functions (we briefly introduced this when discussing characteristic functions).

Definition 24.8. We define the moment generating function of X as $M(\lambda) = \mathbb{E}[e^{\lambda X}]$.

This moment generating function might be infinite — it's always nonnegative, but there's no reason it has to always be finite.

What w're interested in is $\psi(\lambda) = \log M(\lambda)$ — this will play a crucial role in identifying our limit. In fact, the limit (as a function of a) will be constructed using ψ; we're going to see this in a few minutes.

Let's look at properties of this function ψ (the logarithm of the moment generating function).

Fact 24.9 — We have $\psi(0) = 0$.

This is clear, because if we set $\lambda = 0$, we get $M(0) = \mathbb{E}[e^0] = 1$.

Also, by Markov's inequality applied to exponentials, we have

$$\mathbb{P}[S_n \geq an] = \mathbb{P}[e^{\lambda S_n} \geq e^{\lambda na}] \leq e^{-\lambda na}\mathbb{E}[e^{\lambda S_n}].$$

And $\mathbb{E}[e^{\lambda S_n}]$ is just $M(\lambda)^n$, because $e^{\lambda S_n} = \prod e^{\lambda X_i}$, and by independence the expectation of the product is the product of expectations. SO we get

$$\mathbb{P}[S_n \geq an] \leq (e^{-\lambda A}M(\lambda))^n = \exp(-n(\lambda a - \psi(\lambda))).$$

So we've found an upper bound on this probability, and this is true for all $\lambda \geq 0$.

Now we can take logarithms of both sides, which gives

$$\log \mathbb{P}[S_n \geq an] \leq -n(\lambda a - \psi(\lambda)),$$

and dividing by $-n$ on both sides (which reverses the inequality), we get that

$$-\frac{1}{n}\log \mathbb{P}[S_n \geq an] \geq \lambda a - \psi(\lambda)$$

for all $\lambda \geq 0$. So we've obtained a lower bound on the quantity that we wanted to find a limit for. And this is true for all $\lambda \geq 0$; so in other words, this means

$$-\frac{1}{n}\log \mathbb{P}[S_n \geq an] \geq \psi^*(a),$$

where we define

$$\psi^*(a) = \sup_{\lambda \geq 0}(\lambda a - \psi(\lambda)).$$

Definition 24.10. We call ψ^* the Legendre transform of ψ.

So we've obtained that our sequence $-\frac{1}{n}\log \mathbb{P}[S_n \geq na]$ is bounded from below by the Legendre transform of ψ. This depends on ψ, but in lots of cases it's not explicit. This is a lower bound. It turns out that we're extremely lucky, and our sequence actually converges to $\psi^*(a)$. And that's what we're going to prove.

This is really remarkable because we got a lower bound of $\psi^*(a)$ just by using Markov's inequality, but it turns out it's actually the true limit. This gives a complete description of the asymptotics of our probability. Unfortunately, this ψ^* can be extremely complicated; but in some cases we can still compute it.

First let's state the main theorem.

Theorem 24.11 (Cramer's theorem)

Let (X_i) be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \overline{x}$, and let $S_n = \sum_{i=1}^n X_i$. Then

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}[S_n \geq an] = \psi^*(a) \quad \text{for all } a \geq \overline{x}.$$

In other words, what this means is that $\mathbb{P}[S_n \geq an]$ behaves like $e^{-\psi^*(a)n}$ as $n \to \infty$ — we indeed have exponential decay, but we also have a complete characterization of that decay due to $\psi^*(a)$. The exponential decay is unsurprising because averages of i.i.d. random variables tend to concentrate around the mean, and here we want the asymptotics of the probability that the average is far away from the mean.

Student Question. *Have we used the fact that $a > \overline{x}$?*

Answer. So far, in our lower bound argument, we haven't. But we'll see it in the proof of the other direction.

§24.5 Some examples

In the last ten minutes of today's lecture, we'll give a few examples where we can actually compute $\psi^*(a)$, and therefore give an explicit description of this limit. One example is the Gaussian, where we have an explicit description of $\psi^*(a)$; it'll be $a^2/2$, as we discussed earlier.

Example 24.12

Suppose that $X_i \sim \mathcal{N}(0,1)$.

Then we can explicitly compute ψ^* — we have $M(\lambda) = \mathbb{E}[e^{\lambda X}]$, and we know the density of a normal Gaussian, so we can explicitly compute this as

$$M(\lambda) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{\lambda x}\, dx.$$

And by standard integration techniques, this is $e^{\lambda^2/2}$, which implies $\psi(\lambda) = \lambda^2/2$ for all $\lambda \geq 0$. Then

$$\psi^*(a) = \sup_{\lambda \geq 0} \left(\lambda a - \frac{\lambda^2}{2}\right).$$

We want to maximize this quantity — we want to find λ for which $\lambda a - \frac{\lambda^2}{2}$ is maximized. And it's maximized when its derivative is 0, which occurs when $\psi'(\lambda) = a$. And we have $\psi'(\lambda) = \lambda$, so this means $\lambda = a$, which means $\psi^*(a) = a^2/2$.

Example 24.13

Suppose that $X_i \sim \text{Exp}(1)$ is the exponential random variable with parameter 1 — this is the random variable with density $f(s) = e^{-s}$ for all $s \geq 0$.

The exponential random variable is quite common; it appears all the time. Let's try to compute its Legendre transform. If $\lambda < 1$, then we have

$$M(\lambda) = \int_0^\infty e^{\lambda x - x}\, dx = \frac{1}{1-\lambda}.$$

(If $\lambda \geq 1$, then this blows up and is infinite.) This implies

$$\psi(\lambda) = -\log(1-\lambda)$$

for $\lambda < 1$. So we also have an explicit form here. We want to minimize $\lambda a - \psi(a)$, so we want its derivative to be 0, which means we're looking for λ for which

$$a = \psi'(\lambda) = \frac{1}{1-\lambda}.$$

This means $\lambda = 1 - \frac{1}{a}$, which implies

$$\psi^*(a) = a - 1 - \log(a).$$

So for the exponential random variable with parameter 1, we again have an explicit form of the Legendre transform, and this limit is again explicit and easy to compute.

Example 24.14

Suppose that $X_i \sim \text{Poisson}(1)$ — this means X_i takes values on nonnegative integers, with $\mathbb{P}[X = k] = \frac{e^{-1}}{k!}$ for every nonnegative integer k.

Now we have a discrete random variable with explicit density, so its moment generating function is

$$M(\lambda) = \sum_k \frac{1}{k!} e^{\lambda k - 1} = e^{e^\lambda - 1}.$$

So this is again explicit, and so

$$\psi(\lambda) = e^\lambda - 1$$

for all $\lambda \in \mathbb{R}$. We need to solve the equation $a = \psi'(\lambda)$, and that gives $a = e^\lambda$, so $\lambda = \log a$. Then plugging this into the definition of ψ^*, we obtain

$$\psi^*(a) = a \log a - a + 1.$$

So in this case as well (the Poisson random variable with parameter 1), we have an explicit form of ψ^*. But in general we don't — we can see it's defined in a complicated way. In order to find the explicit form, we have to find the value of λ maximizing $\lambda a - \psi(a)$; this means there's an optimization problem hidden here, and in most cases it's not easy. So the limit exists and is described explicitly, but the explicit form is not easy to obtain.

In the next lecture, we'll prove this theorem, and also define Brownian motion.

§25 December 5, 2024

§25.1 Cramer's theorem

Last lecture we started the chapter on large deviations. The main purpose of this chapter is to prove Cramer's theorem. We also gave an intuitive interpretation — suppose we have i.i.d. random variables with finite second moment. The cenral limit theorem states that if we take the average of this sequence, this will behave roughly as a Gaussian — so the probability the average is far away from the mean tends to 0 as $n \to \infty$. But the central limit theorem doesn't give you the decay rate of this probability. The purpose of Cramer's theorem is to explicitly describe the asymptotic behavior of the probability that the average is far away from the mean. And it makes sense because in general, i.i.d. things have the tendency of concentrating around their mean.

Theorem 25.1 (Cramer's theorem)

Let (X_i) be an i.i.d. sequence of random variables with mean $\mathbb{E}[X_i] = \overline{x}$, and let $S_n = \sum_{i=1}^n X_n$. Then

$$\lim_{n\to\infty} -\frac{1}{n} \cdot \log \mathbb{P}[S_n \geq na] = \psi^*(a)$$

for all $a \geq \overline{x}$, where $\psi^*(a) = \sup_{\lambda\geq 0}(\lambda a - \psi(\lambda))$, where $\psi(\lambda) = \log M(\lambda)$ (where M is the moment generating function of each X_i, i.e., $M(\lambda) = \mathbb{E}[e^{\lambda X_1}]$).

Here the only assumption we make is that all the random variables are in L^1, so their expectation is well-defined.

We said last lecture that the Legendre form ψ^* is complicated — it's hard to find a nice explicit form — but in certain cases we can find such a form (e.g., for Gaussians, Poisson random variables, and exponential random variables, as we saw last lecture). Last lecture we also proved that for all n, we have

$$-\frac{1}{n}\log \mathbb{P}[S_n \geq an] \geq \psi^*(a).$$

So what remains to do is prove an upper bound.

Before we do this, we'll need a lemma which states some regularity properties of the moment generating function.

Lemma 25.2

The functions M and ψ are continuous on the set $\mathcal{D} = \{\lambda \geq 0 \mid M(\lambda) < \infty\}$. Furthermore, they are also differentiable in the interior of $\mathcal{D}$, i.e.,

$$\mathring{\mathcal{D}} = \{x \in \mathcal{D} \mid \text{exists } \varepsilon > 0 \text{ such that } (x-\varepsilon, x+\varepsilon) \subseteq \mathcal{D}\},$$

with $M'(\lambda) = \mathbb{E}[X_1 e^{\lambda X_1}]$ and $\psi'(\lambda) = M'(\lambda)/M(\lambda)$.

The moment generating function might be infinite, so we consider all the values at which it's finite; and this lemma says that on that set, both M and its logarithm ψ are continuous and differentiable (of course, it doesn't make sense to talk about continuity or differentiability where they're infinite).

Proof. We're not going to prove continuity; we just apply the dominated convergence theorem to change between integration (here, expectation) and limits. So continuity follows from the dominated convergence theorem.

Now we're going to prove differentiability. The first thing to observe is that $\mathcal{D}$ is an interval (a sub-interval of $\mathbb{R}$) — this is because if we take any $\lambda_1 < \lambda < \lambda_2$ and we have that both λ_1 and λ_2 are in $\mathcal{D}$, then we also have that $\lambda \in \mathcal{D}$, since

$$e^{\lambda x} \leq e^{\lambda_1 x} + e^{\lambda_2 x}$$

for all $x \in \mathbb{R}$, which implies that

$$\mathbb{E}[e^{\lambda X_1}] \leq \mathbb{E}[e^{\lambda_1 X_1}] + \mathbb{E}[e^{\lambda_2 X_1}],$$

and the right-hand side is finite (because $\lambda_1, \lambda_2 \in \mathcal{D}$). And the only subsets of $\mathbb{R}$ with this property are intervals. So $\mathcal{D}$ is indeed an interval (because whenever we take any two points, any point in between these two points is also in our set; and this is a property characterizing intervals).

In order to show differentiability (and that the derivative has the desired form), we'll take limits — we'll write down the definition of the derivatives, and try to justify that you can interchange between expectations

and limits using the dominated convergence theorem. For this, note that

$$\frac{M(\lambda+h)-M(\lambda)}{h} = \mathbb{E}\left[\frac{e^{(\lambda+h)X_1}-e^{\lambda X_1}}{h}\right]$$

(we think of λ as fixed, and $h \neq 0$). The term on the inside converges to $X_1 \cdot e^{\lambda X_1}$ as $h \to 0$. So what we need to show is that as $h \to 0$, we can interchange between limits and expectation. For this, we need an upper bound on the inner function which is uniform in h when h is very small.

If we take h such that $2|h| \leq \min_{i=1,2}|\lambda - \lambda_i| = 2\varepsilon$, then we have

$$\left|\frac{e^{(\lambda+h)x}-e^{\lambda x}}{h}\right| = |x| \cdot e^{\overline{\lambda}}$$

for some $\overline{\lambda} \in [\lambda x, (\lambda+h)x]$ (by the mean value theorem). But this is at most $(e^{\lambda_1 x}+e^{\lambda_2 x}) \cdot \varepsilon^{-1}$. This means if we fix any λ_1, λ_2, and λ with this property and take h sufficiently small, then the term inside is bounded by this. But this term has finite expectation, because $\lambda_1, \lambda_2 \in \mathcal{D}$. So this function dominates the thing inside our expectation. And since that converges to $X_1 e^{\lambda X_1}$ as $h \to 0$, we're done. (And the same thing works for ψ.) □

Here the main tool is just the dominated convergence theorem. We'll see how we're going to use this in the proof.

Now let's proceed with the proof of Cramer's theorem. First, recall that we showed last lecture that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}[S_n \geq na] \geq \psi^*(a)$$

(for all $a \geq \overline{x}$, where $\overline{x}$ is the mean). We also justified why this limit exists (by proving that the sequence of logarithms is subadditive, and there's a fact that subadditive sequences of real numbers converge).

So the harder direction is to prove an upper bound, and that will be slightly complicated. Now the first thing we'd like to do is reduce to the case where $a = 0$, to get rid of a (so that we're just considering $\mathbb{P}[S_n \geq 0]$). This is relatively simple — if we set $\widetilde{X}_i = X_i - a$, then we have

$$\mathbb{P}[S_n \geq na] = \mathbb{P}[\widetilde{S}_n \geq 0]$$

(where $\widetilde{S}_n = \sum_{i=1}^n \widetilde{X}_i$). Also, if we take the moment generating function $\widetilde{M}$ of $\widetilde{X}_1$, this is given by $\widetilde{M}(\lambda) = e^{-\lambda a}M(\lambda)$, which implies that $\widetilde{\psi}(\lambda) = \psi(\lambda) - \lambda a$. So if we just shift everything by a, then we have these two equalities. This implies we need to show that

$$-\frac{1}{n}\log \mathbb{P}[\widetilde{S}_n \geq 0] \to \sup_{\lambda \geq 0} -\widetilde{\psi}(\lambda) = \widetilde{\psi}^*(0).$$

So from now on, we can assume $a = 0$, and that the mean of each X_i is at most 0. We'll see that it's useful because it makes some things a bit easier.

From now on, we'll drop the tildes; we'll treat all the X_i as having mean 0 (because we've seen here how to get the general case from this specific case). So from now on, we'll assume that $\overline{x} \leq 0$ and $a = 0$.

Now it suffices to show that

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}[S_n \geq 0] \geq \inf_{\lambda \geq 0} \psi(\lambda).$$

Note that we wanted to show our original limit was at most $\psi^*(a)$ (since we proved the opposite). But it originally had a minus sign (and the limit is equal to the liminf, because the limit exists). And then the right-hand side is the negative of $\psi^*(0)$, so this is exactly what we wanted to prove.

Now there are some technical arguments, but we'll go through them carefully.

The first observation, which makes our lives a bit easier:

First, consider the case where $\mathbb{P}[X_1 \le 0] = 1$. Then we can write

$$\mathbb{E}[e^{\lambda X_1}] = \mathbb{E}[e^{\lambda X_1}\mathbf{1}_{X_1<0}] + \mathbb{E}[e^{\lambda X_1} \cdot \mathbf{1}_{X_1=0}].$$

The first term remains as it is. and the second term — when $X_1 = 0$, it doesn't matter what λ is — it's always equal to 1 — so we just get $\mathbb{P}[X_1 = 0]$. So we get

$$M(\lambda) = \mathbb{E}[e^{\lambda X_1}\mathbf{1}_{X_1<0}] + \mathbb{P}[X_1 = 0].$$

And the first term converges to 0 as $\lambda \to \infty$ (because it's always bounded by 1, and for any fixed X_1 it converges to 0 because the exponent is negative; so we can use the dominated convergence theorem). So this means

$$\inf_\lambda \psi(\lambda) \le \lim_{\lambda\to\infty} \psi(\lambda) \le \log \mathbb{P}[X_1 = 0].$$

But we also have

$$\mathbb{P}[S_n \ge 0] = \mathbb{P}[X_1 = 0, \dots, X_n = 0]$$

(since all of them are always negative). And these events are independent, so the right-hand side is $\mathbb{P}[X_1 = 0]^n$. Taking logarithms, we get

$$\frac{1}{n}\log \mathbb{P}[S_n \ge 0] \ge \log \mathbb{P}[X_1 = 0]$$

for all n, and the latter term is at least $\inf_{\lambda\ge 0}\psi(\lambda)$ (as we just saw). This is true for all n. So the required bound holds, in the case that X_1 is nonpositive with probability 1 — in that case, we get the statement relatively easily by this analysis.

So to prove the required bound, it suffices to consider the case where X_1 takes positive values with positive probability. (We know X_1 has negative expectation, but certainly it can take positive values with positive probability.) So from now on, we'll assume that $\mathbb{P}[X_1 > 0] > 0$; this is the hard case.

This will have two steps, and we'll see how we use the lemma from before. In the first step, we'll assume $M(\lambda)$ is always finite. (This is in general not true — for example, with the exponential random variable with parameter 1, it's infinite for $\lambda \ge 1$.)

Case 1 ($M(\lambda) < \infty$ for all $\lambda \ge 0$). The idea is to reduce to the case where we have mean 0. In our case, it's not 0 — it's $\overline{x}$. So how do we construct a mean-0 random variable starting with a variable of nonzero mean? The idea is to change the law of X — if X has law μ, we're going to weight it by some other function. And we'll do this by using Radon–Nikodym derivatives.

What we mean by this is that for some $\theta \in \mathbb{R}$ which we'll determine later, we define μ_θ to be the Borel measure defined by

$$\frac{d\mu_\theta}{d\mu} = \frac{e^{\theta X}}{M(\theta)}$$

(the left-hand side is the Radon–Nikodym derivative of μ_θ with respect to μ, where μ is the law of X_1). Recall that this means

$$\mu_\theta(A) = \int_A \frac{e^{\theta X}}{M(\theta)}$$

(from the definition of a Radon–Nikodym derivative). The idea is that for an appropriate value of θ, the random variable with law μ_θ has mean 0, which will allow us to reduce to the mean-0 case. (This is a standard technique in general when we want to shift the mean of a random variable — to weight its law by an appropriate function, as we did here.)

This means that for any bounded measurable f defined on $\mathbb{R}$, we have

$$\mathbb{E}_\theta[f(X_1)] = \int_{\mathbb{R}} f(x)\frac{e^{\theta x}}{\mu(\theta)}\,d\mu(x)$$

(where $\mathbb{E}_\theta$ is the expectation with respect to μ_θ). In particular, if we define $g(\theta) = \mathbb{E}_\theta[X_1]$ (this is a function of θ), we get

$$g(\theta) = \int_{\mathbb{R}} x \cdot \frac{e^{\theta x}}{M(\theta)}\,d\mu(x).$$

And the idea is to show that for a particular value of θ, this is 0.

What's the idea here? We know that X_1 takes positive values with positive probability. If we look at the term on the inside, if we take $\theta \geq 0$, then this is at least

$$\int_0^\infty x\,d\mu(x) \cdot \frac{1}{M(\theta)},$$

because $e^{\theta x} \geq 1$ when we integrate over all $x \in (0,\infty)$. And $\int_0^\infty x\,d\mu(x)$ is positive, because $\mathbb{P}[X_1 > 0] > 0$. This means

$$\lim_{\theta\to\infty} \frac{g(\theta)}{>} 0,$$

because $\lim_{\theta\to\infty} M(\theta) = \mathbb{P}[X_1 > 0] > 0$. So $g(\theta)$ is bounded by a number (depending on θ) which tends to something positive, which means this limit is positive.

But we also know that

$$g(0) = \mathbb{E}[X_1] = \overline{x} \leq 0$$

(where $\mathbb{E}$ denotes expectation with respect to μ). So we have a continuous function g whose limit at ∞ is positive and whose starting point is nonpositive, and this means there exists some $\theta \geq 0$ such that $g(\theta) = 0$.

This means $\mathbb{E}_\theta[X_1]$ has been shifted, and now it's 0 — so now we have a new random variable with mean zero. And for that random variable with zero mean, we can apply known theorems — the central limit theorem and so on. (We'll see soon how this is used.)

(This is a standard technique, called the *weighting* technique — we've weighted μ by the function $e^{\theta x}$.)

Student Question. *How did we get that* $\lim_{\theta\to\infty} M(\theta) = \mathbb{P}[X_1 > 0]$*?*

Answer. There is some issue here; what we wrote earlier is not correct. We should write

$$g(\theta) = \frac{\int_{\mathbb{R}} xe^{\theta x}\,d\mu(x)}{\int_{\mathbb{R}} e^{\theta x}\,d\mu(x)}.$$

What we actually ahve is that this ratio tends to something positive as $\theta \to \infty$ — if we take $\theta \to \infty$, in the case where $\mathbb{P}[X_1 > 0] > 0$, this quantity will tend to some positive number. The point is that if x is a positive value, then $xe^{\theta x}$ will be larger than $e^{\theta x}$; so when we take $\theta \to \infty$, we'll get something which is positive. And we have a continuous function starting from 0, whose limit is positive; this means we can find some θ such that $\mathbb{E}_\theta[X_1] = 0$.

From now on, we fix such a θ; and let's see how to lower-bound $\mathbb{P}[S_n \geq 0]$. For this, we'll do the following. We fix $\varepsilon > 0$, which we'll eventually take tending to 0. Then we have that

$$\mathbb{P}[S_n \geq 0] \geq \mathbb{P}[S_n \in [0, n\varepsilon]].$$

And now is the tricky part — this probability is at least

$$\mathbb{E}[e^{\theta(S_n - \varepsilon n)}\mathbf{1}_{S_n \in [0,n\varepsilon]}].$$

This is because if $S_n \in [0, n\varepsilon]$ then this exponent is nonpositive.

And now the point is that this expectation here can be estimated using the central limit theorem. That's the main advantage, and why we've shifted the mean — this can be interpreted as the expectation with respect to our new measure μ_θ, and for this we can use the central limit theorem — this is

$$(M(\theta))^n \mathbb{P}_\theta[S_n \in [0, \varepsilon n]] e^{-\theta\varepsilon n}.$$

And the idea is that we know the asymptotics of $\mathbb{P}_\theta[S_n \in [0, \varepsilon n]]$, because we can apply the central limit theorem — we have i.i.d. random variables which have mean 0 (under this measure), so we can apply the central limit theorem. That'll give us a lower bound on this probability. Remember that we couldn't have directly applied the central limit theorem to $\mathbb{P}[S_n \geq 0]$ because the mean isn't 0; but it can help us with this probability.

And by the central limit theorem, this will tend to the probability that a Gaussian random variable with mean 0 is nonnegative, and this probability is $\frac{1}{2}$ (by symmetry). So by the central limit theorem, we have that

$$\mathbb{P}_\theta[S_n \in [0, \varepsilon n]] \to \frac{1}{2}$$

as $n \to \infty$. This means if we take logarithms, then

$$\frac{1}{n} \log \mathbb{P}[S_n \in [0, \varepsilon n]] \approx \frac{\log 1/2}{n} \to 0$$

as $n \to \infty$. And this means

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[S_n \geq 0] \geq \psi(\theta) - \theta\varepsilon.$$

And since ε was arbitrary, we can now take $\varepsilon \to 0$, and we obtain that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[S_n \geq 0] \geq \psi(\theta) \geq \inf_{\lambda \geq 0} \psi(\lambda).$$

So if we assume the moment generating function is always finite, then we are done — we've completed the proof, using this trick. The whole point of shifting the mean to 0 was to be able to apply the central limit theorem — we couldn't have directly applied it to the original, whose mean was negative.

Case 2 (The general case). Now we've already seen the main ideas of the proof; but we need to generalize it to the general case, where $M(\lambda)$ is not always finite; and there are some slightly annoying issues. In that case, you have to truncate — you have to condition on the event that your random variable takes values on a fixed compact set. We've done this many times in the course, where we truncate a random variable.

So let ν be the law of X_1 conditioned on the event that $|X_1| \leq k$ (where k is some fixed positive number that we'll eventually let tend to ∞). The point is that under this measure ν the moment generating function of X_1 (where we're conditioning on this event) is finite. So the theorem holds when X_1 has law ν, from the previous step.

In particular, suppose we define

$$\psi_k(\lambda) = \log \int_{-k}^{k} e^{\lambda x} \, d\mu(x).$$

This quantity is some finite real number (because the integral is truncated), so if we consider the moment generating function of X_1 with respect to ν, this is

$$\log \int_{\mathbb{R}} e^{\lambda x} \, d\nu(x) = \psi_k(\lambda) - \log \mu([-k, k]).$$

And this is a real number, with this expression.

And the idea is that the theorem statement holds for X_1 if it has law ν, and somehow we need to use this by taking $k \to \infty$.

In particular, if we set μ_n as the law of S_n under μ (the original measure), and ν_n to be the law of S_n under ν, then we have that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n([0,\infty)) \geq \log \mu([-k,k]) + \frac{1}{n} \log \liminf_{n\to\infty} \nu_n([0,\infty))$$

(recall the left-hand side is just the probability in our statement — $\mu_n([0,\infty)) = \mathbb{P}[S_n \geq 0]$ by definition), by the definitions of μ_n and ν_n. And this is at least

$$\inf_{\lambda \geq 0} \psi_k(\lambda).$$

Call this quantity $\mathcal{J}_k$. That's because this is at least $\log \int_{\mathbb{R}} e^{\lambda x}\, d\nu(x)$, by the previous step. And this logarithm can be expressed as the desired sum.

But now ψ_k converges increasingly to ψ as $k \to \infty$. This means $\mathcal{J}_k$ (as $k \to \infty$) will also converge increasingly to some number $\mathcal{J}$. And since this is true for all k, this implies

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n([0,\infty)) \geq \mathcal{J}.$$

In order to complete the proof, we need to show that $\mathcal{J} \geq \inf_\lambda \psi(\lambda)$; if we manage to prove this, then we're done. So that's the remaining step.

There's only one final step, which is to show this.

Remark 25.3. Explicitly, we have

$$\nu(A) = \frac{\mu(A \cap [-k,k])}{\mu([-k,k])}.$$

So ν_n is μ_n conditioned on all the X_is being between $-k$ and k. This means

$$\nu_n([0,\infty)) = \frac{\mu_n(\{S_n \geq 0\} \cap \{X_i \in [-k,k]\})}{\mu([-k,k])^n}.$$

And that's how you obtain our inequality. So ν_n is just μ_n conditioned on all of $X_1, \ldots, X_n$ being in $[-k,k]$.

The rest is a compactness argument, which is brief but not conceptually easy. However, we have seen it before. The main idea is the following: consider $\bigcap_{k\geq 0}\{\lambda \geq 0 \mid \psi_k(\lambda) \leq \mathcal{J}\}$. What is this set? If we fix k and consider all λ for which $\psi_k(\lambda) \leq \mathcal{J}$, this is a closed set (because ψ_k is a continuous function). And these sets, because ψ_k increases to ψ, are decreasing in k. So this means we have nonempty compact sets, and we take their intersection; this means the intersection is also nonempty (whenever you have a sequence of nested compact sets, its intersection is nonempty; we also showed this several lectures ago, maybe when constructing the Lebesgue measure). This means you can find $\lambda_0 \geq 0$ such that $\psi_k(\lambda_0) \leq \mathcal{J}$ for every $k \in \mathbb{N}$. And since $\psi_k(\lambda_0) \to \psi(\lambda_0)$, this means $\psi(\lambda_0) \leq \mathcal{J}$. And this implies $\inf_{\lambda\geq 0} \psi(\lambda) \leq \psi(\lambda_0) \leq \mathcal{J}$. Therefroe the liminf we were interested in, which is exactly this $\frac{1}{n} \log \mathbb{P}[S_n \geq 0]$, is at least this infimum; and that's the end of the proof.

So the theorem has three main steps. The first is to prove it when X_1 is nonpositive with probability 1 (the easy case). The next is to prove it in the case where the moment generating function is finite always. Iin the third step, we handle the general case; for this, we truncate by conditioning on the event that X_1 lies in the large compact set $[-k,k]$. This defines a new measure ν (the conditional law of X_1 given that it lies in the interval with endpoints $\pm k$). And then we use a slightly complicated argument to actually complete the proof.

This is a complicated proof, and you don't have to have followed all the details. But it's one of the proofs that needs to be done at least once in life, and the result is quite important — Cramer's theorem is quite a strong theorem.

This is more or less all we'll mention about large deviations.

We have one more lecture, next Tuesday. Then we'll define Brownian motion, and if possible, prove some of its properties. If you want to learn more about it, attend Stochastic Calculus. It was first constructed by Wiener using Fourier restriction theory; we'll give a more elegant proof due to Kolgomorov, using his continuity criterion.

§26 December 10, 2024

Today is the final lecture. Prof. Kavvadias thought it would be a good idea to rigorously define and construct Brownian motion, and possibly mention a few properties of this object. Unfortunately we don't have much time. But you'll see lots of properties of Brownian motion in 18.676.

Last lecture we finished a very complicated proof of Cramer's theorem; but it's useful to know. (We will not be asked for that proof on the exam; but he may ask other results, like Kolgomorov's 0-1 law or the dominated convergence theorem — things whose proofs are relatively short. So we should be aware of those proofs. But there will be problems as well, so we'll have to think a bit and be a bit original.)

§26.1 Brownian motion

Brownian motion, as the name suggests, was named after Browne, who observed that the erratic motion of small particles in water follows a certain patter. So it was first invented to describe how these particles behave. Then there was a physical model developed by Stein (1905), without a rigorous proof. The rigorous construction was done later in 1923. But this proof is complicated because it uses Fourier theory, specifically random Fourier series. We'll construct it using Kolgomorov's continuity criterion (this is due to Kolgomorov and Levy).

First, let's define what exactly Brownian motion is. (We'll work in one dimension, but this also applies in higher dimensions.)

Definition 26.1. Let $B = (B_t)_{t \geq 0}$ be a continuous process on $\mathbb{R}$ (indexed by nonnegative real numbers, and taking real values). We say B is a Brownian motion started from x (where x is a fixed real number) if the following conditions hold:

(i) $B_0 = x$ almost surely.

(ii) We have $B_t - B_s \sim \mathcal{N}(0, t-s)$ for all $0 \leq s < t$.

(iii) B has independent increments, and these increments are also independent of B_0. In other words, for any increasing sequence $0 \leq t_1 < t_2 < \cdots < t_n$, the n random variables B_{t_1}, $B_{t_2} - B_{t_1}$, $B_{t_3} - B_{t_2}$, $\ldots$, $B_{t_n} - B_{t_{n-1}}$ are independent.

So a Brownian motion is a random process which is continuous (we'll work with the one-dimensional case) and satisfies these three properties. We're going to construct such a process; so such a process exists.

It's also unique. If we assume the process is continuous, (ii) and (iii) uniquely determine the law of the entire process — that requires a bit of work, but it's not extremely hard to show. However, if you *don't* assume continuity, this is not the case — you can have two different processes which satisfy (i), (ii), and (iii), but are not the same (where they're not continuous). To see this, we have the following example.

Definition 26.2. A standard Brownian motion is one where $B_0 = 0$ almost surely.

Example 26.3

Suppose we start with a standard Brownian motion $B = (B_t)_{t \geq 0}$, and let U be a random variable, independent of B, which is uniformly distributed on $[0, 1]$. Define the process

$$\widetilde{B}_t = \begin{cases} B_t & t \neq U \\ 0 & \text{otherwise.} \end{cases}$$

Then this process has the same finite-dimensional distributions as B — i.e., if we take any fixed $t_1 < \cdots < t_n$ and consider the vectors $(B_{t_1}, \ldots, B_{t_n})$ and $(\widetilde{B}_{t_1}, \ldots, \widetilde{B}_{t_n})$, then these two vectors have the same law. In particular, this means (ii) and (iii) are satisfied by $\widetilde{B}$.

But $\widetilde{B}$ is not continuous at 0 — the point is that B_U is nonzero almost surely.

So $\widetilde{B}$ is not continuous, but satisfies (i), (ii), and (iii). This means these three conditions aren't enough to define the process; you need continuity.

To see why B_U is nonzero almost surely, we can condition on U — we can write

$$\mathbb{P}[B_U = 0] = \mathbb{E}[\mathbb{E}[\mathbf{1}_{B_U = 0} \mid U]].$$

And if we condition on U, then B_U is just a normal random variable with mean 0 and variance U. And if you have a normal distribution, then the probability it takes any fixed value is 0. So $\mathbb{E}[\mathbf{1}_{B_U=0} \mid U] = 0$ for any U, which means this probability is 0.

So with probability 1, $\widetilde{B}$ is not continuous; but it satisfies (2) and (3), showing that we do need continuity in order to get uniqueness.

§26.2 Existence of Brownian motion

Theorem 26.4 (Wiener)

There exists a Brownian motion on some probability space.

(Wiener was the first person to rigorously construct Brownian motion, but the proof we'll see is not due to him.)

Let's try to prove this; we'll come back to Kolgomorov's continuity criterion (which we said earlier we're going to use).

The first thing to do is construct the process in $[0, 1]$; then by translating, we can define it for all real numbers.

First, we'll define the process on the set of dyadic rationals. Then we'll prove this process satisfies a certain Hölder continuity condition (the one in Kolgomorov's continuity criterion), and use that to extend it continuously to the entire unit interval.

Now let's proceed with the details. Define $\mathcal{D}_0 = \{0, 1\}$ as the set of dyadic rationals of order 0, and for every $n \geq 1$, let $\mathcal{D}_n = \{k \cdot 2^{-n} \mid 0 \leq k \leq 2^n\}$ be the set of dyadic rationals of order n (as in the proof of Kolgomorov's continuity criterion). Let $\mathcal{D} = \bigcup_n \mathcal{D}_n$ be the set of all dyadic rationals in $[0, 1]$. The idea is to construct a process indexed by this set satisfying (ii) and (iii) when s and t are dyadic rationals, and use Kolgomorov's continuity criterion to extend the process to the entire unit interval.

For that, we start with a sequence $(Z_d)_{d \in \mathcal{D}}$ (indexed by $\mathcal{D}$) of independent random variables, which are Gaussians with mean 0 and variance 1, all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Remember that

here we have a countable collection of random variables with some fixed distribution, and we've shown (at the beginning of the course) that for any such situation, we can find a probability space such that all of them are defined on the same probability space.

The idea is to use this process to construct Brownian motion inductively, using induction on n (to define it on all dyadic rationals).

The first step, where we define the process on $\mathcal{D}_0$, is quite easy — we just set $B_0 = 0$ and $B_1 = Z_1$.

Now, we use induction in the following way. Suppose we've already constructed the Brownian motion on $\mathcal{D}_{n-1}$ — so we've already constructed $(B_d)_{d\in\mathcal{D}_{n-1}}$. We'll build the process $(B_d)_{d\in\mathcal{D}_n}$ as follows. We fix a dyadic rational $d \in \mathcal{D}_n \setminus \mathcal{D}_{n-1}$ (if $d \in \mathcal{D}_{n-1}$, then we already know the process there, by the induction hypothesis). The set $\mathcal{D}_{n-1}$ partitions the unit interval, so d lies in a unique interval with endpoints in $\mathcal{D}_{n-1}$; let d_- and d_+ be the endpoints of that interval, i.e.,

$$d_- = d - 2^{-n} \quad \text{and} \quad d_+ = d + 2^n.$$

In other words, imagine we draw the unit interval and partition it using a_1, a_2, a_3, and a_4; and imagine d is in the middle of a_2 and a_3.

And we know the Brownian motion at a_2 and a_3, so we define the Brownian motion at d as it should be, taking (ii) and (iii) into consideration. We set B_d to be the average of the Brownian motion at the two endpoints, plus a correction term — so we set

$$B_d = \frac{B_{d_+} + B_{d_-}}{2} + \frac{Z_d}{2^{(n+1)/2}}.$$

What does this result in? If we consider $B_d - B_{d_-}$, this difference is

$$B_d - B_{d_-} = \frac{B_{d_+} - B_{d_-}}{2} + \frac{Z_d}{2^{(n+1)/2}},$$

and similarly

$$B_{d_+} - B_d = \frac{B_{d_+} - B_{d_-}}{2} - \frac{Z_d}{2^{(n+1)/2}}.$$

So we have these explicit forms. And the idea is to show that indeed, these two increments $B_d - B_{d_-}$ and $B_{d_+} - B_d$ are independent, so that (iii) is satisfied. By choosing B_d in this way, indeed we see these satisfy (ii) — that's why we added the correction term $\frac{Z_d}{2^{(n+1)/2}}$. So what we need is these two increments are independent.

And remember this is a linear combination of Gaussians, so it's also Gaussian. And both have mean 0. So it remains to verify that their covariances are 0 — because the vector of the two differences is a Gaussian, so they're independent if and only if their covariance is 0 (this is something we've already done).

For that, we need the following simple observation. If we set

$$N_d = \frac{B_{d_+} - B_{d_-}}{2} \quad \text{and} \quad N_{d'} = \frac{Z_d}{2^{(n+1)/2}},$$

then we see by the induction hypothesis that that N_d and N'_d are independent centered Gaussian random variables, with the same distribution $\mathcal{N}(0, 2^{-(n+1)})$. (This follows by the induction hypothesis, which states that N_d has this law; and N'_d does because $Z_d \sim \mathcal{N}(0,1)$, and when you divide by a number, the variance is divided by its square.)

What this means is that if you take the covariance of $B_d - B_{d_-}$ and $B_{d_+} - B_d$, then this is just

$$\mathrm{Cov}[B_d - B_{d_-}, B_{d_-} - B_d] = \mathrm{Cov}[N_d + N'_d, N_d - N'_d].$$

And since N_d and N'_d are independent, this is just

$$\mathrm{Var}(N_d) - \mathrm{Var}(N'_d).$$

But their variances are equal, so this is 0. This means indeed our increments are independent. And then we proceed inductively. So far, we've constructed a process indexed by $\mathcal{D}_n$ such that it satisfies (ii) and (iii), when s and t are restricted to $\mathcal{D}_n$. This shows indeed we can define the process on the set of all dyadic rationals $\mathcal{D}$ (inductively), and this satisfies (ii) and (iii) when s and t are dyadic rationals.

(The point is that at every step, you define the Brownian motion as it should be in order to satisfy (ii) and (iii).)

Now that we've defined the process on $\mathcal{D}$, we need to define it on the entire unit interval, and for that we need Kolgomorov's continuity criterion. For that, we need some regularity condition on the increments — we need the pth moment of $B_t - B_s$ to satisfy a bound with respect to $|t-s|$. But this is not hard to see — if we fix any $s < t$ in $\mathcal{D}$, then if we take any $p > 0$, we have

$$\mathbb{E}[|B_t - B_s|^p] = |t-s|^{p/2} \cdot \mathbb{E}\left|\frac{B_t - B_s}{\sqrt{t-s}}\right|^2.$$

And by (ii), this random variable inside the moment is a standard Gaussian. So this is the same as

$$|t-s|^{p/2} \cdot \mathbb{E}[|Z|^p]$$

where Z is a standard Gaussian. And the $\mathbb{E}[|Z|^p]$ term is a constant, only depending on p.

So now we can apply Kolgomorov's continuity criterion, since this is truee for all p. This implies by Kolgomorov's continuity criterion that $(B_d)_{d\in\mathcal{D}}$ is α-Hölder continuous for all $\alpha < \frac{1}{2}$, almost surely (this is exaxctly what Kolgomorov's continuity criterion states). It doesn't matter which exponent you have for the construction; it's enough to have Hölder continuity for some exponent.

Now if you fix any $t \in [0,1]$ and consider any $(t_n) \subseteq \mathcal{D}$ such that $t_n \to t$ (recall that the dyadic rationals are dense in the unit interval), we can just set $B_t = \lim_{n\to\infty} B_{t_n}$. And the Hölder continuity condition guarantees that this limit exists; and that the process we obtain is continuous as well.

And now we have a continuous process. So what remains to prove is that this continuous process we've constructed satisfies (ii) and (iii).

Now that we have constructed our process, let's prove that it satisfies (ii) and (iii). For that, fix an increasing sequence $0 = t_0 < t_1 < t_2 < \cdots < t_k$; and we'll just approximate them. We want to know that if we take B and restrict to these times, the increments are independent. And we'll do this by approximating them by dyadic rationals — because for dyadic rationals, we know we really do have independent increments.

So we also let $t_0^n < t_1^n < \cdots < t_k^n$ be numbers in $\mathcal{D}$ such that $t_n^n \to t_j$ as $n \to \infty$, for each $0 \le j \le k$. In other words, we're just approximating our numbers by dyadic rationals.

Now we have continuity, so if you take the vector $(B_{t_1^n}, \ldots, B_{t_k^n})$, we have that this vector converges to the corresponding vector $(B_{t_1}, \ldots, B_{t_k})$ as $n \to \infty$, almost surely.

And for every fixed n, we have that the increments $B_{t_j^n} - B_{t_{j-1}^n}$ (for $1 \le j \le k$) are independent Gaussians with mean 0 and variances $t_j^n - t_{j-1}^n$. Somehow we need to prove that the corresponding increments in $(B_{t_1}, \ldots, B_{t_k})$ are also Gaussians and independent. For that, we're just going to compute the characteristic function of the increments of B_{t_j}, and use this almost sure convergence and the fact that the increments for all n are independent Gaussians with the specified variances.

This means if we take

$$\mathbb{E}\left[\exp\left(i\sum_{j=1}^k u_j(B_{t_j^n} - B_{t_{j-1}^n}\right)\right],$$

we know this for all n because we've got independent Gaussians wth known variances. So this expectation is given by

$$\prod_{j=1}^{k} e^{(t_j^n - t_{j-1}^n)u_j^2/2}.$$

And as $n \to \infty$, the first term converges to

$$\mathbb{E}\left[\exp\left(i\sum_{j=1}^{k} u_j(B_{t_j} - B_{t_{j-1}})\right)\right],$$

and the second term converges to

$$\prod_{j=1}^{k} e^{-(t_j - t_{j-1})u_j^2/2}.$$

This is true for any $u_1, \ldots, u_k$. So if you take the increments of the variables, we've computed the characteristic functions. And this is the characteristic function of independent Gaussians with variances $t_j - t_{j-1}$. So if we take increments, then our vector $(B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}})$ has the same characteristic function, and therefore the same law, as $(\sqrt{t_1 - t_0} \cdot Z_0, \sqrt{t_2 - t_1} \cdot Z_1, \ldots, \sqrt{t_n - t_{n-1}} \cdot Z_{n-1})$, where here the $Z_j \sim \mathcal{N}(0,1)$ are independent standard Gaussians. So we've seen that the increments of B (on any arbitrary increasing sequence) are independent Gaussians, with variance given by the gap. So this shows (ii) and (iii) are satisfied.

So we have constructed a process in $[0,1]$ which is continuous and satisfies (i), (ii), and (iii). The last step is to just define the process for the entire real line. (This is quite easy.)

Now to complete the proof, somehow the idea is we partition $\mathbb{R}$ with intervals whose endpoints are consecutive integers; and in every such interval, we just sample an independent Brownian motion — we just translate by n units of time. And somehow we glue everything together so that we have a single process.

To do this rigorously, we take a sequence (B_t^i) where $t \in [0,1]$ and $i \in \mathbb{N} \cup \{0\}$, of independent Brownian motions. And then we set

$$B_t = B_{t-\lfloor t \rfloor}^{\lfloor t \rfloor} + \sum_{i=0}^{\lfloor t \rfloor - 1} B_1^{(i)}$$

(note that $t - \lfloor t \rfloor$ is always in $[0,1)$; the sum corresponds to gluing the rest). It's not hard to see this is a continuous process which satisfies (ii) and (iii).

So the whole point is to construct the process on $[0,1]$, and that's all the hard work that we've done (the essence is defining it on $[0,1]$ — first defining it on the dyadic rationals, proving the regularity condition, and invoking Kolgomorov). That constructs the process in one dimension. So indeed such a process exists, and we have rigorously constructed it. (If we want to construct the process in higher dimensions, we take d independent Brownian motions, and consider their vector.)

§26.3 Some comments

One quick comment is that we get from the proof that the process is Hölder continuous with any $\alpha < \frac{1}{2}$ — so we don't just have continuity. One could ask what other properties it has. It is highly non-differentiable — it's 'fractal' and has a lot of corners. If you try to draw the function, you'll get something whose graph is not smooth. It's quite an irregular, messy function.

(The reason we needed to do $[0,1]$ first and then extend to $\mathbb{R}_{\geq 0}$ is that Kolgomorov's criterion only works for compact sets.)

Student Question. *Is this construction unique — can we show two Brownian motions started at the same time have the same law?*

Answer. Unfortunately we won't show it. But uniqueness is hidden in the definition. If you restrict to probability measures on the set of continuous functions, and you know their marginals (if you take any fixed vector and you know its law), then you determine the law of the process. And we have determined the law of the vector restricted to any fixed times. So this guarantees uniqueness. This is not obvious, but it's some usual theory about π-systems and so on.

But you need continuity — otherwise you don't have uniqueness, as we showed with our example earlier.

As mentioned earlier, if we want to define the process in d dimensions, we can take $B = (B^1, \dots, B^d)$ to be a vector of d independent one-dimensional Brownian motions.

§26.4 Some properties

People are interested in studying Brownian motion because it has some very nice invariance properties. We'll state some of these properties without proof, in d dimensions.

Proposition 26.5

Let B be a standard Brownian motion on $\mathbb{R}^d$ (i.e., one that starts from 0). Then we have the following properties:

(1) (Rotational invariance) It is invariant under multiplication by orthogonal matrices — i.e., if U is an orthogonal matrix on $\mathbb{R}^d$, then if we consider the process $U \cdot B = (UB_t)_{t\geq 0}$, this is also a Brownian motion. In particular, the Brownian motion is symmetric (i.e., $-B$ is still a Brownian motion).

(2) (Scaling property) For any fixed $\lambda > 0$, if we consider the process $(\lambda^{-1/2} B_{\lambda t})_{t\geq 0}$ (where we scale time, and partialy scale the values), then this is also a Brownian motion.

(3) (Markov property) For every $t \geq 0$, $(B_{t+s} - B_t)_{s\geq 0}$ is still a Brownian motion, and is independent of $(B_u)_{0\leq u\leq t}$.

So (1) says we have rotational invariance and so on. And (2) is a scaling property, where we scale time by λ. And (3) is called the Markov property — if we run the process up to time t, and then subtract the value of the process at time t from the remainder, what you get is independent of what you saw before. As a picture, imagine we've run the process up to some point 1, and then we continue the process in a yellow part. If you translate the yellow part so that its starting point goes to 0, then (3) states that the yellow part has the same distribution as the white part.

These are some basic but very important properties of the Brownian motion.

We unfortunately don't have time to do many things. So the last property is *time inversion* of the Brownian motion.

Theorem 26.6 (Time inversion)

Suppose that $(B_t)_{t\geq 0}$ is a standard Brownian motion. Then if we define a process $X = (X_t)_{t\geq 0}$ by

$$X_t = \begin{cases} 0 & t = 0 \\ t \cdot B_{1/t} & t \neq 0, \end{cases}$$

then X is also a standard Brownian motion.

So if you invert time — you take the process $tB_{1/t}$ — this process is *still* a Brownian motion. Again, proving this is simple. We have to prove two things. The first is that the increments of X are independent and are Gaussians with specified variances (this is conditions (ii) and (iii)). And you have to prove continuity. Clearly X is continuous for $t > 0$, so what we have to show is that it is continuous at 0. So these are the two steps: continuity at 0, and specifying the laws of the increments.

Proof. First, if we fix any $0 \le t_1 < \cdots < t_n$, then

$$\operatorname{Cov}[B_{t_{j-1}}, B_{t_j}] = t_{j-1}$$

for all $1 \le j \le n$. (This follows by (ii) and (iii).) So in order to specify the marginals of X (or any finite list of increments), it suffices to show that this property holds when we replace B by X. For this, we have

$$\operatorname{Cov}[X_{t_{j-1}}, X_{t_j}] = t_{j-1}t_j \operatorname{Cov}[B_{1/t_{j-1}}, B_{1/t_j}] = t_{j-1}t_j \cdot \frac{1}{t_j}$$

(note that order gets reversed, because we've taken inverses). And this is just t_{j-1}, for all $1 \le j \le n$. So this means if we take the vector

$$(B_{t_1}, B_{t_2}, \ldots, B_{t_n}),$$

then this vector has the same law as the corresponding vector

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$$

(because we have explicit formulas for their covariance).

But we also need continuity (this alone isn't enough). What it remains to show is that $\lim_{t\to 0} X_t = 0$ almost surely (that's the only thing it remains to show). How do we show this? The first observation is that since (B_{t_j}) and (X_{t_j}) have the same law for any (t_j), we get that if we restrict to positive rationals, $(X_t)_{t\in\mathbb{Q}}$ and $(B_t)_{t\in\mathbb{Q}}$ have the same law (because $\mathbb{Q}$ is countable, and this is true for any *finite* collection of rationals by what we've shown).

This means if you take the limit as $t \to 0$ when t is a positive rational, you get $\lim_{t\to 0, t\in\mathbb{Q}} X_t = \lim_{t\to 0, t\in\mathbb{Q}} B_t$. But the latter lmit is the same as $\lim_{t\to 0} B_t$, which is 0 almost surely.

So if you restrict t to positive rationals tending to 0, then $X_t \to 0$. From that, we want to deduce that the limit is indeed 0 along *any* sequence. This follows by the density of rationals. What does it mean to have $\lim_{t\to 0} X_t \neq 0$? Negating the definition of a limit, this means you can find $\varepsilon > 0$ and a sequence $t_n \to 0$ in $\mathbb{R}_{>0}$ such that $|X_{t_n}| \ge \varepsilon$ for every $n \in \mathbb{N}$. But the point is that these t_n's can be approximated by positive rationals — you can always find a positive rational arbitrarily close to this t_n.

And our limit over rationals was 0 almost surely, which means you *can* find $\delta > 0$ such that $|X_t| < \varepsilon/2$ for every $t \in \mathbb{Q}_+$ with $t < \delta$ (by the definition of the earlier limit).

And now we have a contradiction. Why? If you take $n \in \mathbb{N}$ such that $t_n < \delta$ (we can do that because $t_n \to 0$), then we can find $q \in \mathbb{Q}_+$ with $q < \delta$ such that $|X_{t_n} - X_q| < \varepsilon/2$, because X is continuous at t_n and $\mathbb{Q}$ is dense in $\mathbb{R}$. By the triangle inequality, this means $|X_{t_n}| \le |X_{t_n} - X_q| + |X_q|$. But the first term is less than $\varepsilon/2$ by the choice of $\mathbb{Q}$. And for the second term, since $q < \delta$ is rational, this is also less than $\varepsilon/2$. So we get $|X| < \varepsilon$, which is a contradiction.

So indeed the limit has to be 0. □

Student Question. *Why is the covariances being equal enough to mean the covariances are the same?*

Answer. Because if you have a vector of Gaussians, then the law is determined by the covariance matrix. The fact

$$\operatorname{Cov}(B_{t_{j-1}}, B_{t_j}) = t_{j-1}$$

determines the covariance matrix of $(B_{t_1}, \ldots, B_{t_n})$; and we showed the same is true of X, so they have

> the same covariance matrix, and therefore the same law. (Here the right thing to do is actually to consider t_j and t_k, rather than just t_{j-1} and t_j.)

(Brownian motion will not be on the exam. Everything else can be. All the material is in summaries, so we know exactly what was covered, and what he can or cannot ask. For the proofs, we should look at the proofs of basic results; we won't be asked for Cramer's theorem's proof, as that's impossible. We should also pay attention to the problem sets; this might be helpful. Some of the problems might be very similar. But there will be some problems which cannot necessarily be derived from problem sets. The exam is arranged so 40% is bookwork, 20 or 25% is some variation of a problem from psets, and 25% will be something different that you have to think a bit more for.)