Math 6710 lecture notes

Nate Eldredge

November 29, 2012

Caution! These lecture notes are very rough. They are mainly intended for my own use during lecture. They almost surely contain errors and typos, and tend to be written in a stream-of-consciousness style. In many cases details, precise statements, and proofs are left to Durrett's text, homework or presentations. But perhaps these notes will be useful as a reminder of what was done in lecture. If I do something that is substantially different from Durrett I will put it in here.

Thursday, August 23

Overview: Probability: study of randomness. Questions whose answer is not "yes" or "no" but a number indicating probability of "yes". Basic courses: divide into "discrete" and "continuous", and are mainly restricted to "finite" or "short term" problems - involving a finite number of events or random variables. To escape these bounds: measure theory (introduced to probability by Kolmogorov). Unify discrete/continuous, and enable the study of long term or limiting properties. Ironically, many theorems will be about how randomness disappears or is constrained in the limit.

1 Intro

1. Basic objects of probability: events and their probabilities, combining events with logical operations, random variables: numerical quantities, statements about them are events. Expected values.

2. Table of measure theory objects: measure space, measurable functions, almost everywhere, integral,. Recall definitions. Ex: Borel/Lebesgue measure on $\mathbb{R}^n$.

3. What's the correspondence? Sample space: all possible "outcomes" of an "experiment", or "states of the world". Events: set of outcomes corresponding to "event happened". σ-field $\mathcal{F}$: all "reasonable" events—measurable sets. Logic operations correspond to set operations. Random variables: measurable functions, so "statement about it" i.e pullback of Borel set, is an event.

4. 2 coin flip example.

5. Fill in other half of the table. Notation matchup.

6. Suppressing the sample space

Tuesday, August 28

2 Random variables

2.1 Random vectors

Definition 2.1. A n-dimensional random vector is a measurable map $\mathbf{X} : \Omega \to \mathbb{R}^n$. (As usual $\mathbb{R}^n$ is equipped with its Borel σ-field.)

Fact: If $\mathbf{X}$ has components $\mathbf{X} = (X_1, \ldots, X_n)$ then $\mathbf{X}$ is a random vector iff the X_i are random variables.

More generally we can consider measurable functions X from Ω to any other set S equipped with a σ-field $\mathcal{S}$ (measurable space). Such an X could be called an S-valued random variable. Examples:

- Set (group) of permutations of some set. For instance when we shuffle a deck of cards, we get a random permutation, which could be considered an S_{52}-valued random variable.
- A manifold with its Borel σ-field. Picking random points on the manifold.
- Function spaces. E.g. $C([0, 1])$ with its Borel σ-field. Brownian motion is a random continuous path which could be viewed as a $C([0, 1])$-valued random variable.

Many results that don't use the structures of $\mathbb{R}^n$ (e.g. arithmetic, ordering, topology) in any obvious way will extend to any S-valued random variable. In some cases problems can arise if the σ-field $\mathcal{S}$ is bad. However good examples are Polish spaces (complete separable metric spaces) with their Borel σ-fields. It turns out that these measurable spaces behave just like $\mathbb{R}$ and so statements proved for $\mathbb{R}$ (as a measurable space) work for them as well.

2.2 Sequences of random variables

If $X_1, X_2, \ldots$ is a sequence of random variables then:

- $\limsup_{n\to\infty} X_n$ and $\liminf X_n$ are random variables. (To prove, use the fact that X is a random variable, i.e. measurable, iff $\{X < a\} = X^{-1}((-\infty, a)) \in \mathcal{F}$ for all $a \in \mathbb{R}$. This is because the sets $\{(-\infty, a)\}$ generate $\mathcal{B}_{\mathbb{R}}$.)
- $\{\lim X_n \text{ exists}\}$ is an event. (It is the event $\{\limsup X_n = \liminf X_n\}$. Or argue directly.) If this event has probability 1 we say $\{X_n\}$ converges almost surely, or $X_n \to X$ a.s. (In this case the limit X is a random variable because it equals the lim sup.)

2.3 Distributions

Most important questions about X are "what values does it take on, with what probabilities"? Here's an object that encodes that information.

Definition 2.2. The distribution (or law) of X is a probability measure μ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ defined by $\mu(B) = P(X \in B)$. In other notation $\mu = P \circ X^{-1}$. It is the pushforward of P onto $\mathbb{R}$ by the map X. We write $X \sim \mu$. Easy to check that μ is fact a probability measure.

Note μ tells us the probability of every event that is a question about X.

Note: Every probability measure μ on $\mathbb{R}$ arises as the distribution of some random variable on some probability space. Specifically: take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}_{\mathbb{R}}$, $P = \mu$, $X(\omega) = \omega$.

Example 2.3. $\mu = \delta_c$ a point mass at c (i.e. $\mu(A) = 1$ if $c \in A$ and 0 otherwise). Corresponds to a constant r.v. $X = c$ a.s.

Example 2.4. Integer-valued (discrete) distributions: $\mu(B) = \sum_{n \in B \cap \mathbb{Z}} p(n)$ for some probability mass function $p : \mathbb{Z} \to [0, 1]$ with $\sum_n p(n) = 1$. Such X takes on only integer values (almost surely) and has $P(X = n) = p(n)$. Other notation: $\mu = \sum_n p(n)\delta_n$.

- Bernoulli: $p(1) = p$, $p(0) = 1 - p$.
- Binomial, Poisson, geometric, etc.

Example 2.5. Continuous distributions: $\mu(B) = \int_B f\, dm$ for some $f \geq 0$ with $\int_{\mathbb{R}} f\, dm = 1$. f is called the density of the measure μ. Radon–Nikodym theorem: μ is of this form iff $m(B) = 0$ implies $\mu(B) = 0$.

- Uniform distribution $U(a, b)$: $f = \frac{1}{b-a} 1_{[a,b]}$
- Normal distribution $N(\mu, \sigma^2)$ (note different μ): $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$
- Exponential, gamma, chi-square, etc.

2.4 Distribution function (CDF)

Definition 2.6. Associated to each probability measure μ on $\mathbb{R}$ (and hence each random variable X) is the (cumulative) distribution function $F_\mu(x) := \mu((-\infty, x])$. We write F_X for F_μ where $X \sim \mu$; then $F_X(x) = P(X \leq x)$.

Fact: F is monotone increasing and right continuous; $F(-\infty) = 0$, $F(+\infty) = 1$. (*)

Proposition 2.7. *F_μ uniquely determines the measure μ.*

We'll use the proof of this as an excuse to introduce a very useful measure-theoretic tool: Dynkin's π-λ lemma. We need two ad-hoc definitions.

Definition 2.8. Let Ω be any set. A collection $\mathcal{P} \subset 2^\Omega$ is a π-system if for all $A, B \in \mathcal{P}$ we have $A \cap B \in \mathcal{P}$ (i.e. closed under intersection). A collection $\mathcal{L} \subset 2^\Omega$ is a λ-system if:

1. $\Omega \in \mathcal{L}$;
2. If $A, B \in \mathcal{L}$ with $A \subset B$ then $B \setminus A \in \mathcal{L}$ (closed under subset subtraction)
3. If $A_n \in \mathcal{L}$ where $A_1 \subset A_2 \subset A_3 \subset \ldots$ and $A = \bigcup A_n$ (we write $A_n \uparrow A$) then $A \in \mathcal{L}$ (closed under increasing unions).

Theorem 2.9. *If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system, and $\mathcal{P} \subset \mathcal{L}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

Proof: See Durrett A.1.4. It's just some set manipulations and not too instructive.
Here's how we'll prove our proposition:

Proof. Suppose μ, ν both have distribution function F. Let

$$\mathcal{P} = \{(-\infty, a] : a \in \mathbb{R}$$
$$\mathcal{L} = \{B \in \mathcal{B}_\mathbb{R} : \mu(B) = \nu(B)\}.$$

Notice that $\mathcal{P}$ is clearly a π-system and $\sigma(\mathcal{P}) = \mathcal{B}_\mathbb{R}$. Also $\mathcal{P} \subset \mathcal{L}$ by assumption, since $\mu((-\infty, a]) = F(a) = \nu((-\infty, a])$. So by Dynkin's lemma, if we can show $\mathcal{L}$ is a λ-system we are done.

1. Since μ, ν are both probability measures we have $\mu(\mathbb{R}) = \nu(\mathbb{R}) = 1$ so $\mathbb{R} \in \mathcal{L}$.

2. If $A, B \in \mathcal{L}$ with $A \subset B$, we have $\mu(B \setminus A) = \mu(B) - \mu(A)$ by additivity (B is the disjoint union of A and $B \setminus A$), and the same for ν But since $A, B \in \mathcal{L}$ we have $\mu(B) = \nu(B)$, $\mu(A) = \nu(A)$. Thus $\mu(B \setminus A) = \nu(B \setminus A)$.

3. If $A_n \uparrow A$, then it follows from countable additivity that $\mu(A) = \lim \mu(A_n)$. (This is called continuity from below and can be seen by writing $B_n = A_n \setminus A_{n-1}$, so that A is the disjoint union of the B_n.) Likewise $\nu(A) = \lim \nu(A_n)$. But $\mu(A_n) = \nu(A_n)$ so we must have $\mu(A) = \nu(A)$.

□

This is a typical application. We want to show some property Q holds for all sets in some σ-field $\mathcal{F}$. So we show that the collection of all sets with the property Q (whatever it may be) is a λ-system. Then we find a collection of sets $\mathcal{P}$ for which we know that Q holds, and show that it's a π-system which generates $\mathcal{F}$.

A similar, simpler technique that you have probably used before: show that the collection of all sets with property Q is a σ-field. Then find a collection of sets for which Q holds and show that it generates $\mathcal{F}$. However this won't work for this problem. In general, if μ, ν are two probability measures, the collection of all B with $\mu(B) = \nu(B)$ is always a λ-system but need not be a σ-field.

Fact: Any F satisfying the conditions (*) above is in fact the cdf of some probability measure μ on $\mathbb{R}$ (which as we have shown is unique). Proof: Use Caratheodory extension theorem (see appendix). Given F it's obvious how to define μ on a half-open interval $(a, b]$ (as $F(b) - F(a)$) and hence on any finite disjoint union of intervals. Call this latter collection $\mathcal{A}$; it's an algebra. Check that μ is countably additive on $\mathcal{A}$. Caratheodory says that μ extends to a countably additive measure on $\sigma(\mathcal{A})$ which is of course $\mathcal{B}_{\mathbb{R}}$. (The extension is constructed as an outer measure: $\mu(B) = \inf\{\mu(A) : A \in \mathcal{A}, B \subset A\}$.)

Notation: If two random variables X, Y have the same distribution, we say they are identically distributed and write $X \stackrel{d}{=} Y$.

Thursday, August 30

2.5 Joint distribution

If $\mathbf{X}$ is an n-dimensional random vector, its distribution is a probability measure on $\mathbb{R}^n$ defined in the same way. (It's possible to define multi-dimensional cdfs but messier and I am going to avoid it.)

Given random variables $X_1, \ldots, X_n$, their joint distribution is the probability measure on $\mathbb{R}^n$ which is the distribution of the random vector $(X_1, \ldots, X_n)$. This is not determined solely by the distributions of the X_n! Dumb example: $X = \pm 1$ a coin flip, $Y = -X$. Then both X, Y have the distribution $\mu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$. But (X, X) does not have the same joint distribution as (X, Y). For instance, $P((X, X) = (1, 1)) = 1/2$ but $P((X, Y) = (1, 1)) = 0$.

Moral: The distribution of X will let you answer any question you have that is only about X itself. If you want to know how it interacts with another random variable then you need to know their joint distribution.

3 Expectation

3.1 Definition, integrability

Expectation EX or $E[X]$.

Defining:

1. For $X = 1_A$, obvious: $EX = P(A)$.

2. For X simple, define via linearity.

3. For $X \geq 0$, $EX = \lim EX_n$ for simple $X_n \uparrow X$. (The limit could be infinite. Here we allow $X \in [0, \infty]$.) Fact: The value of the limit does not depend on the sequence of X_n chosen.

4. For real-valued X, we set $EX = EX^+ - EX^-$ provided both terms are finite. This happens iff $E|X| < \infty$, and in this case we say X is *integrable*.

Remark 3.1. In general we will only write EX when X is known to be either nonnegative (in which case $EX \in [0, \infty]$) or integrable (in which case $EX \in (-\infty, \infty)$). Of course we could also consider nonpositive random variables, or random variables for which at least one of EX^+, EX^- is finite. Usually we won't bother because it will be obvious how to extend any statement to handle these extra cases.

Unfortunate terminology: the expectation (i.e. integral) of X may exist (with a value of $\pm\infty$) even when X is not "integrable".

3.2 Inequalities

Theorem 3.2. *Jensen's inequality, Durrett Theorem 1.5.1. If φ is convex, then $\varphi(EX) \leq E\varphi(X)$. (Costandino will present.)*

Remark 3.3. Examples of convex functions: $|t|$ (useful for remembering which way the inequalities go), $\exp(t)$, $|t|^p$ for $1 \leq p < \infty$, any C^2 function φ with $\varphi'' \geq 0$.

Definition 3.4. The L^p norm of a random variable X is $\|X\|_p := E[|X|^p]^{1/p}$. (Could be infinite.) L^p is the space of all random variables with finite L^p norm.

Theorem 3.5. *Hölder's inequality: If $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$, then $E|XY| \leq \|X\|_p \|Y\|_q$.*

Proof. First, if $\|X\|_p = 0$ then $X = 0$ a.s. (homework) and the inequality is trivial; likewise if $\|Y\|_q = 0$.

We have the following inequality for all nonnegative real numbers x, y:

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}. \tag{1}$$

(Calculus exercise: find where the difference is maximum.) Now take $x = |X|/\|X\|_p$, $y = |Y|/\|Y\|_q$, and take expectations:

$$\frac{1}{\|X\|_p \|Y\|_q} E|XY| \leq \frac{E[|X|^p]}{p\|X\|_p^p} + \frac{E|Y|^q}{q\|Y\|_q^q} = \frac{1}{p} + \frac{1}{q} = 1. \tag{2}$$

□

Special cases:

1. Take $p = q = 2$; this is the Cauchy–Schwarz inequality.

2. Take $Y = 1$; this says $E|X| \leq \|X\|_p$. In particular, if $E|X|^p < \infty$ then $E|X| < \infty$, i.e. $L^p \subset L^1$. (We can also get this from Jensen.) Note it is essential that we are using a finite measure!

3. Let $1 \leq r \leq r'$; take $X = |Z|^r$, $p = r'/r$, $Y = 1$; this says $\|Z\|_r \leq \|Z\|_{r'}$, and in particular $L^{r'} \subset L^r$. (More moments is better.)

Theorem 3.6. *Markov inequality: For $X \geq 0$, $P(X \geq a) \leq \frac{1}{a}EX$. (An integrable random variable has a small probability of taking on large values.)*

Remark 3.7. Due to Chebyshev. Named in accordance with Stigler's law (due to Merton).

Proof. Trivial fact: $X \geq a1_{\{X \geq a\}}$ (draw picture: a box under a hump). Now take expectations and rearrange. □

Corollary 3.8. *Chebyshev's inequality (due to Bienaymé): If $X \in L^2$, then $P(|X - EX| \geq a) \leq \frac{1}{a^2}\operatorname{Var}(X)$. Recall $\operatorname{Var}(X) = E[(X - EX)^2] = E[X^2] - (EX)^2$ which exists whenever $X \in L^2$. This says an L^2 random variable has a small probability of being far from its mean.*

Proof. Apply Markov with $X \to (X - EX)^2$, $a \to a^2$. □

3.3 In terms of distribution

Theorem 3.9 (Change of variables)**.** *Let X be a random variable with distribution μ. For any measurable $f : \mathbb{R} \to \mathbb{R}$,*

$$Ef(X) = \int_{\mathbb{R}} f \, d\mu \tag{3}$$

where the expectation on the left exists iff the integral on the right does.

Proof. Will's presentation. □

Remark 3.10. The same holds for random vectors. This is an illustration of the principle that any statement about a single random variable should only depend on its distribution (and any statement about several should only depend on their joint distribution).

Tuesday, September 4

4 Modes of convergence

4.1 Almost sure, L^p

Definition 4.1. $X_n \to X$ a.s. means what it says. I.e., $P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$. Idea: In the long run, X_n is *guaranteed* to be close to X. (Note this is a pointwise statement: the speed of convergence can depend arbitrarily on ω. We rarely deal with uniform convergence of random variables.)

Definition 4.2. $X_n \to X$ in L^1 means $E|X_n - X| \to 0$. Idea: In the long run, X_n is *on average* close to X.

Note by the triangle inequality, if $X_n \to X$ in L^1 then $EX_n \to EX$. (Expectation is a continuous linear functional on L^1.)

Example 4.3. Let $U \sim U(0, 1)$, and set

$$X_n = \begin{cases} n, & U \leq \frac{1}{n} \\ 0, & \text{otherwise.} \end{cases}$$

Then $X_n \to 0$ a.s. but $EX_n = 1$ so $X_n \not\to 0$ in L^1.

Definition 4.4. $X_n \to X$ in L^p means $E[|X_n - X|^p] \to 0$. A sort of weighting: places where X_n and X are far apart contribute more to the average when p is bigger.

Consequence of Hölder: if $p \le p'$ and $X_n \to X$ in $L^{p'}$ then $X_n \to X$ in L^p. In particular L^p convergence implies L^1 convergence.

4.2 Big 3 convergence theorems

When does $X_n \to X$ a.s. imply $EX_n \to EX$?

Theorem 4.5. *Monotone convergence theorem*

Theorem 4.6. *Dominated convergence theorem*

Corollary 4.7. *Bounded convergence theorem (use 1 as dominating function)*

Corollary 4.8. *DCT gives you L^1 convergence. If dominating function is L^p then you get L^p convergence. (Homework.)*

Theorem 4.9. *Fatou's lemma. (You can only lose mass in the limit, not gain it.)*

Corollary 4.10. *If $X_n \to X$ a.s. and $E|X_n| \le C$ then $E|X| \le C$.*

4.3 Convergence i.p.

Definition 4.11. We say $X_n \to X$ in probability (i.p.) if for all $\epsilon > 0$, $P(|X_n - X| \ge \epsilon) \to 0$ or equivalently $P(|X_n - X| < \epsilon) \to 1$. Idea: In the long run, X_n is *very likely* to be close to X.

Proposition 4.12. *If $X_n \to X$ a.s. then $X_n \to X$ i.p.*

Proof. If $X_n \to X$ a.s., then for any ϵ we have, almost surely, $|X_n - X| < \epsilon$ for sufficiently large n. That is, $P(\liminf_{n\to\infty}\{|X_n - X| < \epsilon\}) = 1$. By homework, $\liminf_{n\to\infty} P(|X_n - X| < \epsilon) \ge P(\liminf_{n\to\infty}\{|X_n - X| < \epsilon\}) = 1$. Of course since these are probabilities the limsup has to be at most 1, so the limit exists and is 1. □

Proposition 4.13. *If $X_n \to X$ in L^p for any $p \ge 1$ then $X_n \to X$ i.p.*

Proof. Using Chebyshev,

$$P(|X_n - X| \ge \epsilon) = P(|X_n - X|^p \ge \epsilon^p) \le \frac{1}{\epsilon^p} E|X_n - X|^p \to 0. \tag{4}$$

□

Example 4.14. Let Y_n be uniformly distributed on $\{1, \dots, n\}$ (i.e. $P(Y_n = k) = 1/n$, $k = 1, \dots, n$). (We don't care about their joint distribution; they don't have to be independent.) Consider the triangular array of random variables $X_{n,k}$, $1 \le k \le n$, defined by $X_{n,k} = 1_{\{Y_n = k\}}$. Think of this as a sequence of random variables $X_{1,1}, X_{2,1}, X_{2,2}, X_{3,1}, \dots$ where we traverse the rows of the array one by one. (If you like you could take $\tilde{X}_m = X_{n,k}$ where $m = \binom{n}{2} + k$.) Take $X = 0$. Note that for any $\epsilon < 1$, we have $P(|X_{n,k} - X| \ge \epsilon) = P(X_{n,k} = 1) = P(Y_n = k) = 1/n \to 0$, thus $X_{n,k} \to 0$ i.p. We also have $E|X_{n,k}|^p = 1/n$ so $X_{n,k} \to 0$ in L^p for all p. But with probability 1, the sequence $X_{n,k}$ contains infinitely many 1s and also infintely many 0s, so $X_{n,k}$ does not converge a.s.

In this case, the "large" discrepancies (i.e. events when $X_{n,k} = 1$) become less and less likely to occur at any given time, and their average effect is also becoming small. But they still happen infinitely often so almost sure convergence is impossible.

Proposition 4.15. *Suppose $X_n \to X$ i.p. and $f : \mathbb{R} \to \mathbb{R}$ is continuous. Then $f(X_n) \to f(X)$ i.p. (The corresponding statement for a.s. convergence is obvious.)*

Proof. Proved in Durrett Theorem 2.3.4 using the double subsequence trick. Homework: give a direct proof. □

4.4 Borel–Cantelli

This is our main tool for proving almost sure convergence.

Theorem 4.16 (First Borel–Cantelli lemma)**.** *Let $A_1, A_2, \ldots$ be any sequence of events. Suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then $P(\limsup A_n) = 0$.*

Proof. By continuity from below, we have

$$P(\limsup A_n) = P\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n\right) = \lim_{m\to\infty} P\left(\bigcup_{n=m}^{\infty} A_n\right). \tag{5}$$

But by union bound

$$P\left(\bigcup_{n=m}^{\infty} A_n\right) \le \sum_{n=m}^{\infty} P(A_n). \tag{6}$$

Since $\sum_{n=1}^{\infty} P(A_n) < \infty$ we must have $\sum_{n=m}^{\infty} P(A_n) \to 0$ as $m \to \infty$. □

Corollary 4.17. *Suppose for any $\epsilon > 0$ we have $\sum_{n=1}^{\infty} P(|X_n - X| \ge \epsilon) < \infty$. Then $X_n \to X$ a.s. (Note this hypothesis looks similar to convergence i.p. except that we require $P(|X_n - X| \ge \epsilon)$ to go to zero a little bit faster. $1/n$ is not fast enough but $1/n^2$ or 2^{-n} is.)*

Proof. Take $\epsilon = 1/m$. Borel–Cantelli implies that $P(\limsup\{|X_n - X| \ge 1/m\}) = 0$. This says that, almost surely, $|X_n - X|$ is eventually less than $1/m$, i.e. $\limsup |X_n - X| < 1/m$. So let $B_m = \{\limsup |X_n - X| < 1/m\}$; we just showed $P(B_m) = 1$. If $B = \bigcap_m B_m$ then $P(B) = 1$ as well. But on B we have $\limsup |X_n - X| < 1/m$ for every m, which is to say $\limsup |X_n - X| = 0$, which is to say $X_n \to X$. □

4.5 The subsequence trick

Lemma 4.18. *If $X_n \to X$ i.p. then there is a subsequence $X_{n_k} \to X$ a.s.*

Proof. For less writing, let's assume $X = 0$ (by replacing X_n with $X_n - X$).

We construct the subsequence inductively. Let $n_1 = 1$. Suppose $n_1 < \cdots < n_{k-1}$ have been chosen. By convergence i.p. we have $P(|X_n| > 1/k) \to 0$, so we may choose n_k so large that for all $n > n_k$, we have $P(|X_n| > 1/k) < 2^{-k}$. If necessary, take n_k larger so that $n_k > n_{k-1}$.

Now we check this subsequence works. Let $\epsilon > 0$. Choose K so large that $1/K < \epsilon$. Then for any $k \ge K$ we have

$$P(|X_{n_k}| \ge \epsilon) \le P(|X_{n_k}| \ge 1/K) \le P(|X_{n_k}| \ge 1/k) \le 2^{-k}.$$

So by the comparison test we have $\sum_{k=1}^{\infty} P(|X_{n_k}| \ge \epsilon) < \infty$. Hence by our previous lemma, $X_{n_k} \to 0$ a.s. □

Thursday, September 6

This fact can often be used to take a theorem about a.s. converging sequences, and prove it under the weaker assumption of i.p. convergence. For instance:

Theorem 4.19 (Upgraded DCT)**.** *Suppose $X_n \to X$ i.p., and there exists Y with $|X_n| \le Y$ and $E|Y| < \infty$. Then $EX_n \to EX$ (and $X_n \to X$ in L^1).*

Proof. Suppose not. Then there is an $\epsilon > 0$ and a subsequence X_{n_m} such that $|EX_{n_m} - EX| > \epsilon$ for all m. We still have $X_{n_m} \to X$ i.p. so there is a further subsequence $X_{n_{m_k}}$ such that $X_{n_{m_k}} \to X$ a.s. We still have $|EX_{n_{m_k}} - EX| > \epsilon$ for all k. But our classic DCT applies to $X_{n_{m_k}}$ so $EX_{n_{m_k}} \to EX$. This is a contradiction. □

We can get upgraded MCT and Fatou in the same way.

4.6 Uniform integrability, Vitali

The biggest possible hammer for proving L^1 convergence is uniform integrability.

Definition 4.20. A set $\mathcal{S}$ of random variables is *uniformly integrable* (ui) if for every $\epsilon > 0$ there exists $M > 0$ such that for every $X \in \mathcal{S}$ we have $E[|X|; |X| \ge M] \le \epsilon$.

(Notation: $E[X; A]$ means $E[X1_A]$ or if you like $\int_A X\,dP$. It should not be confused with the conditional expectation $E[X \mid A]$ which we will discuss later, although it is related.)

Lemma 4.21. *If $\mathcal{S}$ is ui then* $\sup_{X\in\mathcal{S}} E|X| < \infty$.

Proof. Take $\epsilon = 17$ and choose M so large that $E[|X|; |X| \ge M] \le 17$ for all $X \in \mathcal{S}$. Then $E|X| = E[|X|; X \ge M] + E[|X|; |X| < M] \le 17 + M$. □

Lemma 4.22. *If $X \in L^1$ then $\{X\}$ is ui.*

Proof. Suppose $X \in L^1$. Set $X_n = |X|1_{\{|X|\ge n\}}$. Then $X_n \to 0$ a.s. and $|X_n| \le |X|$. So by DCT $EX_n \to 0$. Thus given any $\epsilon > 0$ we may choose M so large that $EX_M \le \epsilon$. But EX_M is exactly $E[|X|; X \ge M]$. □

Theorem 4.23 (Useful half of Vitali convergence theorem)**.** *Let X_n, X be random variables with $X_n \to X$ i.p. If $\{X_n\}$ is ui, then $X_n \to X$ in L^1.*

Proof. For this proof, let $\phi_M(x) = x \vee -M \wedge M$. (Draw a picture.) Note ϕ_M is continuous and bounded. Observe that $|X - \phi_M(X)| \le |X|1_{\{|X|\ge M\}}$.

Suppose $X_n \to X$ i.p. and $\{X_n\}$ is ui. We know that $\sup_n E|X_n| < \infty$; by upgraded Fatou we thus have $X \in L^1$, so $\{X\}$ is ui as well. Take $\epsilon > 0$. For any M we can write

$$E|X_n - X| \le E|X_n - \phi_M(X_n)| + E|\phi_M(X_n) - \phi_M(X)| + E[\phi_M(X) - X| \tag{7}$$

by the triangle inequality. By uniform integrability, we can take M so large that for all n,

$$E|X_n - \phi_M(X_n)| \le E[|X_n|; |X_n| \ge M] < \epsilon \tag{8}$$

and by taking M larger we can get the same to hold for X. So for such large M we have

$$E|X_n - X| \le 2\epsilon + E|\phi_M(X_n) - \phi_M(X)|.$$

By continuous mapping $\phi_M(X_n) \to \phi_M(X)$ i.p., and by bounded convergence also in L^1, so $E|\phi_M(X_n) - \phi_M(X)| \to 0$. Taking the limsup we have $\limsup E|X_n - X| \le 2\epsilon$. But ϵ was arbitrary, so $\limsup E|X_n - X| = 0$ and we are done. □

Remark 4.24. The converse of this statement is also true: if $X_n \to X$ in L^1 then $X_n \to X$ i.p. (this is just Chebyshev) and $\{X_n\}$ is ui. So the condition of ui for L^1 convergence is in some sense optimal because it is necessary as well as sufficient.

Remark 4.25. Vitali implies DCT since a dominated sequence is ui. (Homework.)

Lemma 4.26 (Crystal ball)**.** *Suppose for some $p > 1$ we have $\sup_{X \in S} E|X|^p < \infty$ (i.e. S is bounded in L^p). Then S is ui.*

Proof. Homework. □

5 Smaller σ-fields

Our probability space comes equipped with the σ-field $\mathcal{F}$ consisting of all "reasonable" events (those whose probability is defined). A feature which distinguishes probability from other parts of measure theory is that we also consider various sub-σ-fields $\mathcal{G} \subset \mathcal{F}$.

Interpretation: if an event A is a "question" or a piece of binary data about an experiment, a σ-field $\mathcal{G}$ is a collection of "information", corresponding to the information that can answer all the questions $A \in \mathcal{G}$. This makes sense with the σ-field axioms: if you can answer the questions A, B, then with simple logic you can answer the questions A^c ("not A") and $A \cup B$ ("A or B"). If you can answer a whole sequence of questions $A_1, A_2, \ldots$, you can answer the question $\bigcup A_n$ ("are any of the A_n true?").

We say a random variable X is $\mathcal{G}$-measurable if $X^{-1}(B) \in \mathcal{G}$ for every Borel $B \subset \mathbb{R}$. As an abuse of notation, we also write $X \in \mathcal{G}$. Idea: you have been given enough information from the experiment that you can deduce the value of the numerical data X.

Definition 5.1. If X is a random variable, $\sigma(X)$ denotes the smallest σ-field with respect to which X is measurable; we call it the σ-field generated by X. (As usual, "smallest" means that if $\mathcal{G}$ is a σ-field with $X \in \mathcal{G}$, then $\sigma(X) \subset \mathcal{G}$. This σ-field is obviously unique, and it also exists because it is the intersection of all sub-σ-fields $\mathcal{G}$ of $\mathcal{F}$ for which $X \in \mathcal{G}$. The intersection is nonempty because $X \in \mathcal{F}$ by definition of random variable.)

Think of $\sigma(X)$ as "all the information that can be learned by observing X".

Proposition 5.2. $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}_{\mathbb{R}}\}$.

Proof. Let $\mathcal{G}$ denote the collection on the right. Since X is $\sigma(X)$-measurable, we must have $X^{-1}(B) \in \sigma(X)$ for all Borel sets B. This proves $\mathcal{G} \subset \sigma(X)$. Conversely, since preimages preserve unions and complements, $\mathcal{G}$ is a σ-field, and since it contains $X^{-1}(B)$ for all Borel B, we have $X \in \mathcal{G}$. By the minimality of $\sigma(X)$, we must have $\sigma(X) \subset \mathcal{G}$. □

Intuitively, $Y \in \sigma(X)$ should mean "knowing X is enough to determine Y". The next proposition makes this explicit.

Proposition 5.3 (Doob-Dynkin lemma)**.** *$Y \in \sigma(X)$ if and only if there exists a measurable $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$.*

Proof. Presentation. □

(Note that such f is unique only if $X : \Omega \to \mathbb{R}$ is surjective.)

Given several random variables $X_1, \ldots, X_n$, we denote by $\sigma(X_1, \ldots, X_n)$ the smallest σ-field "containing" all of $X_1, \ldots, X_n$. (Fact: This is the same as the σ-field generated by the random vector $\mathbf{X} = (X_1, \ldots, X_n)$.) The same definition goes for an infinite set of random variables.

Tuesday, September 11

6 Independence

Definition 6.1. A sequence of events $A_1, A_2, \ldots$ (finite or infinite) is **independent** if, for any distinct indices $i_1, \ldots, i_n$, we have

$$P(A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) \ldots P(A_{i_n}). \tag{9}$$

To understand this, start with just two events. Then this just says $P(A_1 \cap A_2) = P(A_1)P(A_2)$. This may be best understood in terms of conditional probability: it says $P(A_1|A_2) = P(A_1)$, i.e. knowing that A_2 happened doesn't give you any information about whether A_1 happened, and doesn't let you improve your estimate of its probability.

More generally, for any $i_0, i_1, \ldots, i_n$ we have

$$P(A_{i_0}|A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_0}).$$

This says that if you want to know something about one of the events, A_{i_0}, then knowing that *any number of the other events* happened is irrelevant information.

This is *not* a pairwise statement; it is not sufficient that any two A_i, A_j are independent. We really need to be allowed to consider subsequences of any (finite) size. It could be that no *single* A_j is useful in learning about A_i, but maybe several of them together are. You can't look at just two of the A_i at a time. However, you can see from the definition that it *is* sufficient to only look at *finitely many* at a time.

Example 6.2. Flip two fair coins. Let A_1 be the event that the first coin is heads, A_2 the second coin is heads, A_3 the two coins are the same. Then any two of the A_i are independent, but all three together are not. If you want to know whether A_1, then neither A_2 nor A_3 by itself helps you at all, but A_2 and A_3 together answer the question completely.

Note we could extend this to infinite subsequences: if $A_1, A_2, \ldots$ are independent, then for any distinct $i_1, i_2, \ldots$ we have

$$P(\bigcap_{j=1}^{\infty} A_{i_j}) = \prod_{j=1}^{\infty} P(A_{i_j}).$$

(Use continuity from above.)

We can generalize this to sequences of collections of events C_n. Think of each C_n as some database of information (corresponding to the set of questions, i.e. events, that the database can answer). In practice the C_n will generally be σ-fields. Independence of the collections means that access to some of the databases won't give you any clues about the information contained in any of the others.

Definition 6.3. A (finite or infinite) sequence of collections of events $C_1, C_2, \cdots \subset \mathcal{F}$ is **independent** if for any distinct indices $i_1, \ldots, i_n$ and any choice of events $A_{i_1} \in C_{i_1}, \ldots, A_{i_n} \in C_{i_n}$, we have

$$P(A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) \ldots P(A_{i_n}).$$

Definition 6.4. We say random variables $X_1, X_2, \ldots$ are independent if the σ-fields $\sigma(X_1), \sigma(X_2), \ldots$ are independent in the sense of the previous definition.

To understand independence in other ways, this lemma will be very useful.

Lemma 6.5. *Let $A \in \mathcal{F}$ be any event, and let*

$$\mathcal{L}_A := \{B \in \mathcal{F} : P(A \cap B) = P(A)P(B)\}$$

be the collection of all events which are independent of A. Then $\mathcal{L}_A$ is a λ-system.

Proof. (Don't do in class?)

1. $P(A \cap \Omega) = P(A) = P(A)P(\Omega)$, so $\Omega \in \mathcal{L}_A$.

2. If $B_1, B_2 \in \mathcal{L}_A$ and $B_1 \subset B_2$, then we have

$$\begin{aligned} P(A \cap (B_2 \setminus B_1)) &= P((A \cap B_2) \setminus (A \cap B_1)) \\ &= P(A \cap B_2) - P(A \cap B_1) && \text{since } A \cap B_1 \subset A \cap B_2 \\ &= P(A)(P(B_2) - P(B_1)) \\ &= P(A)P(B_2 \setminus B_1). \end{aligned}$$

3. If $B_1 \subset B_2 \subset \ldots$ is an increasing sequence of events in $\mathcal{L}_A$, and $B = \bigcup B_n$ then

$$\begin{aligned} P(A \cap B) &= P(\bigcup_n (A \cap B_n)) \\ &= \lim P(A \cap B_n) \\ &= \lim P(A)P(B_n) \\ &= P(A)P(B). \end{aligned}$$

□

Corollary 6.6. *If C is any collection of events, then $\mathcal{L}_C := \{B \in \mathcal{F} : P(A \cap B) = P(A)P(B)$ for all $A \in C\}$ is a λ-system.*

Proof. $\mathcal{L}_C = \bigcap_{A \in C} \mathcal{L}_A$, and it is simple to check from the definition that an arbitrary intersection of λ-systems is another λ-system. □

Proposition 6.7. *If $\mathcal{P}_1, \mathcal{P}_2, \ldots$ are independent π-systems, then $\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \ldots$ are independent σ-fields.*

Proof. Let

$$\mathcal{P} = \left\{A_{i_1} \cap \cdots \cap A_{i_n} : A_{i_j} \in \mathcal{P}_{i_j},\ i_j \geq 2\right\}$$

be the collection of all finite intersections of events from $\mathcal{P}_2, \mathcal{P}_3, \ldots$. By the assumption of independence, for any $A \in \mathcal{P}_1$ and $B = A_{i_1} \cap \cdots \cap A_{i_n} \in \mathcal{P}$ we have

$$P(A \cap B) = P(A \cap A_{i_1} \cap \cdots \cap A_{i_n}) = P(A)P(A_{i_1}) \ldots P(A_{i_n}) = P(A)P(B).$$

So A, B are independent. This shows $A \in \mathcal{L}_{\mathcal{P}}$, so we have $\mathcal{P}_1 \subset \mathcal{L}_{\mathcal{P}}$. By the π-λ lemma we have $\sigma(\mathcal{P}_1) \subset \mathcal{L}_{\mathcal{P}}$, which is to say that $\sigma(\mathcal{P}_1), \mathcal{P}_2, \ldots$ are independent.

This is still a sequence of π-systems, so we can repeat the argument, using $\mathcal{P}_2$ instead of $\mathcal{P}_1$, to see that $\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \mathcal{P}_3, \ldots$ are independent. Indeed, for any n, we can repeat this n times, and learn that $\sigma(\mathcal{P}_1), \ldots, \sigma(\mathcal{P}_n), \mathcal{P}_{n+1}, \ldots$ are independent. Since the definition of independence only looks at a finite number of the $\mathcal{P}_i$ at a time, this in fact shows that all the σ-fields $\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \ldots$ are independent. □

Proposition 6.8. *Suppose $\mathcal{G}_1, \mathcal{G}_2, \dots$ are independent σ-fields. Let $\mathcal{H}_1 = \sigma(\mathcal{G}_{i_1}, \mathcal{G}_{i_2}, \dots)$, $\mathcal{H}_1 = \sigma(\mathcal{G}_{j_1}, \mathcal{G}_{j_2}, \dots)$, where the (finite or infinite) sets of indices $\{i_1, i_2, \dots\}$ and $\{j_1, j_2, \dots\}$ are disjoint. Then $\mathcal{H}_1, \mathcal{H}_2$ are independent σ-fields.*

Intuitively: independence of $\mathcal{G}_1, \mathcal{G}_2, \dots$ means that information from any of the $\mathcal{G}_i$ is irrelevant to any of the others. This says even when you pool together information from several of the $\mathcal{G}_i$, it's irrelevant to information pooled from any of the remaining $\mathcal{G}_j$.

Proof. As in the previous proof, let $\mathcal{P}_1$ be the collection of all finite intersections of events from the $\mathcal{G}_i$, and $\mathcal{P}_2$ the same for the $\mathcal{G}_j$. $\mathcal{P}_1, \mathcal{P}_2$ are independent π-systems, so $\mathcal{H}_1 = \sigma(\mathcal{P}_1), \mathcal{H}_2 = \sigma(\mathcal{P}_2)$ are independent σ-fields. □

We could extend this by partitioning the sequence of σ-fields any way we like (rather than just into 2 groups); these disjoint subsets will generate independent σ-fields.

It's worth noting that when dealing with

7 Independent random variables

Proposition 7.1. *If $X_1, X_2, \dots$ are independent, and $f_1, f_2, \dots$ are measurable functions, then $f_1(X_1), f_2(X_2), \dots$ are independent.*

Proof. Since $f_i(X_i)$ is $\sigma(X_i)$-measurable, we have $\sigma(f_i(X_i)) \subset \sigma(X_i)$. This implies that $\sigma(f_1(X_1)), \sigma(f_2(X_2)), \dots$ are independent. □

Independence of random variables can be expressed in terms of their joint distribution. Recall the following facts about product measures:

Definition 7.2. Let $(S_1, \mathcal{S}_1), \dots, (S_n, \mathcal{S}_n)$ be measurable spaces, and let $S = S_1 \times \cdots \times S_n$. The product σ-field $\mathcal{S}$ on S is the σ-field generated by all "rectangles" of the form $A_1 \times \cdots \times A_n$, $A_i \in \mathcal{S}_i$; we abuse notation and write $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$.

Theorem 7.3. *If $\mu_1, \dots, \mu_n$ are finite measures on $S_1, \dots, S_n$ respectively, there is a unique measure μ on the product σ-field of S which satisfies $\mu(A_1 \times \cdots \times A_n) = \mu_1(A_1) \dots \mu_n(A_n)$. μ is called the **product measure** of $\mu_1, \dots, \mu_n$ and we write $\mu = \mu_1 \times \cdots \times \mu_n$.*

This is a standard theorem of measure theory and I won't prove it here. The existence is an application of the Carathéodory extension theorem; the uniqueness follows immediately using the π-λ theorem (the collection of rectangles is a π-system which generates $\mathcal{S}$.)

Also recall the Fubini–Tonelli theorem: suppose $\mu = \mu_1 \times \cdots \times \mu_n$ is a product measure on $(S, \mathcal{S})$ and $f : S \to \mathbb{R}$ is measurable. If either $f \geq 0$ or $\int |f| \, d\mu < \infty$, then we have

$$\int f \, d\mu = \int_{S_1} \dots \int_{S_n} f(x_1, \dots, x_n) \mu_n(dx_n) \dots \mu_1(dx_1)$$

and the integrals on the right may be interchanged at will. (Note that it is often useful to apply the nonnegative case to $|f|$ to verify that $\int |f| \, d\mu < \infty$.

Theorem 7.4. *Suppose $X_1, \dots, X_n$ are random variables with distributions $\mu_1, \dots, \mu_n$. Then $X_1, \dots, X_n$ are independent if and only if their joint distribution $\mu_{X_1, \dots, X_n}$ is the product measure $\mu_1 \times \cdots \times \mu_n$ on $\mathbb{R}^n$.*

Proof. (Worth doing in class?) By Proposition 5.2 the events in $\sigma(X_i)$ are precisely those of the form $\{X_i \in B\}$ for Borel sets B.

Suppose $X_1, \dots, X_n$ are independent. Let $B_1, \dots, B_n$ be Borel subsets of $\mathbb{R}$. Then the events $\{X_1 \in B_1\}, \dots, \{X_n \in B_n\}$ are in $\sigma(X_1), \dots, \sigma(X_n)$ respectively, hence are independent. So

$$\begin{aligned}\mu(B_1 \times \cdots \times B_n) &= P((X_1, \dots, X_n) \in B_1 \times \cdots \times B_n) \\ &= P(\{X_1 \in B_1\} \cap \cdots \cap \{X_n \in B_n\}) \\ &= P(X_1 \in B_1) \dots P(X_n \in B_n) \\ &= \mu_1(B_1) \dots \mu_n(B_n).\end{aligned}$$

Therefore $\mu = \mu_1 \times \cdots \times \mu_n$.

Conversely, suppose $\mu = \mu_1 \times \cdots \times \mu_n$. Let $A_1 \in \sigma(X_1), \dots, A_n \in \sigma(X_n)$. By Proposition 5.2 above, we must have $A_1 = \{X_1 \in B_1\}, \dots, A_n = \{X_n \in B_n\}$ for Borel sets $B_1, \dots, B_n$. Thus

$$\begin{aligned}P(A_1 \cap \cdots \cap A_n) &= P(\{X_1 \in B_1\} \cap \cdots \cap \{X_n \in B_n\}) \\ &= P((X_1, \dots, X_n) \in B_1 \times \cdots \times B_n) \\ &= \mu(B_1 \times \cdots \times B_n) \\ &= \mu_1(B_1) \dots \mu_n(B_n) \\ &= P(A_1) \dots P(A_n).\end{aligned}$$

□

Thursday, September 13

Proposition 7.5. *If $X_1, \dots, X_n$ are independent nonnegative random variables, then $E[X_1 \dots X_n] = E[X_1] \dots E[X_n]$. If $X_1, \dots, X_n$ are independent integrable random variables, then $X_1 \dots X_n$ is integrable and $E[X_1 \dots X_n] = E[X_1] \dots E[X_n]$. (Note that without independence, a product of integrable random variables need not be integrable.)*

Proof. We'll just write out the case $n = 2$. If X, Y are independent and nonnegative then we have

$$\begin{aligned}E[XY] &= \int_{\mathbb{R}^2} xy\mu_{X,Y}(dx, dy) && \text{change of variables} \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} xy\mu_X(dx)\mu_Y(dy) && \text{Tonelli} \\ &= \int_{\mathbb{R}} x\mu_X(dx) \int_{\mathbb{R}} y\mu_Y(dy) \\ &= E[X]E[Y].\end{aligned}$$

If instead X, Y are independent and integrable, then $|X|, |Y|$ are independent and nonnegative. By the previous case we have $E|XY| = E|X|E|Y| < \infty$. So $\int_{\mathbb{R}^2} |xy|\mu_{X,Y}(dx, dy) < \infty$ and we can use the same argument as above, with Fubini in place of Tonelli. □

Corollary 7.6. *If $X_1, \dots, X_n \in L^2$ are independent, then* $\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$.

Proof. Again we just do $n = 2$. If $X, Y \in L^2$, then by a simple computation we have $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, where $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$. But by the previous proposition we have $\mathrm{Cov}(X, Y) = 0$. □

8 Independence and limiting behavior

Notation 8.1. If $\mathcal{G}_1, \mathcal{G}_2$ are independent σ-fields we will write $\mathcal{G}_1 \perp\!\!\!\perp \mathcal{G}_2$; likewise for independent events or random variables. (We will not use this notation when more than two things are independent.)

Let us start with a result that looks very abstract but is actually profound.

Theorem 8.2 (Kolmogorov 0-1 law)**.** *Let $\mathcal{G}_1, \mathcal{G}_2, \dots$ be a sequence of independent σ-fields. Define the **tail σ-field***

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\mathcal{G}_n, \mathcal{G}_{n+1}, \dots).$$

Every event $A \in \mathcal{T}$ has probability 0 or 1.

Proof. Set $\mathcal{F}_n = \sigma(\mathcal{G}_1, \dots, \mathcal{G}_n)$. By Proposition 6.8, for any n, we have $\mathcal{F}_n \perp\!\!\!\perp \sigma(\mathcal{G}_{n+1}, \dots)$ are independent. But $\mathcal{T} \subset \sigma(\mathcal{G}_{n+1}, \dots)$ so $\mathcal{F}_n \perp\!\!\!\perp \mathcal{T}$. Now $\{\mathcal{F}_n\}$ is an increasing sequence of σ-fields, so by this week's homework, we have $\mathcal{T} \perp\!\!\!\perp \sigma(\mathcal{F}_1, \mathcal{F}_2, \dots) = \sigma(\mathcal{G}_1, \mathcal{G}_2, \dots)$. But $\mathcal{T} \subset \sigma(\mathcal{G}_1, \mathcal{G}_2, \dots)$ so we must actually have $\mathcal{T} \perp\!\!\!\perp \mathcal{T}$, i.e. $\mathcal{T}$ is independent of itself! In particular, every $A \in \mathcal{T}$ is independent of itself, so $P(A) = P(A \cap A) = P(A)^2$, which can only happen if $P(A)$ is 0 or 1. □

Corollary 8.3. *As shown in your homework, every random variable $X \in \mathcal{T}$ is almost surely constant, i.e. not really random.*

The events and random variables in $\mathcal{T}$ represent "long term behavior"; they are events that, for any n, don't depend on the short run behavior up to time n which is described by $\mathcal{G}_1, \dots, \mathcal{G}_n$. So the Kolmogorov zero-one law says that, given independence, long-term behavior is deterministic; there is no randomness in the limit.

At first glance it may not look like $\mathcal{T}$ contains anything except Ω and $\emptyset$ in which case this theorem would be trivial. But actually it contains many interesting events:

1. Suppose $A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2, \dots$. Then $\liminf A_n$ and $\limsup A_n$ are both in $\mathcal{T}$.

   To see this for lim inf, set $B_m = \bigcap_{n=m}^{\infty} A_n$; clearly $B_m \in \sigma(\mathcal{G}_m, \mathcal{G}_{m+1})$ since it is a countable intersection of events from this σ-field. Now by definition $\liminf A_n = \bigcup_{m=1}^{\infty} B_m$. But this is an increasing union; for any k we in fact have $\liminf A_n = \bigcup_{m=k}^{\infty} B_m$. This is a countable union of events in $\sigma(\mathcal{G}_k, \mathcal{G}_{k+1}, \dots)$, hence $\liminf A_n \in \sigma(\mathcal{G}_k, \mathcal{G}_{k+1}, \dots)$. But k was arbitrary, so we have $\liminf A_n \in \sigma(\mathcal{G}_k, \mathcal{G}_{k+1}, \dots)$ for every k, i.e. $\liminf A_n \in \mathcal{T}$.

   This should make intuitive sense; if I want to know if all but finitely many of the events A_n happened, for any given k, I don't need to know whether $A_1, \dots, A_k$ happened, because that is only finitely many events. This really is a long-run statement.

   The lim sup case is similar.

2. Suppose $X_1, X_2, \dots$ are random variables and let $\mathcal{G}_n = \sigma(X_n)$. Then the following events and random variables are in $\mathcal{T}$:

   (a) $\limsup_{n\to\infty} X_n$, $\liminf_{n\to\infty} X_n$

   (b) $\{\lim_{n\to\infty} X_n \text{ exists}\}$

   (c) $\lim X_n$ when it exists (since it equals the limsup and the liminf)

(d) $\left\{\sum_{n=1}^{\infty} X_n \text{ converges}\right\}$

On the other hand, the following apparently "limiting" objects are *not* in $\mathcal{T}$:

(a) $\sup_n X_n$, $\inf_n X_n$ (for instance, the first term could affect the supremum if it is larger than all the remaining terms)

(b) $\sum_{n=1}^{\infty} X_n$ when it exists. (For instance, think about the case that all $X_n \geq 0$, so the infinite sum definitely exists. The value of X_1 cannot affect whether or not the value of the sum is ∞, but it *can* affect whether the value is, say, 1 or 2.)

So, for instance, the Kolmogorov zero-one law says that if I have an *independent* sequence of random variables, there is no element of chance affecting whether or not X_n converges. Depending on the distributions of the X_n, either it almost surely converges, or almost surely diverges; there is no middle ground.

Example 8.4. Percolation. Consider the integer lattice $\mathbb{Z}^2$ in the plane, and produce a random graph by turning each edge on independently with probability p. You get a big graph. It is probably (certainly) disconnected, so it has a bunch of components (or clusters). What's the probability that one of those components is infinite (we say "percolation occurs")?

If you think about it, if I hide any finite number of the edges (say, those in some ball), you can still tell whether there's an infinite component or not; the presence or absence of any given finite set of edges can't change that event. So this is a tail event; by the Kolmogorov zero-one law it must have probability zero or one. Depending on the value of p, an infinite component is either guaranteed or impossible.

But which is it, for which values of p? Well, it's intuitively clear that increasing p should make it easier to have an infinite component, and this can be made rigorous with a coupling argument (explain?). So $P_p(\text{percolation})$ must jump from 0 to 1 at some "critical" value $p = p_c$, i.e. for all $p < p_c$ percolation does not happen, and for all $p > p_c$ it does.

This leaves two questions: what is the value of p_c, and what happens at p_c itself? In 1960 Harris showed there is no percolation at $p = 1/2$, so $p_c \geq 1/2$. The other direction was open for 20 years until Harry Kesten famously showed in 1980 that for the 2-dimensional integer lattice, $p_c \leq 1/2$. So the "critical threshold" is $1/2$, and percolation does not occur at this threshold.

What about higher dimensions, i.e. the integer lattice $\mathbb{Z}^d$? There's no reason to expect a nice value for p_c, but it's been estimated numerically by simulations, and asymptotics as $d \to \infty$ are also known. In particular it can be shown it is always strictly between 0 and 1. Is there percolation at p_c itself? It is conjectured the answer is no in every dimension. This is Harris and Kesten's result in $d = 2$, and Hara and Slade in the early 1990s showed it also holds for all $d \geq 19$. For $3 \leq d \leq 18$ this remains one of the most notorious open problems in probability today.

Let's think back to the situation of a sequence of events $A_1, A_2, \ldots$. The (first) Borel–Cantelli lemma gave us a sufficient condition to ensure $P(\limsup A_n) = 0$, i.e. almost surely only finitely many A_n happen; namely, that $\sum_n P(A_n) < \infty$. This didn't require anything about the relationship between the sets, and in general this sufficient condition is not necessary. (Example: let $U \sim U(0,1)$, $A_n = \{U < 1/n\}$. Then $P(A_n) = 1/n$ so $\sum P(A_n) = \infty$ but $P(\limsup A_n) = P(U = 0) = 0$.) But in the presence of independence this sufficient condition is also necessary.

Theorem 8.5 (Second Borel–Cantelli lemma)**.** *Let $A_1, A_2, \ldots$ be* independent *events. If $\sum_n P(A_n) = \infty$, then $P(\limsup A_n) = 1$.*

Proof. Yipu's presentation. □

This is sometimes also combined with the first Borel–Cantelli lemma and stated as the **Borel zero-one law** (what happened to Cantelli?):

Corollary 8.6 (Borel zero-one law)**.** *If $A_1, A_2, \dots$ are independent events, then $P(\limsup A_n)$ is 0 or 1, according to whether $\sum P(A_n) < \infty$ or $\sum P(A_n) = \infty$.*

Example 8.7. If you flip a fair coin infinitely many times, you will get infinitely many heads and infinitely many tails. (What else?)

Tuesday, September 18

9 Strong law of large numbers

Formally we define expectation as a Lebesgue integral, but surely that isn't how you would define it intuitively. If I asked you "what's the expected number of green M&M's in a bag", you'd go buy a bunch of bags, count the number of green ones in each bag, and average them. So if X_i is the number in bag i, you'd compute $\frac{1}{n}(X_1 + \cdots + X_n)$ for some large n, and you'd expect that would be a good approximation to the abstract "expected value".

If probability theory is going to be a useful mathematical machine, it had better agree with our intuition on this point: that expectation is a long-run average. Luckily it does, via the following fundamental theorem.

Theorem 9.1 (Strong law of large numbers)**.** *Let $X_1, X_2, \dots$ be independent and identically distributed (we write "iid" for short), and integrable. Then*

$$\frac{X_1 + \cdots + X_n}{n} \to E[X_1], \quad \textit{almost surely}.$$

(A "weak law" is a statement of this kind, in which the conclusion is only convergence in probability.)

It looks a bit weird that we have singled out X_1 to appear in the conclusion; but of course since the X_i are identically distributed, they all have the same expected value. So we mean that the averages of the iid random variables converge to their common expected value.

The above is the "optimal" version of the SLLN: the conclusion only uses the expectation of X_1, and we don't assume any higher moments. Proving this sharp version is a bit involved and I don't intend to do it. Anyway the full strength is rarely needed in practice; most of the random variables we meet in everyday life have many more finite moments, usually even all of them. So we'll prove some results assuming more integrability.

For the rest of this section, S_n is the sum $S_n = X_1 + \cdots + X_n$, so we are interested in the convergence of $\frac{S_n}{n}$.

Theorem 9.2 (L^2 WLLN)**.** *Let $X_1, X_2, \dots$ be independent, identically distributed, and L^2. Then $\frac{S_n}{n} \to E[X_1]$ in L^2 (hence also in probability).*

We are assuming more than the classic SLLN, i.e. L^2 instead of L^1, and getting only convergence ip. But the proof will be very easy.

Proof. For short, set $\mu = E[X_1]$, $\sigma^2 = \mathrm{Var}(X_1)$. Since the X_n are iid, we have $E[S_n] = n\mu$ and $\mathrm{Var}(S_n) = n\sigma^2$. Now

$$\left\| \frac{S_n}{n} - \mu \right\|_2^2 = E\left[\left(\frac{S_n}{n} - \mu \right)^2 \right] = \mathrm{Var}\left(\frac{S_n}{n} \right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \to 0.$$

□

The only place we used independence was the fact that $\mathrm{Var}(S_n) = n\,\mathrm{Var}(X_1)$, which follows from $E[X_iX_j] = E[X_i]E[X_j]$, $i \neq j$ or in other words $\mathrm{Cov}(X_i, X_j) = 0$. That is, we only used that the X_i are **uncorrelated**. In particular, it is sufficient for them to be *pairwise* independent.

Note that using Chebyshev's inequality with the above computation tells us that

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{\epsilon n}$$

which is nice because it gives you a quantitative bound.

Here's a strong law of large numbers, under even stronger hypotheses:

Theorem 9.3. *Suppose $X_1, X_2, \ldots$ are iid and L^4. Then $\frac{S_n}{n} \to E[X_1]$ almost surely.*

Proof. Evan's presentation; also Durrett 2.3.5 page 66. □

Here's the opposite extreme: we can't do without at least a first moment.

Theorem 9.4. *Suppose $X_1, X_2, \ldots$ are iid and $E|X_1| = \infty$, i.e. the X_i are not integrable. Then $P(\frac{S_n}{n}$ diverges$) = 1$.*

Proof. Lemuel's presentation; also Durrett Theorem 2.3.7, page 67. □

Note this proof has shown that if $X_n \geq 0$ and $EX_1 = \infty$, then $\frac{S_n}{n} \to +\infty$ almost surely. So again we have a result that applies to either integrable or nonnegative random variables.

This doesn't lose all hope for convergence in the non-integrable case; for instance, we could try putting something else in the denominator, which grows a bit faster than n. See Durrett Example 2.2.7 for an example where you use this idea to get convergence in probability, and estimate the "average size" of a non-integrable distribution (discuss in class?); see Theorem 2.5.9 for a proof that you cannot get almost sure convergence in such cases, no matter how you normalize.

Example 9.5 (The St. Petersburg paradox)**.** Consider the following game: we flip a coin repeatedly until heads comes up. If the first heads is on flip number k, you win 2^k dollars (so your winnings double with each tails). Thus if X is the amount you win, we have $P(X = 2^k) = 2^{-k}$ for $k \geq 1$. How much should you pay to get into such a game?

We can easily compute $E[X] = \sum_{k=1}^{\infty} 2^k P(X = 2^k) = \sum_{n=1}^{\infty} 1 = \infty$. So your expected winnings are infinite. Suppose you get to play repeatedly, i.e. $X_1, X_2, \ldots$ are iid with the distribution of X, and $S_n = X_1 + \cdots + X_n$ is your total winnings up to time n, we showed above that $\frac{S_n}{n} \to +\infty$ almost surely. If you pay an amount c for each play, your net winnings after time are $S_n - cn = n(\frac{S_n}{n} - c) \to \infty$ a.s.; you will eventually make back everything you paid.

On the other hand, this may not happen very fast. Say you pay $c = 200$ dollars for each play. You only profit if at least 7 tails are flipped, which happens only $1/128$ of the time. So the vast majority of plays will see you lose money. It's just that, every once in a great while, you'll win a huge amount that will, on average, make up for all those losses.

The point is that if you pay a fixed amount for each play, the total amount you paid grows linearly with time, while your winnings grow faster than linearly. How much faster? Durrett works out an example to show that $\frac{S_n}{n \lg n} \to 1$ i.p. So after a large number of plays, you have paid out cn dollars, and with high probability, you have won about $n \lg n$ dollars. Thus you break even after about 2^c plays. For $c = 200$ dollars that is a very long time.

Durrett's Theorem 2.5.9 shows that we cannot get almost sure convergence of $\frac{S_n}{n \lg n}$. For large n, S_n is very likely to be close to $n \lg n$, but there will be occasional excursions that take it further away.

10 Existence of probability spaces

We have been blithely talking about iid sequences of random variables and such. But a priori it is not clear that our development was not vacuous. Could there actually be a sequence of random variables on some probability space with that complicated set of properties? We certainly hope so because the idea of an iid sequence is so intuitive. But not just any old probability space $(\Omega, \mathcal{F}, P)$ will work. (For example, you can show that it won't work for Ω to be a countable set.) The natural spaces to use are infinite product spaces, which we will now discuss.

But first an easy example.

Theorem 10.1. *Let $\Omega = [0,1]$ and P be Lebesgue measure. Let $Y_n(\omega)$ be the nth bit in the binary expansion of ω. Then the random variables Y_n are iid Bernoulli.*

Proof. Diwakar presents. □

Remark 10.2. (Normal numbers) In other words, if you choose a number uniformly at random in $[0,1]$, the bits in its binary expansion look like fair coin flips. In particular, by the SLLN, asymptotically you will see equal numbers of 0 and 1 bits (in the sense that the fraction of the first n bits which are 0 goes to $1/2$ as $n \to \infty$). We say a number with this property is **normal in base 2**, and we've just shown that almost every number in $[0,1]$ is normal in base 2.

Of course, not every number has this property; for example, $1/2 = 0.1000\ldots_2$ is not normal in base 2. But $1/3 = 0.01010101\ldots_2$ is.

We could ask the same for other bases. $1/3 = 0.1000\ldots_3$ is not normal in base 3. But the same argument shows that almost every number is normal in base 3. Taking an intersection of measure 1 sets, almost every number is normal in bases 2 and 3 simultaneously.

In fact, taking a countable intersection, almost every number is normal in *every* base simultaneously, or in other words is a **normal number**. This is remarkable because, as far as I know, no explicit example of a normal number is known. Rational numbers are not normal (p/q is not normal in base q because its base q expansion terminates). It is a famous conjecture that π is normal but this has never been proved. So this is one of those cases in mathematics where it is hard to find a single example, but easy to show there must be lots of them.

For this section, $I = [0,1]$ denotes the unit interval.

Definition 10.3. $\mathbb{R}^{\mathbb{N}}$ is the infinite Cartesian product of $\mathbb{R}$ with itself. You can think of it as the set of all sequences of real numbers, or alternatively as the set of all functions $x : \mathbb{N} \to \mathbb{R}$; we will use the latter notation. We define $I^{\mathbb{N}}$ similarly.

Definition 10.4. A **cylinder set** is a subset of $\mathbb{R}^{\mathbb{N}}$ which is of the form $A = B \times \mathbb{R} \times \mathbb{R} \times \ldots$, where for some n we have $B \subset \mathbb{R}^n$ and B is Borel. That is, $x \in A$ iff $(x(1), \ldots, x(n)) \in B$.

We equip $\mathbb{R}^{\mathbb{N}}$ with the **infinite product σ-field** $\mathcal{B}^{\mathbb{N}}$ which is the σ-field generated by the cylinder sets. We remark that $\mathcal{B}^{\mathbb{N}}$ contains more than cylinder sets; for example it contains all products of Borel sets $B_1 \times B_2 \times \ldots$ (this is the countable intersection of the cylinder sets $B_1 \times \cdots \times B_n \times \mathbb{R} \times \ldots$).

We can use $\mathbb{R}^{\mathbb{N}}$ to talk about infinite joint distributions. Suppose $X_1, X_2, \ldots$ is an infinite sequence of random variables on some probability space $(\Omega, \mathcal{F}, P)$. Define a map $\mathbf{X} : \Omega \to \mathbb{R}^{\mathbb{N}}$ as

$$(\mathbf{X}(\omega))(k) = X_k(\omega)$$

where we are again thinking of $\mathbb{R}^{\mathbb{N}}$ as a space of functions. Or in terms of sequences, $\mathbf{X}(\omega) = (X_1(\omega), X_2(\omega), \dots)$. It is simple to check that $\mathbf{X}$ is measurable. Thus we can use it to push forward P to get a measure μ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$, which can be viewed as the joint distribution of the entire sequence $X_1, X_2, \dots$.

Much as the joint distribution of a finite number of random variables completely describes how they interact, the same is true for an infinite sequence. It even describes "infinite" properties. For example, suppose we have $X_n \to 3$ a.s. This says that the subset of $\mathbb{R}^{\mathbb{N}}$ consisting of sequences which converge to 3 (you can check this set is in $\mathcal{B}^{\mathbb{N}}$) has μ-measure 1.

Somewhat surprisingly, at this point we cease to be able to do all our work in the language of measure theory alone; we have to bring in some topology.

$\mathbb{R}^{\mathbb{N}}$ and $I^{\mathbb{N}}$ carry natural topologies: the product topology on $\mathbb{R}^n$ is *generated by* the sets of the form $U \times \mathbb{R} \times \mathbb{R} \times \dots$ where $U \subset \mathbb{R}^n$ for some n and U is open.[1] A better way to understand this topology is via convergent sequences: a sequence $x_1, x_2, \dots \in \mathbb{R}^{\mathbb{N}}$ converges to some x with respect to the product topology iff it converges pointwise, i.e. iff for every k we have $\lim_{n\to\infty} x_n(k) = x(k)$. Actually $\mathbb{R}^{\mathbb{N}}$ with the product topology is really a metric space, so we can do everything in terms of sequences. The metric is:

$$d(x, y) = \sum_{k=1}^{\infty} 2^{-k}(|x - y| \wedge 1). \tag{10}$$

(When dealing with $I^{\mathbb{N}}$ we can drop the $\wedge 1$ since it is redundant.)

It's worth mentioning that the Borel σ-field generated by the product topology on $\mathbb{R}^{\mathbb{N}}$ is in fact the product σ-field defined above.

Thursday, September 20

A very important fact from topology is:

Theorem 10.5 (Baby Tychonoff theorem)**.** *$I^{\mathbb{N}}$ is compact.*

This is a special case of Tychonoff's theorem which says that an arbitrary (finite, countable, or even uncountable) product of compact spaces is compact. But we can prove this special case without resorting to the full strength of Tychonoff, and without needing nets, ultrafilters, or the full axiom of choice / Zorn lemma. Anyway, this special case is all you ever "really" need. Pretty much every useful compactness property in real analysis follows from the compactness of $I^{\mathbb{N}}$; those that don't are usually just abstract nonsense.

We can prove baby Tychonoff with a picture. (I'll draw the picture in class but I'm too lazy to TeX it for these notes.)

Here's a reminder of some equivalent definitions of compactness in metric spaces:

Theorem 10.6. *Let (X, d) be a metric space. The following are equivalent:*

1. *Every open cover of X has a finite subcover. (The usual definition of compactness.)*
2. *If $\{E_j\}_{j\in J}$ is a family of closed sets, and for every $j_1, \dots, j_n \in J$ the finite intersection $E_{j_1} \cap \dots \cap E_{j_n}$ is nonempty, then $\bigcap_{j\in J} E_j$ is also nonempty. (This is just the "open cover" definition, after taking complements.)*
3. *Every sequence in X has a convergent subsequence. (The Bolzano–Weierstrass theorem.)*

[1] A previous version of these notes erroneously stated that *every* open set was of this form.

4. *X is complete (every Cauchy sequence converges) and totally bounded (for any $\epsilon > 0$, we can cover X with finitely many balls of radius ϵ).*

Topology and measure in $\mathbb{R}$ and similar spaces interact in interesting ways. Here is one that we will use.

Theorem 10.7. *Any probability measure μ on $\mathbb{R}^n$ is regular, i.e. for all Borel B and all $\epsilon > 0$, there exist closed F and open U with $F \subset B \subset U$ and $\mu(U \setminus F) < \epsilon$. F can even be taken compact.*

Proof. Homework. □

Now we have enough tools to prove a big theorem that lets us construct pretty much any conceivable measure on $I^{\mathbb{N}}$.

Theorem 10.8 (Kolmogorov extension theorem)**.** *For each n, suppose μ_n is a probability measure on I^n, and the μ_n are **consistent** in that for any Borel $B \subset I^n$, we have $\mu_n(B) = \mu_{n+1}(B \times I)$. Then there exists a probability measure μ on $I^{\mathbb{N}}$ (with its product σ-field) such that for each n and each Borel $B \subset I^n$, we have*

$$\mu_n(B) = \mu(B \times I \times I \times \dots).$$

"One measure to rule them all!" Or, in fancier words, a projective limit.

Corollary 10.9. *This also works if we replace I by $\mathbb{R}$.*

Proof. We first observe that we could do it for $(0, 1)$. Any probability measure μ_n on $(0, 1)^n$ extends to one $\tilde{\mu}_n$ on $[0, 1]^n$ in the obvious way (don't put any mass at the edges; or to say that in a fancier way, push the measure forward under the inclusion map). It is easy to see that if $\{\mu_n\}$ is a consistent sequence then so is $\{\tilde{\mu}_n\}$, and the Kolmogorov extension theorem gives us a probability measure $\tilde{\mu}$. One can now check that $(0, 1)^{\mathbb{N}}$ is a Borel subset of $[0, 1]^{\mathbb{N}}$, the restriction μ of $\tilde{\mu}$ to $(0, 1)^{\mathbb{N}}$ is a probability measure, and μ interacts with the μ_n as desired.

But $\mathbb{R}$ is homeomorphic to $(0, 1)$ so we can push measures back and forth between them (under a homeomorphism and its inverse) at will. □

Actually this would still work if we replaced $(0, 1)$ by any Borel subset $B \subset [0, 1]$, and $\mathbb{R}$ by any measurable space X for which there is a measurable $\phi : X \to B$ with a measurable inverse. It turns out that this can be done, in particular, whenever X is any Polish space (a complete separable metric space, or any topological space homeomorphic to one) with its Borel σ-algebra. So $[0, 1]$ is universal in that sense; in fact any uncountable Polish space will do here. We chose $[0, 1]$ mainly because it is compact.

Corollary 10.10. *Given any sequence $\{\mu_n\}$ of consistent measures on $\mathbb{R}^n$, there exists a probability space $\{\Omega, \mathcal{F}, P\}$ and a sequence of random variables $X_1, X_2, \dots$ defined on Ω such that $(X_1, \dots, X_n) \sim \mu_n$.*

Proof. Let μ be the measure on $\mathbb{R}^{\mathbb{N}}$ produced by the Kolmogorov extension theorem. Take the probability space $(\Omega, \mathcal{F}, P) = (\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, \mu)$, and for each n let $X_n : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ be the projection map $X_n(x) = x(n)$. It is simple to check that each X_n is measurable (hence a random variable on $(\Omega, \mathcal{F}, P)$) and that the joint distribution of $(X_1, \dots, X_n)$ is μ_n. □

Corollary 10.11. *If $\nu_1, \nu_2, \dots$ are probability measures on $\mathbb{R}$, there exists a probability space with a sequence of* independent *random variables $X_1, X_2, \dots$ such that $X_n \sim \nu_n$.*

Proof. Let $\mu_n = \nu_1 \times \cdots \times \nu_n$. This is a consistent family by definition of product measure. By the previous corollary we can find $X_1, X_2, \ldots$ with $(X_1, \ldots, X_n) \sim \mu_n = \nu_1 \times \cdots \times \nu_n$. This means $X_1, \ldots, X_n$ are independent, for any n. But the definition of independence only takes finitely many objects at a time, so in fact this means the entire sequence $X_1, X_2, \ldots$ is independent. □

In other words, infinite product measure exists. (This actually holds in more generality than Kolmogorov's theorem.) It is then clear that a sequence of random variables is independent iff their joint distribution is an infinite product measure.

Now let's prove Kolmogorov's extension theorem.

Proof. We are going to use Carathéodory's extension theorem to produce the measure μ; it's the only nontrivial tool we have to construct measures.

Let $\mathcal{A}$ be the algebra of all cylinder sets, i.e. those of the form $B \times I \times \ldots$, $B \subset I^n$ for some n. Define μ on $\mathcal{A}$ in the obvious way: $\mu(B \times I \times \ldots) = \mu_n(B)$. We have to check this is well defined: for example, we could also write $B \times I \times I \times \ldots$ as $(B \times I) \times I \times \ldots$ in which case our definition should say its measure should be $\mu_{n+1}(B \times I)$. But by our consistency condition this is the same value. By induction we can see we get the same value no matter how we express the cylinder set.

It is not hard to see that μ is finitely additive on $\mathcal{A}$. Suppose that A_1, A_2 are two disjoint cylinder sets, where $A_1 = B_1 \times I \times \ldots$ for some Borel $B_1 \subset I^{n_1}$ and $A_2 = B_2 \times I \times \ldots$ for some Borel $B_2 \subset I^{n_2}$. Without loss of generality, assume $n_1 \geq n_2$; then we can rewrite A_2 as $B_2' \times I \times \ldots$ where $B_2' = B_2 \times I^{n_2 - n_1} \subset I^{n_1}$. If A_1, A_2 are disjoint then so are B_1 and B_2', and we have $A_1 \cup A_2 = (B_1 \cup B_2') \times I \times \ldots$. So we have

$$\mu(A_1 \cup A_2) = \mu_{n_1}(B_1 \cup B_2') = \mu_{n_1}(B_1) + \mu_{n_1}(B_2').$$

But $\mu_{n_1}(B_1) = \mu(A_1)$ by definition of μ, and by consistency we have $\mu_{n_1}(B_2') = \mu_{n_2}(B_2) = \mu(A_2)$. So we have finite additivity.

For countable additivity, suppose that $A_1, A_2, \cdots \in \mathcal{A}$ are disjoint and that $A := \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ as well; we want to show $\sum_{n=1}^{\infty} \mu(A_n) = \mu(A)$. It suffices to consider the case $A = I^{\mathbb{N}}$. (If this is shown, then for any other A we may take $A_0 = A^c$ and see that $1 = \mu(I^{\mathbb{N}}) = \mu(A^c) + \sum_{n=1}^{\infty} \mu(A_n)$, but by finite additivity $\mu(A^c) = 1 - \mu(A)$). One inequality is easy: finite additivity gives $\sum_{n=1}^{N} \mu(A_n) \leq 1$ so the same holds in the limit.

For the reverse inequality, fix $\epsilon > 0$. For each n we can write $A_n = B_n \times I \times \ldots$ for some Borel $B_n \subset I^n$. As mentioned above, μ_n is a regular measure, so there is an open $U_n \supset B_n$ with

$$\mu_n(U_n) \leq \mu_n(B_n) + 2^{-n}\epsilon. \tag{11}$$

Set $V_n = U_n \times I \times \ldots$; then V_n is an open subset of $I^{\mathbb{N}}$ and $A_n \subset V_n$. Since $\bigcup_n A_n = I^{\mathbb{N}}$ we also have $\bigcup_n V_n = I^{\mathbb{N}}$. So $\{V_n\}$ is an open cover of $I^{\mathbb{N}}$. By compactness there is a finite subcover, say $\{V_1, \ldots, V_N\}$, which is to say that $V_1 \cup \cdots \cup V_N = I^{\mathbb{N}}$. By finite (sub)additivity we must have $\mu(V_1) + \cdots + \mu(V_N) \geq 1$. But then

$$1 \leq \sum_{n=1}^{N} \mu(V_n) = \sum_{n=1}^{N} \mu_n(U_n) \leq \sum_{n=1}^{N} (\mu_n(B_n) + \epsilon 2^{-n}) = \sum_{n=1}^{N} (\mu(A_n) + \epsilon 2^{-n}) \leq \sum_{n=1}^{\infty} \mu(A_n) + \epsilon.$$

Since ϵ was arbitrary we have $\sum_{n=1}^{\infty} \mu(A_n) \geq 1$.

□

Remark 10.12. One can also prove Kolmogorov's extension theorem for *uncountable* products of $\mathbb{R}$ with itself (i.e. the space of real-valued functions on an arbitrary uncountable set). ($\mathbb{R}$ can again be replaced by

any standard Borel space.) However this is more complicated and less useful. It's more complicated because we have to use the full version of Tychonoff's theorem on the compactness of uncountable products, and this requires Zorn's lemma, ultrafilters, etc; it is much more abstract and nonconstructive. Also, uncountable product space is a nasty measurable space; it has in some sense too many measurable sets (the product σ-field is not generated by any countable collection) and not enough (for instance, finite sets are not measurable). And every "practical" theorem in probability I've ever seen can be done with just the countable version.

One might think the uncountable version of Kolmogorov would be useful when trying to construct uncountable families of random variables, such as when constructing continuous-time stochastic processes such as Brownian motion. Indeed, some textbooks actually take this approach (see for instance Karatzas and Shreve). But when you do this it turns out the random variables you constructed don't actually do everything you wanted, and you have to "fix" them (this is the idea of a modification of a process), and your life is complicated by the nastiness of using uncountable product space as a probability space. All this trouble can be avoided by first constructing a countable family of random variables with countable Kolmogorov, and then describing the rest of the family in terms of them.

We'll see more applications of Kolmogorov's extension theorem when we construct stochastic processes such as Markov chains and Brownian motion.

Tuesday, September 26

11 Weak convergence

Our next result will be the central limit theorem. The SLLN tells us approximately how large S_n is; it's about $nE[X_1]$, to first order. The central limit theorem tells us about its distribution, or in some sense its shape: exactly how likely it is to be far from its expected value. Specifically, it says that for large n, the distribution of S_n is "approximately" normal.

The word "approximately" in the previous sentence needs to be interpreted in terms of some notion of convergence. This is so-called weak convergence, which we will now discuss. Since we are interested in distributions (i.e. probability measures on $\mathbb{R}$ or $\mathbb{R}^n$) being close, weak convergence is fundamentally a notion of convergence of measures.

My goal here is to take a somewhat different approach from Durrett, that places a bit less emphasis on working on $\mathbb{R}$, and uses techniques that, where feasible, apply more generally.

Notation 11.1. $C_b(\mathbb{R}^d)$ denotes the set of all bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$. $C_c(\mathbb{R}^d)$ is all the continuous $f : \mathbb{R}^d \to \mathbb{R}$ that have compact support.

Definition 11.2. Let $\mu_1, \mu_2, \ldots, \mu$ be probability measures on $\mathbb{R}^d$. We say $\mu_n \to \mu$ **weakly** if, for every bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$, we have $\int f \, d\mu_n \to \int f \, d\mu$. (Durrett would write $\mu_n \Rightarrow \mu$.) [This definition makes sense if $\mathbb{R}^d$ is replaced by any topological space X equipped with its Borel σ-field.]

Note: Durrett defines weak convergence in terms of the distribution functions $F_n(x) = \mu((-\infty, x])$. We will see later that both definitions are equivalent. I prefer the definition in terms of bounded continuous functions because it seems cleaner and it is not tied to the structure of $\mathbb{R}$, so it is more general.

Example 11.3. For $x \in \mathbb{R}^d$, let δ_x be the Dirac delta measure which puts one unit of mass at x. If $x_1, x_2, \cdots \in \mathbb{R}$, we have $\delta_{x_n} \to \delta_x$ weakly if and only if $x_n \to x$.

Since $\int f \, d\delta_x = f(x)$, we have $\delta_{x_n} \to \delta_x$ iff $f(x_n) \to f(x)$ for every bounded continuous f. If $x_n \to x$, this is immediate from continuity. Conversely, if x_n does not converge to x then there is a neighborhood U

of x such that $x_n \notin U$ infinitely often. Construct a continuous bump function which is 1 at x and 0 outside U. Then the sequence $\{f(x_n)\}$ is 0 infinitely often and cannot converge to $f(x) = 1$.

Remark 11.4. If you were guessing what would be a good notion of convergence for probability measures, perhaps the most obvious guess would be to require that $\mu_n(B) \to \mu(B)$ for every measurable set B. This would not satisfy the above property, since if we take $B = \{x\}$ and choose a sequence x_n with $x_n \neq x$, we would have $\delta_{x_n}(B) = 0$ for all n while $\delta_x(B) = 1$. This notion of convergence fails to respect the topological structure on $\mathbb{R}^d$; it doesn't know that nearby points of $\mathbb{R}^d$ are close. Likewise, requiring that $\int f\,d\mu_n \to \int f\,d\mu$ for all bounded *measurable* f would be problematic for the same reason.

A useful fact:

Lemma 11.5. *If $U \subset \mathbb{R}^d$ is open, there is a sequence f_k of nonnegative bounded continuous functions increasing to 1_U. (I.e. $f_k = 0$ outside U and $f_k \uparrow 1$ inside U.).*

Proof. Notice that U^c is closed, so $d(x, U^c) = \inf\{d(x, y) : y \in U^c\}$ is continuous in x. Then just take $f_k(x) = kd(x, U^c) \wedge 1$. [This works on any metric space.[2]] □

Corollary 11.6. *If μ, ν are two probability measures on $\mathbb{R}^d$ and $\int f\,d\mu = \int f\,d\nu$ for all bounded continuous f, then $\mu = \nu$.*

Proof. Fix an open set U and choose continuous $f_k \uparrow 1_U$. Then by monotone convergence we get

$$\mu(U) = \int 1_U\,d\mu = \lim \int f_k\,d\mu = \lim \int f_k\,d\nu = \nu(U).$$

Since the open sets are a π-system which generates the Borel σ-field, we must have $\mu = \nu$. □

Actually it is sufficient that $\int f\,d\mu = \int f\,d\nu$ for all continuous, compactly supported f, since using such f we can approximate the indicator of any *bounded* open set U, and the bounded open sets are also a π-system generating the Borel σ-field. [This works in any locally compact, separable metric space.]

Corollary 11.7. *Weak limits are unique; if $\mu_n \to \mu$ weakly and $\mu_n \to \nu$ weakly, then $\mu = \nu$.*

Theorem 11.8. *If $\int f\,d\mu_n \to \int f\,d\mu$ for all* compactly supported *continuous f, then $\mu_n \to \mu$ weakly (i.e. the same holds for all bounded continuous f. (The converse is trivial.) (Homework? Presentation?)*

We can also think of weak convergence as a mode of convergence for random variables: we say $X_n \to X$ weakly if their distributions converge weakly, i.e $\mu_n \to \mu$ where $X_n \sim \mu_n$, $X \sim \mu$. An equivalent statement, thanks to the "change-of-variables theorem": $X_n \to X$ weakly iff for every bounded continuous $f : \mathbb{R} \to \mathbb{R}$, we have $Ef(X_n) \to Ef(X)$.

This idea can be a bit misleading, since weak convergence is really inherently a property of measures, not random variables, it can't tell the difference between random variables that have the same distribution. For example, if $X_n \to X$ weakly, and $X \overset{d}{=} Y$, then we also have $X_n \to Y$ weakly, even though we may not have $X = Y$ a.s. (For instance, X and Y could be independent.) So weak limits are not unique as random variables, they are only unique up to distribution. Moreover, we can even talk about weak convergence of a sequence of random variables which are defined on completely different probability spaces! In this case,

[2] I erroneously said in class that it also works for any completely regular space, i.e. whenever Urysohn's lemma holds, and in particular on locally compact Hausdorff spaces. This is not true and the uncountable ordinal space $\omega_1 + 1$ is a counterexample; you cannot approximate 1_{ω_1} by a sequence of continuous functions. I think the right topological condition for this is "perfectly normal".

statements like almost sure convergence have no meaning, because you can't compare functions that are defined on different spaces.

You'll prove a few properties of weak convergence in your homework. An important one: if $X_n \to X$ in probability then $X_n \to X$ weakly. So this is the weakest mode of convergence yet.

Thursday, September 28

Here are some other equivalent characterizations of weak convergence:

Theorem 11.9 (Portmanteau theorem, named after the suitcase)**.** *Let $\mu_1, \mu_2, \ldots, \mu$ be probability measures on $\mathbb{R}^d$. The following are equivalent:*

1. *$\mu_n \to \mu$ weakly;*
2. *For every open $U \subset \mathbb{R}^d$, $\mu(U) \le \liminf \mu_n(U)$;*
3. *For every closed $F \subset \mathbb{R}^d$, $\mu(E) \ge \limsup \mu_n(E)$;*
4. *For every Borel $B \subset \mathbb{R}^d$ satisfying $\mu(\partial B) = 0$, we have $\mu_n(B) \to \mu(B)$.*

To understand why the inequalities are as they are, think again of our example of taking $\mu_n = \delta_{x_n}$, $\mu = \delta_x$ where $x_n \to x$ in $\mathbb{R}^d$, where we are just pushing a point mass around. If $x_n \in U$ for every n, we could have $x \notin U$ if we push the mass to the boundary of U. Then the mass inside U decreases in the limit. But if $x \in U$ we have to have infinitely many $x_n \in U$, so the mass inside U was already 1. That is, in a weak limit, an open set could lose mass (if it gets pushed to its boundary) but it cannot gain mass. Conversely, a closed set F can gain mass (if mass from outside F reaches the boundary of F in the limit). Statement 4 says that the only way to have a drastic change in the mass inside any set B at the limit is to have some mass wind up on the boundary of B.

Proof. (1 implies 2): Suppose $\mu_n \to \mu$ weakly and U is open. Choose bounded continuous $f_k \uparrow 1_U$ as in our remark. Then for each k and each n we have $\mu_n(U) \ge \int f_k \, d\mu_n$ so for each k we have $\liminf_{n\to\infty} \ge \liminf_{n\to\infty} f_k \, d\mu_n = \int f_k \, d\mu$. Now let $k \to \infty$; by monotone convergence we have $\int f_k \, d\mu \to \mu(U)$.

2 iff 3: Just take complements.

2 and 3 imply 4: If $\mu(\partial B) = 0$ then we have $\mu(B^o) = \mu(\bar{B}) = \mu(B)$. On the other hand, using 2 and 3 and the inclusion $B^o \subset B \subset \bar{B}$ we have

$$\begin{aligned} \mu(B^o) &\le \liminf \mu_n(B^o) \\ &\le \liminf \mu_n(B) \\ &\le \limsup \mu_n(B) \\ &\le \limsup \mu_n(\bar{B}) \\ &\le \mu(\bar{B}). \end{aligned}$$

The first and last are equal, so equality must hold throughout. In particular $\limsup \mu_n(B) = \liminf \mu_n(B) = \mu(B)$.

4 implies 1: Recall in HW 1 you showed that for a nonnegative random variable X, we have $EX = \int_0^\infty P(X \ge t) \, dt$. Recasting this in the case where our probability space is $\mathbb{R}$, it says that for any probability measure μ on $\mathbb{R}$ and any nonnegative measurable f, we have

$$\int f \, d\mu = \int_0^\infty \mu(\{f \ge t\}) \, dt. \tag{12}$$

Let f be a nonnegative, bounded, continuous function on $\mathbb{R}$.

Now for each t, $\{f \geq t\}$ is a closed set, and we have $\partial\{f \geq t\} \subset \{f = t\}$ (since $\{f > t\}$ is an open set contained in $\{f \geq t\}$ and hence contained in $\{f \geq t\}^o$). In particular the sets $\partial\{f \geq t\}$ are all disjoint. Therefore, only countably many of them can have positive measure under μ. So for all but countably many t we have $\mu(\partial\{f \geq t\}) = 0$ and hence $\mu_n(\{f \geq t\}) \to \mu(\{f \geq t\})$. In particular this holds for (Lebesgue) almost every t.

Since f is bounded we have $f < C$ for some C. Thus we have $\mu(\{f \geq t\}) = 0$ for $t > C$. So by dominated convergence, using the dominating function $1_{[0,C]}$, we have

$$\int_0^\infty \mu_n(\{f \geq t\})\, dt \to \int_0^\infty \mu(\{f \geq t\})$$

which thanks to (12) gives us exactly what we want. Finally we can get rid of the assumption that f is nonnegative by considering f^+ and f^-. □

[The proof works in any metric space.]

For probability measures on $\mathbb{R}$ there is a nice characterization of weak convergence in terms of distribution functions. Recall the distribution function F of a measure μ is defined by $F(x) = \mu((-\infty, x])$; F is nondecreasing and right continuous.

Theorem 11.10. *Let $\mu_1, \mu_2, \ldots, \mu$ be probability measures on $\mathbb{R}$ with distribution functions μ_n, μ. Then $\mu_n \to \mu$ weakly if and only if, for every $x \in \mathbb{R}$ such that F is continuous at x, we have $F_n(x) \to F(x)$. That is, $F_n \to F$ pointwise, except possibly at points where F is discontinuous. These correspond to point masses in the measure μ.*

Proof. One direction is easy. Suppose $\mu_n \to \mu$ weakly, and let $x \in \mathbb{R}$. If F is continuous at x, this means $\mu(\{x\}) = 0$. Since $(-\infty, x]$ is a Borel set whose boundary is $\{x\}$, by Portmanteau part 4 we have $\mu_n((-\infty, x]) \to \mu((-\infty, x])$ which is to say $F_n(x) \to F(x)$.

Conversely, suppose the distribution functions F_n converge as described. We will verify Portmanteau part 2.

Let C be the set of points at which F is continuous. Note that F can have at most countably many discontinuities (since μ can have at most countably many point masses, as they are disjoint sets of positive measure), so C is co-countable and in particular dense. Let D be a countable dense subset of C (e.g. choose one element of C in each rational interval).

If $a, b \in D$ then we have $\mu_n((a, b]) = F_n(b) - F_n(a) \to F(b) - F(a) = \mu((a, b])$. Likewise, if A is a finite disjoint union of intervals of the form $(a, b]$, $a, b \in C$, we also have $\mu_n(A) \to \mu(A)$. Actually saying "disjoint" there was redundant because any finite union of such intervals can be written as a disjoint union (if two intervals overlap, merge them into a single interval). So let $\mathcal{A}$ be the class of all such finite unions; note that $\mathcal{A}$ is countable.

I claim any open set U can be written as a countable increasing union of sets of $\mathcal{A}$. If $x \in U$ we can find an interval $(a_x, b_x]$, $a_x, b_x \in D$, containing x and contained in U. The union over all x of such intervals must equal U, and since D is countable it is really a countable union. Then we can write the countable union as an increasing union of finite unions.

So we can find $A_n \in \mathcal{A}$ with $A_n \uparrow U$. In particular, by continuity from below, for any ϵ we can find $A \in \mathcal{A}$ with $A \subset U$ and $\mu(A) \geq \mu(U) - \epsilon$. Now for each n, $\mu_n(U) \geq \mu_n(A)$ so taking the liminf we have

$$\liminf_{n\to\infty} \mu_n(U) \geq \liminf_{n\to\infty} \mu_n(A) = \mu(A) \geq \mu(U) - \epsilon.$$

Taking $\epsilon \to 0$ we have what we want.

□

The same statement can actually be proved on $\mathbb{R}^d$ also, using the notion of multidimensional distribution functions. (Essentially, you have to interpret "increasing" with respect to a partial order on $\mathbb{R}^d$, where $(x_1, \ldots, x_d) \leq (y_1, \ldots, y_d)$ iff $x_i \leq y_i$ for each i.) The discontinuities of such a function no longer need to be countable, but they still have dense complement. And there is another minor complication; the intervals in the above argument have to be replaced with boxes, and it's a little harder to write a finite union of boxes as a disjoint union, but you can still do it.

A key property of weakly converging sequences is that they put most of their mass on a compact set.

Definition 11.11. A sequence of measures $\mu_1, \mu_2, \ldots$ is **tight** if for every $\epsilon > 0$ there is a compact K such that for all n, $\mu_n(K) \geq 1 - \epsilon$.

This is similar in spirit to uniform integrability.

Proposition 11.12. *If $\mu_n \to \mu$ then $\{\mu_n\}$ is tight.*

Proof. Fix $\epsilon > 0$. Since $\mathbb{R}^d = \bigcup_{m=1}^{\infty} B(0, m)$ we can find a m so large that $\mu(B(0, m)) > 1 - \epsilon/2$. Let $K_0 = \overline{B(0, m)}$ which is compact. We have

$$\liminf \mu_n(K_0) \geq \liminf \mu_n(B(0, m)) \geq \mu(B(0, m)) > 1 - \epsilon/2$$

by Portmanteau 2. In particular we must have $\mu_n(K_0) \geq 1 - \epsilon$ for all sufficiently large n, say all $n > N$. But for each $n \leq N$, we can use a similar argument to construct a compact K_n with $\mu_n(K_n) \geq 1 - \epsilon$. Set $K = K_1 \cup \cdots \cup K_N \cup K_0$. As a finite union of compact sets, K is compact, and we have $\mu_n(K) \geq 1 - \epsilon$ for every n. □

Remark 11.13. This proof could be modified slightly to work in a locally compact separable metric space: we have to find a precompact open set of large μ measure. Choose a basis of precompact open sets U_n; by countable additivity, for large enough N we have $\mu(U_1 \cup \cdots \cup U_N) \geq 1 - \epsilon$. Then we can let $K_0 = \overline{U_1} \cup \cdots \cup \overline{U_n}$ which is compact. The theorem can also be proved for any complete separable metric space; see Billingsley.

Tuesday, October 2

A key fact is that this necessary condition is in some sense also sufficient. Namely:

Theorem 11.14 (Prohorov)**.** *If $\mu_1, \mu_2, \ldots$ is tight then it has a weakly convergent subsequence.*

We'll concentrate on the one-dimensional case. Start with the compact case, where tightness is automatic.

Theorem 11.15 (Helly)**.** *If $\mu_1, \mu_2, \ldots$ is a sequence of probability measures on $[0, 1]$ then it has a weakly convergent subsequence.*

It can be shown that the topology of weak convergence on the set $\mathcal{P}([0, 1])$ of probability measures on $[0, 1]$ is metrizable, so this in fact shows that $\mathcal{P}([0, 1])$ is compact.

First proof. Consider the distribution functions $F_1, F_2, \ldots$. Enumerate the rationals in $[0,1]$ as $q_1, q_2, \ldots$. If we set $x_n(i) = F_n(q_i)$ then $\{x_n\}$ is a sequence in $[0,1]^{\mathbb{N}}$; by Tychonoff's theorem it has a subsequence subsequence x_{n_k} converging to some $x \in [0,1]^{\mathbb{N}}$. That is to say, $\lim_{k\to\infty} F_{n_k}(q_i) = x(i)$ for each i.

We have to turn x into an F which will be the distribution function of the limit. F needs to be nondecreasing and right continuous, so let

$$F(t) = \inf\{x(i) : q_i > t\}$$

and $F(1) = 1$. This is clearly nondecreasing. To see it is right continuous, suppose $t_m \downarrow t$. For each rational $q > t$ we can find a t_m with $t < t_m < q$, thus $F(t_m) \le F(q)$. Taking the limit in m, $\lim F(t_m) \le F(q)$. Taking the infimum over $q > t$, $\lim F(t_m) \le F(t)$. The reverse inequality is immediate since F is nondecreasing.

Now suppose F is continuous at t. Fix ϵ and choose rational $q < t < r$ with $F(r) - F(q) < \epsilon$. Now for each k, $F_{n_k}(t) \ge F_{n_k}(q) \to F(q)$, so $\liminf_{k\to\infty} F_{n_k}(t) \ge F(q) \ge F(t) - \epsilon$. Similarly, $\limsup_{k\to\infty} F_{n_k}(t) \le F(r) \le F(t) + \epsilon$. Letting $\epsilon \to 0$ we see that we have $\lim_{k\to\infty} F_{n_k}(t) = F(t)$. □

A similar proof can be used on $[0,1]^d$. It is also possible to extend it to $[0,1]^{\mathbb{N}}$ using the Kolmogorov extension theorem and more work.

What goes wrong if we try to do this on $\mathbb{R}$ instead of $[0,1]$? We can still find a nondecreasing, right continuous F such that $F_{n_k} \to F$ where F is continuous. The only problem is that F may not be a "proper" distribution function; it may fail to have $F(+\infty) = 1$ and/or $F(-\infty) = 0$ (interpreted via limits). This happens, for instance, with our example of $\mu_n = \delta_n$; in some sense we push mass off to infinity. The limiting object could be interpreted as a "sub-probability measure" μ; a measure with total mass less than 1, and we would say the μ_n converge **vaguely** to this degenerate sub-probability measure. (We would have $\int f\,d\mu_n \to \int f\,d\mu$ for all $f \in C_c(\mathbb{R})$, but not for all $f \in C_b(\mathbb{R})$; consider $f = 1$ for instance.) But we are really most interested in the case where the limit is actually an honest probability measure, and this is where we would need the condition of tightness.

Second proof, uses functional analysis. $C([0,1])$ is a separable Banach space, and the Riesz representation theorem says its dual space $C([0,1])^*$ is all the signed finite measures on $[0,1]$, with the total variation norm. We note that weak-* convergence in $C([0,1])^*$ is exactly the same as what we were calling weak convergence. The unit ball B of $C([0,1])^*$, in the weak-* topology, is compact (Alaoglu's theorem, really Tychonoff again) and metrizable (it's enough to check convergence on a countable dense subset of $C([0,1])$, so in fact B embeds in $[0,1]^{\mathbb{N}}$). Thus any sequence of probability measures has a weakly convergent subsequence. It is easy to check the limit μ is a (positive) probability measure by noting that, since $\int f\,d\mu_n \ge 0$ for $f \ge 0$ and $\int 1\,d\mu_n = 1$, the same properties must hold for μ. □

This proof works on any compact metric space.

Remark 11.16. We could apply the Riesz/Alaoglu argument on $[0,1]^{\mathbb{N}}$ to prove the Kolmogorov extension theorem. If μ_n are measures on $[0,1]^n$, extend them to $[0,1]^{\mathbb{N}}$ in some silly way. For example let δ_0 be a Dirac mass at $(0,0,\ldots) \in [0,1]^{\mathbb{N}}$, and set $\tilde{\mu}_n = \mu_n \times \delta_0$. Now some subsequence $\tilde{\mu}_{n_k}$ converges weakly to some measure μ. I claim μ is the measure desired in the Kolmogorov extension theorem. Let $V = U \times I \times I \times \ldots$ be an open cylinder set, where $U \subset I^n$ is open. We have $\tilde{\mu}_m(V) = \mu_n(U)$ for all $m \ge n$, so by weak convergence $\mu(V) \le \mu_n(U)$. Now since I^n is a metric space and U^c is closed, we can find open sets $U_i \downarrow U^c$. If we set $V_i = U_i \times I \times \ldots$ we also have $\mu(V_i) \le \mu_n(U_i)$. Passing to the limit, $\mu(V^c) \le \mu_n(U^c)$, i.e. $\mu(V) \ge \mu_n(U)$. By a π-λ argument the same holds for all cylinder sets, and μ has the desired property.

Now we can prove Prohorov's theorem for $\mathbb{R}$.

Proof. We will actually prove it for $(0,1)$ which is homeomorphic to $\mathbb{R}$. (Here the issue is not pushing mass off to infinity, but pushing it off the edge of the interval toward 0 or 1.) If we have probability measures μ_n on $(0,1)$ then they extend in the obvious way to probability measures on $[0,1]$, and there is a subsequence converging weakly (with respect to $[0,1]$) to a measure μ on $[0,1]$, which we can then restrict to a measure on $(0,1)$ again.

There are two issues. First, we have to check that $\mu((0,1)) = 1$ (conceivably μ could put some of its mass on the endpoints). Second, we only have weak convergence with respect to $[0,1]$, which is to say $\int f\,d\mu_n \to \int f\,d\mu$ for all f which are bounded and continuous on $[0,1]$. We need to know it holds for f which are merely continuous on $(0,1)$, which is a larger class (consider for instance $f(x) = \sin(1/x)$).

So we need to use tightness. Fix $\epsilon > 0$; there is a compact $K \subset (0,1)$ with $\mu_n(K) \geq 1-\epsilon$ for all n. K is also compact and hence closed in $[0,1]$, so by Portmanteau 3, we have $\mu((0,1)) \geq \mu(K) \geq \limsup \mu_n(K) \geq 1-\epsilon$. Letting $\epsilon \to 0$ we see $\mu((0,1)) = 1$.

Next, suppose f is bounded and continuous on $(0,1)$; say $|f| \leq C$. K must be contained in some closed interval $[a,b] \subset (0,1)$. Let us modify f outside $[a,b]$ to get a new function $\tilde{f}$ that's continuous up to $[0,1]$; for instance, let $\tilde{f} = f(a)$ on all of $[0,a]$ and $\tilde{f} = f(b)$ on $[b,1]$. (In a more general setting we would use the Tietze extension theorem here.) We now have $\int \tilde{f}\,d\mu_n \to \int \tilde{f}\,d\mu$. But on the other hand, for each n we have

$$\left| \int \tilde{f}\,d\mu_n - \int f\,d\mu_n \right| \leq \int |\tilde{f} - f|\,d\mu_n \leq 2C(1-\mu_n(K)) \leq 2C\epsilon$$

since f and $\tilde{f}$ agree inside K, and are each bounded by C outside K. The same goes for μ. So we must have $\int f\,d\mu_n \to \int f\,d\mu$ and we have shown $\mu_n \to \mu$ weakly. □

Remark 11.17. This proof would work if we replaced $\mathbb{R}$ by any space homeomorphic to a Borel subset of $[0,1]$ (or $[0,1]^d$ or $[0,1]^{\mathbb{N}}$ if we used a fancier proof of Helly, or any other compact metric space if we use Riesz). It turns out that every complete separable metric space is homeomorphic to a Borel subset of $[0,1]^{\mathbb{N}}$, so this proof applies to all Polish spaces. Actually with a bit more work it can be proved for any metric space at all; again see Billingsley.

Example 11.18. For the central limit theorem, we want to examine the weak convergence of $S_n/\sqrt{n}$. (We assume for simplicity that X_1 has mean 0 and variance 1, which can be accomplished by looking at $(X - EX)/\sqrt{Var(X)}$.) Notice that $\mathrm{Var}(S_n/\sqrt{n}) = 1$ for all n. So by Chebyshev, $P(|S_n|/\sqrt{n} \geq a) \leq 1/a^2$. Given any $\epsilon > 0$, choose a so large that $1/a^2 < \epsilon$; then if we set $K = [-a,a]$, we have for all n that $P(S_n/\sqrt{n} \in K) \geq 1-\epsilon$. So if μ_n is the distribution of $S_n/\sqrt{n}$, we have just shown that $\{\mu_n\}$ is tight.

This is a good sign since we do indeed hope it is going to converge weakly. By Prohorov we know that a subsequence is guaranteed to converge. If we can show all convergent subsequences converge to the same limit, then a double subsequence trick will give us that the sequence $\{\mu_n\}$ itself converges, which is exactly what the CLT requires. We then just have to check that the limit is in fact the normal distribution. We'll pursue this further as we go.

Example 11.19. Random walk measures on $C([0,1])$ are tight, limit is Wiener measure.

Thursday, October 4

The following theorem falls into the category of "cheap trick". It can enable one to give very short proofs of various facts about weak convergence, by reducing to almost sure convergence, but at the expense of gaining very little understanding of weak convergence itself. Durrett likes to use it as much as possible; I prefer not to. You can use it in your homework if you feel you must, but I would encourage you to also try to find proofs that don't use it.

Theorem 11.20 (Skorohod representation theorem)**.** *Suppose μ_n, μ are probability measures on $\mathbb{R}^d$ and $\mu_n \to \mu$ weakly. There exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ and random variables $\tilde{X}_n, \tilde{X}$ defined on $\tilde{\Omega}$ such that $\tilde{X}_n \sim \mu_n$, $\tilde{X} \sim \mu$, and $\tilde{X}_n \to \tilde{X}$ $\tilde{P}$-almost surely.*

We could restate this theorem solely in terms of random variables as follows: if $X_n \to X$ weakly, then there exist random variables $\tilde{X}_n, \tilde{X}$ defined on a *different probability space* $\tilde{\Omega}$ such that $\tilde{X}_n \stackrel{d}{=} X_n$, $\tilde{X} \stackrel{d}{=} X$, and $\tilde{X}_n \to \tilde{X}$ $\tilde{P}$-almost surely. It is essential to keep in mind that the theorem *only* guarantees that the individual distributions of X_n and $\tilde{X}_n$ are the same; in general their *joint* distributions will be different. In particular, any independence that may hold among the X_n in general will *not* hold for the $\tilde{X}_n$.

I will only sketch the proof here; see Durrett Theorem 3.2.2 for more details. The proof is based on the following fact, which is Durrett's Theorem 1.2.2.

Lemma 11.21. *Suppose μ is a probability measure on $\mathbb{R}$ with distribution function F. Define the "inverse" of F by*

$$F^{-1}(t) = \sup\{x : F(x) < t\}.$$

(This definition ensures that $F^{-1} : (0,1) \to \mathbb{R}$ is everywhere defined and nondecreasing.) If $U \sim U(0,1)$ is a uniform random variable, then $F^{-1}(U) \sim \mu$.

Now to prove Skorohod, let F_n, F be the distribution functions of μ_n, μ. We know that $F_n(x) \to F(x)$ at all points x where F is continuous. Essentially by turning our head sideways, we can show that we also have $F_n^{-1}(t) \to F^{-1}(t)$ at all points t where F^{-1} is continuous. In particular, $F_n^{-1} \to F^{-1}$ almost everywhere on $(0,1)$. So if U is a single uniform $U(0,1)$ random variable defined on some probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ (for example, we could use $(0,1)$ with Lebesgue measure and take $U(\omega) = \omega$), we can set $\tilde{X}_n = F_n^{-1}(U)$, $\tilde{X} = F^{-1}(U)$. By the previous lemma, the distributions of $\tilde{X}_n, \tilde{X}$ are as desired, and since $F_n^{-1} \to F^{-1}$ almost everywhere, we have $\tilde{X}_n \to \tilde{X}$ almost surely.

12 Characteristic functions and Fourier transforms

A very powerful way to describe probability measures on $\mathbb{R}^d$ is via their Fourier transforms, also called characteristic functions.

Definition 12.1. Let μ be a probability measure on $\mathbb{R}$. The **Fourier transform** or **characteristic function** or **chf** of μ is the function $\phi_\mu : \mathbb{R} \to \mathbb{C}$ (or $\hat{\mu}$) defined by

$$\phi_\mu = \int_{\mathbb{R}} e^{itx} \mu(dx).$$

Durrett usually names this function $\phi(t)$ or $\phi_\mu(t)$.

Likewise, we can define the characteristic function of a random variable X as the Fourier transform of its distribution, i.e. $\phi_X(t) = E[e^{itX}]$.

Some immediate properties:

- $\phi(0) = 1$ (obvious)
- $\phi(-t) = \overline{\phi(t)}$ (obvious)
- $|\phi(t)| \leq 1$ (triangle inequality)

- $|\phi(t+\delta) - \phi(t)| \le \int |e^{i\delta x} - 1|\, dx \to 0$ as $\delta \to 0$, so ϕ is uniformly continuous. (Factor.)
- $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$. (Obvious.)
- If $\mu_n \to \mu$ weakly then $\phi_{\mu_n}(t) \to \phi_\mu(t)$ pointwise. (Actually with more work we can show the convergence is uniform on compact sets.)

Example 12.2. If μ is the normal distribution $N(c, \sigma^2)$, i.e $d\mu = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-c)^2/2\sigma^2}\, dx$, then $\phi_\mu(t) = e^{ict - \sigma^2 t^2/2}$. (Presentation)

Example 12.3. For many other examples of computing characteristic functions, see Durrett.

The more moments a measure has, the smoother its characteristic function.

Proposition 12.4 (Durrett Exercise 3.3.14)**.** *If $\int |x|^n \mu(dx) < \infty$, then ϕ_μ is C^n and its nth derivative is given by $\phi_\mu^{(n)}(t) = \int (ix)^n e^{itx} \mu(dx)$. (Presentation)*

For a random variable, this says:

Corollary 12.5. *If $E|X|^n < \infty$ then ϕ_X is C^n and its nth derivative is $\phi_X^{(n)}(t) = E[(iX)^n e^{itX}]$.*

When you add two independent random variables, their distributions convolve, which is a bit messy. But their chfs multiply:

Proposition 12.6. *If X, Y are independent, then $\phi_{X+Y} = \phi_X \phi_Y$.*

Proof.

$$E[e^{it(X+Y)}] = E[e^{itX}e^{itY}] = E[e^{itX}]E[e^{itY}]$$

since e^{itX}, e^{itY} are independent random variables for each t. □

Thursday, October 11

Durrett's development is completely in terms of probability. This makes it more self-contained but I think it loses the connection with the wider world of Fourier theory. I'll use it more explicitly.

Definition 12.7. If $f : \mathbb{R} \to \mathbb{C}$ is integrable, we define its Fourier transform as

$$\hat{f}(t) = \int_{\mathbb{R}} e^{itx} f(x)\, dx.$$

Using t as the argument of f is perhaps wrong because it should really be a variable in the frequency domain, not the time domain. However everyone in probability seems to use t for the Fourier transform variable.

A nice class of functions to work with when doing Fourier analysis are the **Schwartz functions**:

Definition 12.8. The **Schwartz class** $\mathcal{S}$ consists of all functions $f : \mathbb{R} \to \mathbb{C}$ which are C^∞ and, for every n, k, we have that $|x|^n f^{(k)}(x)$ is bounded. So f and all its derivatives decay at infinity faster than any polynomial; in particular they are integrable. Examples: any C^∞ function with compact support, $e^{-|x|^2}$.

Theorem 12.9. *If $f \in \mathcal{S}$ then $\hat{f} \in \mathcal{S}$.*

Proof. This follows from two basic facts about the Fourier transform:

1. If f is differentiable and f' is integrable, then $\widehat{f'}(t) = -it\hat{f}(t)$. (Integrate by parts.) In particular, since $\widehat{f'}$ is bounded, we must have $\hat{f}$ decaying at least as fast as $1/|t|$.

2. Conversely, if $xf(x)$ is integrable, then $\hat{f}$ is differentiable and $\hat{f}'(t) = i\widehat{xf(x)}$. (Effectively the same as Proposition 12.4 above.)

□

The Fourier transform has an inverse: let $\check{g}(t) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx} g(t)\,dt$. Then:

Theorem 12.10. *For all $f \in \mathcal{S}$, $\check{\hat{f}} = \hat{\check{f}} = f$.*

Proof. See any Fourier analysis book. The proof is not quite trivial; you have two integrals which you would like to interchange, but the hypotheses of Fubini's theorem are not satisfied. □

Combined with:

Lemma 12.11. *If $\int f\,d\mu = \int f\,d\nu$ for all $f \in C_c^\infty(\mathbb{R})$ then $\mu = \nu$.*

Proof. We previously argued that for any bounded open set U, we can approximate 1_U from below by continuous functions f_n with compact support. The f_n can actually be made smooth, for example, with a convolution. Then monotone convergence gives $\mu(U) = \nu(U)$ for all bounded open U, but the bounded open sets are a π-system which generates the Borel σ-field. □

We can conclude that a measure is determined by its Fourier transform.

Theorem 12.12. *If $\phi_\mu = \phi_\nu$ then $\mu = \nu$.*

Proof. For $f \in \mathcal{S}$,

$$\begin{aligned}\int f(x)\,\mu(dx) &= \frac{1}{2\pi}\iint e^{itx}\check{f}(t)\,dt\,\mu(dx)\\ &= \frac{1}{2\pi}\int \check{f}(t)\int e^{itx}\,\mu(dx)\,dt\\ &= \frac{1}{2\pi}\int \check{f}(t)\phi_\mu(t)\,dt.\end{aligned}$$

The same holds for ν. So we can conclude that $\int f\,d\mu = \int f\,d\nu$ for all $f \in \mathcal{S}$ which by the previous lemma shows $\mu = \nu$. □

It is actually possible to invert the Fourier transform of a measure more explicitly. If the measure μ has a density f with respect to Lebesgue measure, its chf ϕ will be integrable and you will just have $f(x) = \check{\phi}(x) = \frac{1}{2\pi}\int e^{-itx}\phi(t)\,dt$. If μ is not absolutely continuous to Lebesgue measure then ϕ may not be integrable. For instance, take $\mu = \delta_0$, so that $\phi = 1$. We cannot compute $\check{\phi}$ in the obvious way because the Lebesgue integral $\frac{1}{\sqrt{2\pi}}\int e^{-itx}1\,dt$ does not exist. However, this can be worked around with a more complicated approach (essentially using an improper integral). See Durrett's Theorem 3.3.4, which shows how to recover the μ measure of intervals. I don't want to pursue this; knowing that the Fourier transform is one-to-one is all we need for present purposes.

We can now get a powerful result relating weak convergence to convergence of Fourier transforms.

Theorem 12.13. *Let μ_n be a sequence of probability measures on $\mathbb{R}$ with characteristic functions ϕ_n. Suppose that $\phi_n(t)$ converges pointwise to some function $\phi(t)$, and further suppose that $\{\mu_n\}$ is tight. Then μ_n converges weakly to some probability measure μ, whose characteristic function is ϕ.*

Example 12.14. To see we need something besides just pointwise convergence of chfs, let μ_n be a normal distribution with mean 0 and variance n, so that $\phi_n(t) = e^{-nt^2/2}$. Then $\phi_n(t)$ converges pointwise, but the limit is $1_{\{0\}}$ which cannot be the chf of any measure because it is not continuous. This shows that the sequence $\{\mu_n\}$ does not converge weakly, and indeed is not tight.

Proof. By tightness, there is a subsequence μ_{n_k} converging weakly to some probability measure μ. Thus $\phi_{n_k} \to \phi_\mu$ pointwise so we must have $\phi_\mu = \phi$. If there is another convergent subsequence $\mu_{n'_k}$ converging to some other measure μ', the same argument shows that $\phi_{\mu'} = \phi$, whence from the previous theorem we have $\mu' = \mu$. So in fact every convergent subsequence of $\{\mu_n\}$ converges to μ.

It will follow, using a double subsequence trick, that the entire sequence converges to μ. Suppose it does not. There is a bounded continuous f such that $\int f\,d\mu_n \not\to \int f\,d\mu$. We can then find an $\epsilon > 0$ and a subsequence μ_{m_k} so that $\left|\int f\,d\mu_{m_k} - \int f\,d\mu\right| > \epsilon$ for all k. But μ_{m_k} is itself tight so it has a further subsequence $\mu_{m_{k_j}}$ converging weakly, and as argued above the limit must be μ. Then $\int f\,d\mu_{m_{k_j}} \to \int f\,d\mu$ which is a contradiction. □

Combined with a little calculus, we can prove the central limit theorem.

Theorem 12.15 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be iid L^2 random variables. Set $S_n = X_1 + \cdots + S_n$. Then*

$$\frac{S_n - nE[X_1]}{\sqrt{n\,\mathrm{Var}(X_1)}} \to N(0,1) \quad \textit{weakly}$$

where $N(0,1)$ is the standard normal distribution, i.e. the measure μ defined by $d\mu = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$.

Proof. By replacing X_n by $\frac{X_n - E[X_n]}{\sqrt{\mathrm{Var}(X_n)}}$, we can assume without loss of generality that $E[X_n] = 0$ and $\mathrm{Var}(X_n) = 1$, so we just have to show $\frac{S_n}{\sqrt{n}} \to N(0,1)$ weakly. We have already shown the sequence is tight, so by the previous theorem it suffices to show the chfs converge to the chf of the normal distribution, $e^{-t^2/2}$.

If ϕ is the chf of X_n, then the chf of $S_n/\sqrt{n}$ is given by $\phi_n(t) = \phi(t/\sqrt{n})^n$. Since $E|X_n|^2 < \infty$, ϕ is C^2 by Proposition 12.4; we have $\phi(0) = 0$, $\phi'(0) = E[iX] = 0$, and $\phi''(0) = E[-X^2] = -1$. Then Taylor's theorem tell us that

$$\phi(t) = 1 - \frac{1}{2}t^2 + o(t^2)$$

or in other words, $\phi(t) = 1 - \frac{1}{2}t^2 + \epsilon(t)$ where $\epsilon(t)/t^2 \to 0$ as $t \to 0$. Thus

$$\phi_n(t) = \left(1 - \frac{t^2}{2n} + \epsilon(\frac{t}{\sqrt{n}})\right)^n.$$

We just have to compute the limit as $n \to \infty$; we are only asking for pointwise convergence so we can treat t as fixed. The rest of the proof is nothing but calculus and could be assigned to a determined Math 1120 student. Notice if the ϵ term were not there, we would just have the classic limit $\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$ with $x = -t^2/2$. So we just have to show the ϵ term can be neglected.

Set $a_n = -\frac{t^2}{2n} + \epsilon(\frac{t}{\sqrt{n}})$. I claim $na_n \to -\frac{t^2}{2}$ as $n \to \infty$, because if we let $b_n = \frac{t}{\sqrt{n}}$ and note that $b_n \to 0$, we have $n\epsilon(b_n) = t^2\epsilon(b_n)/b_n^2 \to 0$. So now we just need to show:

Claim 12.16. *If $na_n \to a \in (-\infty, \infty)$ then $(1 + a_n)^n \to e^a$.*

We take a log[3] and consider $n \ln(1 + a_n)$. By Taylor's theorem we have $\ln(1 + x) = x + o(x)$. So we are looking at

$$na_n + no(a_n) = na_n + (na_n)\frac{o(a_n)}{a_n} \to a + a \cdot 0 = a.$$

We have thus proved the claim and also the theorem. □

Remark 12.17. Durrett goes to more trouble in estimating the remainder term (which I have called $\epsilon(t)$); see his Lemma 3.3.7. He makes it sound like this is necessary to get the CLT for L^2 random variables, but unless I am really missing something, this is not so. The plain ordinary Taylor theorem says that $\phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2}t^2\phi''(0) + o(t^2)$ (the Peano remainder) provided only that $\phi''(0)$ exists, which we know to be the case by Proposition 12.4; indeed, we know $\phi \in C^2(\mathbb{R})$.

Theorem 12.13 requires convergence of chfs as well as tightness. For the CLT we were able to show the tightness of $\frac{S_n}{\sqrt{n}}$ directly using Chebyshev (see Example 11.18). A nifty theorem due to Lévy says you can get tightness by looking at what the chfs converge to.

Theorem 12.18 (Lévy continuity theorem)**.** *Let μ_n be a sequence of probability measures with chfs ϕ_n. Suppose that $\phi_n(t)$ converges pointwise to some function $\phi(t)$. If ϕ is continuous at $t = 0$, then $\{\mu_n\}$ is tight. It then follows from Theorem 12.13 that μ_n converges weakly to the measure μ whose chf is ϕ.*

Note the requirement that the limit is continuous at 0 is enough to exclude the situation of Example 12.14.

Proof. Let's start with what we know: the continuity of ϕ at 0. Since $\phi(0) = \lim \phi_n(0) = 1$, fix an $\epsilon > 0$ and choose $\delta > 0$ such that $|\phi(t) - 1| < \epsilon$ for all $|t| < \delta$. If we had the ϕ_n converging to ϕ uniformly, we could say something similar was true for ϕ_n; but we don't have uniform convergence, only pointwise.

So let's average instead. If we average ϕ over $(-\delta, \delta)$, we will get something close to 1:

$$\left|\frac{1}{2\delta}\int_{-\delta}^{\delta}\phi(t)\,dt - 1\right| \le \frac{1}{2\delta}\int_{-\delta}^{\delta}|\phi(t) - 1|\,dt < \epsilon.$$

But by dominated convergence, $\int_{-\delta}^{\delta}\phi_n(t)\,dt \to \int_{-\delta}^{\delta}\phi(t)\,dt$. Thus for sufficiently large n, say $n \ge N$, we have

$$\left|\frac{1}{2\delta}\int_{-\delta}^{\delta}\phi_n(t)\,dt - 1\right| < 2\epsilon.$$

Actually, since $\phi_n(-t) = \overline{\phi_n(t)}$, the imaginary part of ϕ_n is an odd function, so the integral in the previous equation is actually real. So we can say

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}\phi_n(t)\,dt > 1 - 2\epsilon.$$

[3]Since we will apply this with complex values of a_n, we should choose a branch of the log function whose branch cut stays away from the positive real axis. Since $a_n \to 0$, for sufficiently large n we will avoid the branch cut.

Now, let's look at this integral:

$$\begin{aligned}\frac{1}{2\delta}\int_{-\delta}^{\delta}\phi_n(t)\,dt &= \frac{1}{2\delta}\int_{-\delta}^{\delta}\int e^{itx}\,\mu_n(dx)\,dt \\ &= \int \frac{1}{2\delta}\int_{-\delta}^{\delta} e^{itx}\,dt\,\mu_n(dx) \qquad \text{(Fubini)} \\ &= \int \frac{\sin(\delta x)}{\delta x}\,\mu_n(dx)\end{aligned}$$

where we just evaluated the dt integral.[4]

Why is this helpful? If you look at a graph of $\frac{\sin u}{u}$, you will see that it is less than 1 everywhere, and less than 1/2 outside $[-2,2]$. So we have $\frac{\sin u}{u} \le \frac{1}{2} + \frac{1}{2}1_{[-2,2]}$. Therefore we have

$$\int \frac{\sin(\delta x)}{\delta x}\,\mu_n(dx) \le \frac{1}{2} + \frac{1}{2}\int 1_{[-2,2]}(\delta x)\,\mu_n(dx) = \frac{1}{2} + \mu_n([-\frac{2}{\delta},\frac{2}{\delta}]).$$

Thus we have shown $\frac{1}{2} + \frac{1}{2}\mu_n([-\frac{2}{\delta},\frac{2}{\delta}]) > 1 - 2\epsilon$, or in other words, $\mu_n([-\frac{2}{\delta},\frac{2}{\delta}]) > 1 - 4\epsilon$. This holds for all $n \ge N$, with the same δ, so we have effectively shown tightness. (We found a compact K_0, namely $K_0 = [-\frac{2}{\delta},\frac{2}{\delta}]$, such that $\mu_n(K_0) > 1 - 2\epsilon$ for all $n \ge N$. As in the proof of Proposition 11.12 we can find a larger K such that $\mu_n(K) > 1 - 4\epsilon$ for all $n \ge 1$.) □

This gives us a quick proof of a problem from the homework:

Proposition 12.19. *If $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly then $\mu_n * \nu_n \to \mu * \nu$ weakly.*

Proof. Let ϕ_n, ψ_n be the chfs of μ_n, ν_n respectively, and ϕ, ψ the chfs of μ, ν; then $\phi_n \to \phi$ and $\psi_n \to \psi$ pointwise. The chf of $\mu_n * \nu_n$ is $\phi_n\psi_n$, which converges pointwise to $\phi\psi$. This is the chf of $\mu * \nu$ and in particular is continuous at 0, so Lévy's theorem lets us conclude that $\mu_n * \nu_n \to \mu * \nu$ weakly. □

Example 12.20. The CLT doesn't apply for sums of iid random variables that are not L^2, but there are other results in this area. See Durrett's section 3.7 on "stable laws" for what happens; we can still estimate the distribution of S_n. In these cases the normalization has to be something other than $\frac{1}{\sqrt{n}}$. The limiting distributions are called "stable laws"; the best way to describe them is in terms of their chfs (their densities and distribution functions don't have closed form expressions). (Presentation)

Remark 12.21. There is also a version of the CLT for multidimensional random vectors. It says that if $\mathbf{X}_n$ is an iid sequence of d-dimensional random vectors with mean zero, then $\frac{1}{\sqrt{n}}(\mathbf{X}_1 + \cdots + \mathbf{X}_n)$ converges weakly to a normal distribution on $\mathbb{R}^d$. The variance σ^2 has to be replaced by a covariance matrix Σ and can't be renormalized away quite so simply. The issue is that $\mathbf{X}_n$ could be supported in a proper subspace of $\mathbb{R}^d$ (i.e. its components are linearly dependent) and in that case the limiting distribution must be supported in that same subspace; it cannot be an absolutely continuous distribution. See Durrett's Section 3.9. A lot of his development of multidimensional distribution functions can be skipped because we have already done it in the language of measures.

[4] To handle the case $x = 0$ we should adopt the convention that $\frac{\sin 0}{0} = 0$, filling in the removable discontinuity.

13 Basic concepts of discrete-time stochastic processes

A major focus of probability theory is the study of **stochastic processes**. The idea is to study random phenomena that evolve over time. Some examples might include:

- The weather
- Stock prices
- Profits in a gambling scheme (perhaps related to the previous example)
- Noise in an electrical circuit
- Small particles moving around in a fluid

In this section we are going to study **discrete-time stochastic processes**, where we model time by the integers, so that it passes in discrete steps. For some applications this is reasonable: e.g. modeling the change in weather from day to day, looking at closing prices for a stock, playing a game that proceeds one play at a time. For other applications (electrical noise, moving particles, etc) it is not so reasonable, and it is better to work in **continuous time**, where we model time by the reals. This can give more realistic models, but it also adds quite a bit of mathematical complexity. Essentially, the issue is that the integers are countable while the reals are not. Probability deals much better with things that are countable, and so continuous-time models tend to build in enough continuity to guarantee that one can do everything on countable dense subsets (such as the rationals).

Definition 13.1. Let $(S, \mathcal{S})$ be any measurable space. An S-valued discrete-time **stochastic process** is simply a sequence $X_0, X_1, \ldots$ of S-valued random variables (i.e. measurable maps from a probability space Ω to S).

We think of a system evolving over time; X_n is what you see when you observe the system at time n. In general the "state space" of possible observations could be any measurable space S. For most of our examples it will be something like $\mathbb{R}^d$.

A sequence of iid random variables $\{X_n\}$ is technically a stochastic process, but this is not really what you should think of, because it doesn't "evolve": the behavior of $X_1, \ldots, X_n$ tells you absolutely nothing about X_{n+1}. In contrast, most interesting real-world events have some dependence. For instance, knowing the weather today tells us a lot about the weather tomorrow. If it is 80 and sunny today, that makes it much less likely that it will be 20 and snowing tomorrow (though in Ithaca this is perhaps not so clear).

So the canonical example is:

Example 13.2. Simple random walk. Let $\{\xi_i\}$ be an iid sequence of coin flips: $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$. Our stochastic process is $X_n = \xi_1 + \cdots + \xi_n$. Think of X_n as the position of a drunkard on the integers: at each time step he moves one unit right or left, chosen independently of all other choices. Note that the X_n themselves are not independent, since knowing X_n leaves only two possible values for X_{n+1}.

We could also take ξ_i to have any other distribution in $\mathbb{R}$, which would lead to a more general random walk.

Example 13.3. If you are gambling, and with each (iid) play of the game your net winnings are given by ξ_i, then X_n is your total net profit at time n. For instance, if you are playing roulette and betting \$1 on black each time, then $P(\xi_i = 1) = \frac{18}{38}$, $P(\xi_i = -1) = \frac{20}{38}$. If you are buying lottery tickets, then the distribution of ξ_i is something like $P(\xi_i = 70000000) = \frac{1}{175711536}$, $P(\xi_i = -1) = \frac{175711535}{175711536}$.

Example 13.4 (Random walk on a group)**.** The notion of "sum (or product) of random variables" makes sense in any group. If we take as our state space S a group G (equipped with a σ-algebra such that the group operation is measurable), then we can take a sequence of G-valued random variables $\xi_1, \xi_2, \dots$ which are iid and multiply them to get $X_n = \xi_1 \dots \xi_n$ (so the walk starts at the identity: $X_0 = e$). Then $\{X_n\}$ is a nice example of a G-valued stochastic process, called a random walk on G. (For simple random walk, $G = \mathbb{Z}$.)

As a concrete example of this, consider shuffling a deck of 52 cards according to the following procedure: choose one of the 52 cards uniformly at random and swap it with the top card. This corresponds to a random walk on the group $G = S_{52}$, the symmetric group on 52 elements. (Specifically, S_{52} is the set of all bijections of the set $\{1, \dots, 52\}$, with composition as the group operation.) If the ξ_i are iid with the distribution $P(\xi_i = (1\,k)) = 1/52$ for $k = 1, \dots, 52$, then the permutation of the deck after n swaps is $X_n = \xi_n \dots \xi_1$. (Note that here we are multiplying on the left instead of the right by the usual convention for function composition; since this is a non-abelian group the order matters, but the theory is the same either way.)

An unsurprising result is that X_n converges weakly to the uniform measure on S_{52}, i.e. the measure which assigns the same measure $1/52!$ to each of the 52! elements of S_{52}. That is, this procedure really does shuffle the deck. But the rate of convergence is important, because it tells you how long it would take before the deck is "mostly" random. This topic is often called "mixing times".

Example 13.5 (Random walk on a graph)**.** Let $G = (V, E)$ be a locally finite graph. We can imagine a random walk on G as follows: start at some vertex X_0, and at each step n, let X_{n+1} be a uniformly chosen neighbor of X_n. There is a tricky detail here though: intuitively we would like the choices of neighbors to be "independent". But they can't actually be independent, because the values of $X_1, \dots, X_n$ affect your location X_n at time n, which affects the possible values for X_{n+1} because it must be a neighbor of X_n. What we really want, it turns out, is that X_{n+1} can depend on the vertex X_n but not on how you got to that vertex. This is the fundamental idea of a Markov chain, which we'll discuss later.

13.1 Filtrations

As the system evolves, we learn more about it; new information is revealed. Since we have thought about encoding "information" in σ-fields, this leads to the notion of a filtration.

Definition 13.6. A **filtration** on a probability space $(\Omega, \mathcal{F}, P)$ is a sequence $\{\mathcal{F}_n\}$ of sub-σ-fields of $\mathcal{F}$ which are increasing: $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$. A probability space equipped with a filtration is sometimes called a **filtered probability space**, i.e. a 4-tuple $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$.

$\mathcal{F}_n$ is interpreted as the information available at time n; an event A is in $\mathcal{F}_n$ if, by time n, we can determine whether or not it happens. The $\mathcal{F}_n$ are increasing, which means that information learned at time n is remembered from then on.

We sometimes define a "last" σ-field in a filtration: $\mathcal{F}_\infty := \sigma(\mathcal{F}_n : n \geq 0)$, which contains all information that will ever be revealed. $\mathcal{F}_\infty$ is not necessarily equal to $\mathcal{F}$ (the probability space could contain additional randomness that we never get to see), but there is often no loss of generality in replacing $(\Omega, \mathcal{F}, P)$ by the probability space $(\Omega, \mathcal{F}_\infty, P)$.

Example 13.7. For random walk, you could think of the filtration $\mathcal{F}_n = \sigma(\xi_1, \xi_2, \dots)$ generated by the iid sequence ξ_i; the information available at time n is everything we have learned from observing the coin flips so far. (Take $\mathcal{F}_0$ to be the trivial σ-field $\{\Omega, \emptyset\}$ since at time 0 nothing has happened yet.) Then, for instance:

- The event that the first coin is heads $\{\xi_1 = 1\}$ is in $\mathcal{F}_1$.

- The event that at least four of the first seven flips are heads is in $\mathcal{F}_7$ (but not $\mathcal{F}_6$).
- The event that there is ever at least one heads is not in any of the $\mathcal{F}_n$ (but it is in $\mathcal{F}_\infty$).

We would like the process $\{X_n\}$ that we are studying to be part of the available information.

Definition 13.8. A stochastic process $\{X_n\}$ is **adapted** to a filtration $\{\mathcal{F}_n\}$ if $X_n \in \mathcal{F}_n$ (i.e. $\{X_n \in B\} \in \mathcal{F}_n$ for all measurable $B \subset S$). Usually the filtration $\{\mathcal{F}_n\}$ is understood (we are working on a *filtered* probability space) and we will just say $\{X_n\}$ is **adapted**.

Example 13.9. In our running (walking?) random walk example, the random walk $X_n = \xi_1 + \cdots + \xi_n$ is adapted.

13.2 Stopping times

A reasonable question about a randomly evolving system is "when does some phenomenon[5] occur"? Since the answer may be random, it should be expressed by a random variable. A **random time** τ is just an $\mathbb{N} \cup \{\infty\}$-valued random variable whose value is interpreted as a time. (The event $\{\tau = \infty\}$ is to be interpreted as "the phenomenon never occurs".)

Example 13.10. Suppose X_n is the state of the weather on day n, taking values in the set $S = \{C, R, S\}$ (clear, rain, snow). Let us think of the filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Here are some examples of random times:

1. $\tau_1 = \min\{n : X_n = S\}$ is the first day on which it snows. (To convince yourself that τ_1 is measurable, notice that for each n, $\{\tau_1 = n\} = \{X_1 \neq S\} \cap \cdots \cap \{X_{n-1} \neq S\} \cap \{X_n = S\}$, and for any Borel set B, $\{\tau_1 = B\} = \bigcup_{n \in B \cap \mathbb{N}} \{\tau_1 = n\}$ is a countable union of such sets.) Day τ_1 is a good time to skip class and go sledding.
2. $\tau_2 = \tau_1 - 1$ is the day before the first snow. Day τ_2 would be an excellent time to install your snow tires and buy some rock salt.
3. $\tau_3 = \min\{n : X_{n-1} = S, X_n = R\}$ is the first time we see rain coming right after snow. Day τ_3 is a good time to buy sandbags because it may flood. (Day $\tau_3 - 1$ would be even better.)

In this example, τ_2 is a bit problematic: since weather forecasting is imperfect, you will not know the date of τ_2 until it has already passed (i.e. on day τ_1 when the snow actually falls, you will know τ_2 was the previous day). So it isn't actually possible to accomplish the plan "install snow tires on day τ_2". To avoid issues like this, we introduce the idea of a stopping time.

Definition 13.11. A random time τ is a **stopping time** if for each n, we have $\{\tau = n\} \in \mathcal{F}_n$.

That is, on day n, using the available information $\mathcal{F}_n$, you can determine whether or not τ happens today.

Proposition 13.12. *τ is a stopping time iff for each n we have $\{\tau \leq n\} \in \mathcal{F}_n$.*

This says, for a stopping time τ, you can tell on day n whether τ has already happened, and this gives an equivalent definition.

[5] I keep wanting to use the word "event" here but that's already in use.

Proof. If τ is a stopping time, for each $k \le n$ we have $\{\tau = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$. But $\{\tau \le n\} = \bigcup_{k=0}^{n}\{\tau = k\}$ so $\{\tau \le n\} \in \mathcal{F}_n$ also. Conversely, assume for every n we have $\{\tau \le n\} \in \mathcal{F}_n$. Then $\{\tau = n\} = \{\tau \le n\} \setminus \{\tau \le n-1\}$. But $\{\tau \le n\} \in \mathcal{F}_n$, and $\{\tau \le n-1\} \in \mathcal{F}_{n-1} \subset \mathcal{F}_n$, so their difference is also in $\mathcal{F}_n$. □

Example 13.13. In Example 13.10, τ_1 is a stopping time, because $\{\tau_1 \le n\} = \{X_0 = S\} \cup \cdots \cup \{X_n = S\}$. This is a finite union of events from $\mathcal{F}_n$. τ_2 is not a stopping time: for example, $\{\tau_2 \le 2\} = \{X_0 \ne S, X_1 \ne S, X_2 = S\}$ but this event is not in $\mathcal{F}_1$ (unless the weather is way more predictable than we think). We cannot know on day 1 whether or not it is going to snow on day 2. τ_3 is a stopping time because

$$\{\tau_3 \le n\} = \bigcup_{k=1}^{n}\{X_k = R, X_{k-1} = S\}$$

which is again a union of events from $\mathcal{F}_n$.

Probably the most important example is stopping times like τ_1.

Proposition 13.14. *If $\{X_n\}$ is an adapted process and $B \subset S$ is measurable, then $\tau_B := \inf\{n : X_n \in B\}$ is a stopping time. We call τ the **hitting time** of the set B, since it's the first time that X_n hits B. (By convention the infimum of the empty set is $+\infty$, so in the event that X_n never hits B we have $\tau_B = \infty$.)*

Proof. As in our example, $\{\tau_B \le n\} = \bigcup_{k=0}^{n}\{X_k \in B\}$ which is a finite union of events from $\mathcal{F}_n$. □

Also note that constants are also stopping times: if $\tau = m$ then $\{\tau = n\}$ is Ω if $n = m$ and $\emptyset$ otherwise; either way it is in $\mathcal{F}_n$. (This corresponds to waiting a deterministic amount of time, which doesn't use any of the revealed information from the filtration.)

Proposition 13.15. *If S, T are stopping times, then so are $S \wedge T$, $S \vee T$, $S + T$. (Sergio's presentation)*

$S \wedge T$ is "wait until S or T, whichever comes first"; $S \vee T$ is "whichever comes second".

The next object to define is a σ-field which corresponds to the information available when the stopping time τ occurs. We denote it by $\mathcal{F}_\tau$ which looks a bit strange: τ is a random variable, a function on Ω, but $\mathcal{F}_\tau$ is not a random object. Rather it is defined as follows:

Definition 13.16. If $\{\mathcal{F}_n\}$ is a filtration and τ is a stopping time, then $\mathcal{F}_\tau$ is the collection of all events $A \in \mathcal{F}$ such that, for every n, $A \cap \{\tau = n\} \in \mathcal{F}_n$.

So A is in $\mathcal{F}_\tau$ if, whenever we are in the event that $\{\tau = n\}$, we can tell at time n whether A happened.

Proposition 13.17. *$\mathcal{F}_\tau$ is a σ-field.*

Proof. Let n be arbitrary. First, $\emptyset \cap \{\tau = n\} = \emptyset \in \mathcal{F}_n$, so $\emptyset \in \mathcal{F}_\tau$. If $A \in \mathcal{F}_\tau$, then

$$A^c \cap \{\tau = n\} = (A \cup \{\tau = n\}^c)^c = ((A \cap \{\tau = n\}) \cup \{\tau = n\}^c)^c.$$

We have $A \cap \{\tau = n\} \in \mathcal{F}_n$ by assumption, and $\{\tau = n\} \in \mathcal{F}_n$ because τ is a stopping time. Thus $A \in \mathcal{F}_\tau$. Finally, if $A_1, A_2, \ldots in \mathcal{F}_\tau$, then

$$\bigcup_k A_k \cap \{\tau = n\} = \bigcup_k (A_k \cap \{\tau = n\}) \in \mathcal{F}_n$$

so $\bigcup A_k \in \mathcal{F}_\tau$. □

A useful thing to do is to observe an adapted process $\{X_n\}$ at a stopping time τ. The result is the random variable X_τ. (Note that X, τ are both random variables so you should think of this at $X_{\tau(\omega)}(\omega)$. If you have measurability worries, you can allay them by writing $X_\tau = \sum_{0 \le n \le \infty} X_n 1_{\{\tau = n\}}$.)

One caution: for this to make sense, we have to know that $\tau < \infty$, at least almost surely, since our stochastic process does not necessarily include an X_∞. One workaround is to add an extra "cemetery" state Δ to the state space S, and define $X_\tau(\omega) = \Delta$ whenever $\tau(\omega) = \infty$.

Proposition 13.18. *If $\{X_n\}$ is adapted and $\tau < \infty$ is a stopping time, then $X_\tau \in \mathcal{F}_\tau$.*

Proof. If $B \subset S$ is measurable, we have

$$\{X_\tau \in B\} \cap \{\tau = n\} = \{X_n \in B\} \in \mathcal{F}_n.$$

□

For example, if $\tau = \tau_D$ is the first time that X_n hits a measurable set D, then X_{τ_D} is the point of D that actually gets hit. (And on the event that D is never hit, our above convention says to take $X_{\tau_D} = \Delta$.

Proposition 13.19. *If σ, τ are stopping times with $\sigma \le \tau$, then $\mathcal{F}_\sigma \le \mathcal{F}_\tau$.*

That is, if τ always happens later than σ, then at time τ you have more information than at time σ.

Proof. Suppose $A \in \mathcal{F}_\sigma$. We have

$$A \cap \{\tau = n\} = \bigcup_{k=0}^{\infty} A \cap \{\sigma = k\} \cap \{\tau = n\}.$$

If $\sigma \le \tau$ then $\{\sigma = k\} \cap \{\tau = n\} = \emptyset$ for $k > n$, so

$$A \cap \{\tau = n\} = \bigcup_{k=0}^{n} (A \cap \{\sigma = k\}) \cap \{\tau = n\}.$$

But $A \cap \{\sigma = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$ when $k \le n$, and $\{\tau = n\} \in \mathcal{F}_n$, so $A \cap \{\tau = n\} \in \mathcal{F}_n$. Thus $A \in \mathcal{F}_\tau$. □

14 Conditional expectation

A big part of the study of stochastic processes is looking at how information accumulates over time, and what we can do with it. The essential tool here is **conditional expectation**, which we now develop.

In elementary probability, one studies conditional probability: if we are looking at the probability of an event B, but we have the additional information that another event A happened, this may let us improve our estimate of $P(B)$ to the conditional probability $P(B \mid A) := P(B \cap A)/P(A)$. In some sense, we are restricting the probability measure P to the event A, and dividing by $P(A)$ to normalize; we only consider those outcomes that are in A, and ignore the others. To think in terms of information, knowing whether or not A happened is one bit of information, and we can adjust our estimate of $P(B)$ based on this bit: $P(B \mid A)$ if A happened, and $P(B \mid A^c)$ if it did not.

Likewise, if we are looking at an (integrable) random variable X and we have no information whatever about the outcome of our experiment, the best guess we can make as to the value of X is its expectation $E[X]$. But if we know whether or not A happened, then we can improve our guess to one of the values

$E[X \mid A] = E[X; A]/P(A)$ or $E[X \mid A^c]$. (Recall the notation: $E[X; A] = E[X1_A] = \int_A X\,dP$ is the expectation of X over the event A.)

Now what if we have more than one bit of information? Say we have an entire σ-field worth of information, call it $\mathcal{G}$. How can we describe the way this information lets us improve our estimate of a random variable X? We could record, for each $A \in \mathcal{G}$, the value of $E[X \mid A]$. One could imagine this as a map $\mathcal{G} \to \mathbb{R}$), but in fact, we can encode all these improved estimates in a single random variable.

Recall the notation: $E[X; A] = E[X1_A]$ is the expectation of X over the event A; this is the same as $\int_A X\,dP$.

Theorem 14.1. *Let X be an integrable random variable, and $\mathcal{G} \subset \mathcal{F}$ a σ-field. There exists a random variable Y such that:*

1. *$Y \in \mathcal{G}$, and*
2. *For every event $A \in \mathcal{G}$, we have $E[Y; A] = E[X; A]$.*

Moreover, Y is unique up to measure-zero events: if Y' also satisfies items 1,2 above, then $Y = Y'$ a.s.

Thus Y is something that only depends on the information in $\mathcal{G}$, but it encodes all the conditional expectations $E[X; A]$ for $A \in \mathcal{G}$; just compute $E[Y; A]$.

We can also view it as the best approximation we can make to X, given the information in $\mathcal{G}$. This parallels the idea of (unconditional) expectation being the best approximation we can make to X given *no* information about the outcome of the experiment.

Definition 14.2. The conditional expectation of X given $\mathcal{G}$, denoted $E[X \mid \mathcal{G}]$, is the unique random variable described in Theorem 14.1. If B is an event, we define $P(B \mid \mathcal{G})$ as the random variable $E[1_B \mid \mathcal{G}]$.

The uniqueness part of the proof is elementary. For existence, I don't know a proof which is completely elementary; all the ones I know involve some deeper theorem or machinery. You don't have to worry too much about this, because the existence can be used as a black box.

Proof. For uniqueness: suppose Y, Y' satisfy items 1,2. Then for any $A \in \mathcal{G}$ we have $E[Y - Y'; A] = 0$. Since Y, Y' are both $\mathcal{G}$-measurable, $\{Y \geq Y'\} \in \mathcal{G}$. Thus $E[Y - Y'; Y \geq Y'] = 0$. This is the expectation of the nonnegative random variable $(Y - Y')1_{\{Y \geq Y'\}}$, but you proved in homework that a nonnegative random variable with zero expectation is zero almost surely. Similarly, $(Y - Y')1_{\{Y < Y'\}} = 0$ a.s. Adding, $Y - Y' = 0$ a.s.

Existence: This uses the Radon–Nikodym theorem. For $A \in \mathcal{G}$ let $\mu(A) = P(A)$ and $\nu(A) = E[X; A]$. Then μ is a probability measure and ν is a signed measure on the measurable space $(\Omega, \mathcal{G})$. The Radon–Nikodym theorem says there is a ($\mathcal{G}$-measurable!) $Y : \Omega \to \mathbb{R}$ such that $d\nu = Y\,d\mu$, i.e. $\nu(A) = \int_A Y\,d\mu$ for every $A \in \mathcal{G}$. But this says precisely that $E[X; A] = E[Y; A]$. □

Let's derive some properties of conditional expectation. They mostly parallel those of unconditional expectation, and are usually proved using the uniqueness in Theorem 14.1.

Proposition 14.3. *Conditional expectation is linear: if X, Y are integrable and $a, b \in \mathcal{R}$ then $E[aX + bY \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$ a.s.*

Proof. Set $Z = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$; we show Z satisfies the two properties in Theorem 14.1. Clearly $Z \in \mathcal{G}$ since it is a linear combination of $\mathcal{G}$-measurable random variables. Now if $A \in \mathcal{G}$, we have

$$E[Z1_A] = aE[E[X \mid \mathcal{G}]1_A] + bE[E[Y \mid \mathcal{G}]1_A] = aE[X1_A] + bE[Y1_A] = E[(aX + bY)1_A]$$

by linearity of E and definition of $E[X \mid \mathcal{G}]$. Thus by the uniqueness, we must have $Z = E[aX + bY \mid \mathcal{G}]$ almost surely. □

Proposition 14.4. *Conditional expectation is monotone: if $X \geq Y$ a.s. then $E[X \mid \mathcal{G}] \geq E[Y \mid \mathcal{G}]$ a.s.*

Proof. Replacing X by $X - Y$ and using linearity, we can assume $Y = 0$. Let $A = \{E[X \mid \mathcal{G}] \leq 0\}$, so that $E[X \mid \mathcal{G}]1_A \leq 0$. But taking expectations and noting that $A \in \mathcal{G}$, we have $E[E[X \mid \mathcal{G}]1_A] = E[X1_A] \geq 0$ since $X1_A \geq 0$. Since $E[X \mid \mathcal{G}]1_A$ is a nonpositive random variable with nonnegative expectation, we must have $E[X \mid \mathcal{G}]1_A = 0$ (you proved this in HW 1), which is to say $E[X \mid \mathcal{G}] \geq 0$ a.s. □

Proposition 14.5. *Triangle inequality: $|E[X \mid \mathcal{G}]| \leq E\,|X| \mid \mathcal{G}]$, almost surely.*

Proof. By monotonicity above, we have $E[X^+ \mid \mathcal{G}]$ and $E[X^- \mid \mathcal{G}]$ are nonnegative. Now

$$\begin{aligned} |E[X \mid \mathcal{G}]| &= \left|E[X^+ \mid \mathcal{G}] - E[X^- \mid \mathcal{G}]\right| \\ &\leq E[X^+ \mid \mathcal{G}] + E[X^- \mid \mathcal{G}] \\ &= E[X^+ + X^- \mid \mathcal{G}] = E[|X| \mid \mathcal{G}]. \end{aligned}$$

□

Proposition 14.6. *Conditional monotone convergence theorem: if $X_n \geq 0$, $X_n \uparrow X$ a.s., then $E[X_n \mid \mathcal{G}] \uparrow E[X \mid \mathcal{G}$ almost surely.*

Proof. By the monotonicity proved in the previous proposition, $E[X_n \mid \mathcal{G}$ is an increasing sequence, hence converges a.s. to some limit, Y. We have to show $Y = E[X \mid \mathcal{G}]$, for which we'll use the uniqueness of Theorem 14.1. As a limit of $\mathcal{G}$-measurable random variables, Y is also $\mathcal{G}$-measurable. If $A \in \mathcal{G}$, we have $X_n1_A \uparrow X1_A$ and $E[X_n \mid \mathcal{G}]1_A \uparrow Y1_A$. Since $E[X_n1_A] = E[E[X_n \mid \mathcal{G}]1_A]$, using (unconditional) monotone convergence on both sides gives $E[X1_A] = E[Y1_A]$. So $Y = E[X \mid \mathcal{G}]$. □

Proposition 14.7. *Conditional Fatou lemma: If $X_n \geq 0$ are integrable and so is $X := \liminf X_n$ then $E[X \mid \mathcal{G}] \leq \liminf E[X_n \mid \mathcal{G}]$, almost surely. (Presentation.)*

The integrability assumptions can be removed using your homework problem which extends conditional expectation to nonnegative random variables.

Proposition 14.8. *Suppose $X_n \to X$ in L^1. Then $E[X_n \mid \mathcal{G}] \to E[X \mid \mathcal{G}]$ in L^1.*

Proof.

$$\begin{aligned} E[|E[X_n \mid \mathcal{G}] - E[X \mid \mathcal{G}]|] &= E[|E[X_n - X \mid \mathcal{G}]|] \\ &\leq E[E[|X_n - X| \mid \mathcal{G}]] \\ &= E[|X_n - X|] \to 0. \end{aligned}$$

In the last equality we used the fact $E[E[Y \mid \mathcal{G}]] = E[Y]$; this comes from the definition of conditional expectation, taking $A = \Omega$. □

Remark 14.9. Even if we have $X_n \to X$ a.s. and in L^1, we cannot conclude that $E[X_n \mid \mathcal{G}] \to E[X \mid \mathcal{G}]$ almost surely. It seems there is a fairly strong negation of this statement due to Blackwell and Dubins. [Nate: add a reference here.]

Example 14.10. Consider the special case where we fix an event A and set $\mathcal{G} = \sigma(\{A\}) = \{\emptyset, \Omega, A, A^c\}$. Then in order for a random variable Y to be $\mathcal{G}$-measurable, it must be of the form $Y = a1_A + b1_A^c$. So to compute $Y = E[X \mid \mathcal{G}]$ for some other random variable X, we just note that $a = E[Y1_A]/P(A) = E[X1_A]/P(A) = E[X \mid A]$, the elementary definition of conditional expectation. Likewise $b = E[X \mid A^c]$. So $E[X \mid \mathcal{G}] = E[X \mid A]1_A + E[X \mid A^c]1_{A^c}$. On A, $E[X \mid \mathcal{G}]$ is constant, and its value is the average of X over A; the conditional expectation just flattens out X to make it conform to the partition $\{A, A^c\}$ of Ω.

By a similar argument, if we partition Ω into a finite or countable sequence of events $A_1, A_2, \dots$ which are pairwise disjoint and whose union is Ω, and set $\mathcal{G} = \sigma(A_1, A_2, \dots)$, then $E[X \mid \mathcal{G}] = \sum_i E[X \mid A_i]1_{A_i}$.

Proposition 14.11. *If $Z \in \mathcal{G}$, and X, ZX are both integrable, then $E[ZX \mid \mathcal{G}] = ZE[X \mid \mathcal{G}]$.*

That is, when conditioning on $\mathcal{G}$, $\mathcal{G}$-measurable random variables act like constants and can be factored out. Since $\mathcal{G}$ is the information we know, it makes sense that a quantity Z depending only on things we know should behave like a constant.

Proof. Presentation. □

Proposition 14.12. *If $X \perp\!\!\!\perp \mathcal{G}$ then $E[X \mid \mathcal{G}] = E[X]$, i.e. the conditional expectation is a constant.*

If you are given some information $\mathcal{G}$ to make a prediction of X, but the information is completely irrelevant (i.e. independent), then your best guess won't actually involve the given information (i.e. $E[X \mid \mathcal{G}]$ is a deterministic constant) and will be the same as what you would guess given no information at all (i.e. $E[X]$).

Proof. We use uniqueness: clearly $E[X]$ is $\mathcal{G}$-measurable (it's a constant!) and if $A \in \mathcal{G}$, then by independence $E[1_A E[X]] = E[1_A]E[X] = E[1_A X]$. □

Conversely:

Proposition 14.13. *If for each Borel set B, $P(X \in B \mid \mathcal{G})$ is a.s. equal to a constant, then $X \perp\!\!\!\perp \mathcal{G}$. (In particular, this holds if $E[f(X) \mid \mathcal{G}]$ is constant for all, say, bounded measurable f.)*

Proof. First note that if $P(X \in B \mid \mathcal{G})$ is a constant then it is equal to its expectation, so $P(X \in B \mid \mathcal{G}) = E[P(X \in B \mid \mathcal{G})] = P(X \in B)$.

Suppose $A \in \mathcal{G}$ and $C \in \sigma(X)$. We know $C = \{X \in B\}$ for some Borel B. Then

$$P(A \cap C) = E[1_A 1_B(X)] = E[1_A E[1_B(X) \mid \mathcal{G}]] = E[1_A P(X \in B \mid \mathcal{G})] = E[1_A P(X \in B)] = P(A)P(C)$$

so $A \perp\!\!\!\perp C$. Hence $\mathcal{G} \perp\!\!\!\perp \sigma(X)$. □

Proposition 14.14. *Suppose $\mathcal{G}_1 \subset \mathcal{G}_2$ are σ-fields. Then $E[E[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = E[E[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = E[X \mid \mathcal{G}_1]$.*

Proof. $E[E[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = E[X \mid \mathcal{G}_1]$ is obvious because $E[X \mid \mathcal{G}_1] \in \mathcal{G}_1 \subset \mathcal{G}_2$. For the other direction, clearly $E[E[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] \in \mathcal{G}_1$, and if $A \in \mathcal{G}_1$, we have

$$E[1_A E[E[X \mid \mathcal{G}_1] \mid \mathcal{G}_2]] = E[E[E[1_A X \mid \mathcal{G}_1] \mid \mathcal{G}_2]] = E[1_A X]$$

using Proposition 14.11 and the fact that $E[E[Z \mid \mathcal{G}]] = E[Z]$ twice. □

The best estimate of a best estimate of X is the best estimate of X.
Mnemonic: smallest σ-field wins.

Proposition 14.15 (Conditional Jensen)**.** *If φ is convex and $X, \varphi(X)$ are integrable then $\varphi(E[X \mid \mathcal{G}]) \leq E[\varphi(X) \mid \mathcal{G}]$ a.s.*

Proof. In the proof of unconditional Jensen, we used the fact that $\varphi(y) \geq \varphi(x) + c \cdot (y - x)$, where c is the slope of the "tangent line" to φ at x. In fact, we argued that

$$\lim_{y \uparrow x} \frac{\varphi(y) - \varphi(x)}{y - x} \leq \lim_{y \downarrow x} \frac{\varphi(y) - \varphi(x)}{y - x}$$

and that c could be taken to be any number between these two limits. We'd like to apply this same idea with $x = E[X \mid \mathcal{G}]$; but this is a random variable, i.e. depends on ω and so c will also depend on ω, and we have to be a little careful to make sure the dependence is measurable.

Set $c_n(x) = \frac{\varphi(x+1/n) - \varphi(x)}{1/n}$. Convex functions are continuous, so c_n is continuous also. As argued, for each x, $c(x) := \lim_{n \to \infty} c_n(x)$ exists and is finite, and c is measurable since it is a pointwise limit of measurable functions. Moreover, as before, for any x, y we have $\varphi(y) \geq \varphi(x) + c(x) \cdot (y - x)$. So taking $y = X$ and $x = E[X \mid \mathcal{G}]$ we can do:

$$\begin{aligned} E[\varphi(X) \mid \mathcal{G}] &\geq E[\varphi(E[X \mid \mathcal{G}]) + c(E[X \mid \mathcal{G}])(X - E[X \mid \mathcal{G}]) \mid \mathcal{G}] \\ &= \varphi(E[X \mid \mathcal{G}]) + cE[X \mid \mathcal{G}](E[X - E[X \mid \mathcal{G}] \mid \mathcal{G}]). \end{aligned}$$

□

15 Martingales

The first main class of stochastic processes we will study are **martingales**. These have a rather special property that, in some sense, they predict their own future behavior. There are a limited number of real-life systems that are reasonable to model using martingales; but there are several very useful theorems that apply to them, and it turns out to be possible to build martingales based on more general processes and use them for analysis. Also, martingales are the cornerstone of the theory of stochastic calculus that will be developed in Math 6720.

We'll first give the definition and then think about what it means.

Definition 15.1. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$ be a filtered probability space. An adapted stochastic process $\{M_n\}_{n \geq 0}$ is a **martingale** if each M_n is integrable and for each $n \geq 0$, we have

$$E[M_{n+1} \mid \mathcal{F}_n] = M_n, \quad \text{a.s.}$$

It is sometimes more convenient to write this as

$$E[M_{n+1} - M_n \mid \mathcal{F}_n] = 0.$$

If instead we have the inequality $E[M_{n+1} \mid \mathcal{F}_n] \geq M_n$ a.s., we call M_n a **submartingale**; if $E[M_{n+1} \mid \mathcal{F}_n] \leq M_n$, we use the word **supermartingale**. (Some authors refer to them collectively as **smartingales**, which I think is pretty cute.)

The way to think of a martingale is as a *fair game*: you are gambling in a casino where every game is fair, and M_n is your fortune at time n. If you look at all the information available to you at time n (including your current fortune M_n) and try to predict your fortune after the next play, your best estimate of M_{n+1} is the same as your current fortune M_n; since the games are fair, on average, you expect to neither win nor lose money on the next play.

A submartingale is a *favorable game*, where you expect, on average, to win money every time (or at worst to break even); likewise a supermartingale represents an *unfavorable game*. The names may seem backwards; they arise because there is a correspondence with subharmonic and superharmonic functions in PDEs and potential theory, but as a mnemonic you can think of the names as referring to the casino's point of view: a submartingale is favorable to the player and hence *sub-optimal* for the casino, while a supermartingale is *super* for the casino.

Notice that if we just look at expectations we see that for a martingale $E[M_{n+1}] = E[E[M_{n+1} \mid \mathcal{F}_n]] = E[M_n]$, i.e. a martingale has constant expectations. For a supermartingale, $E[M_n]$ decreases with n, and for a submartingale it increases.

Also, M_n is a martingale we have by the tower property

$$E[M_{n+2} \mid \mathcal{F}_n] = E[E[M_{n+2} \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = E[M_{n+1} \mid \mathcal{F}_n] = M_n.$$

Iterating, $E[M_{n+k} \mid \mathcal{F}_n] = M_n$. So given the information at time n, your best estimate of the martingale at *any* future time is M_n itself.

Example 15.2 (Independent increments)**.** Let $\{\xi_i\}$ be independent random variables with $E[\xi_i] = 0$ for every i (they do not have to be identically distributed) and take $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$. Let $M_0 = 0$, $M_n = \xi_1 + \cdots + \xi_n$. Then I claim M_n is a martingale; it is clearly adapted, and

$$E[M_{n+1} \mid \mathcal{F}_n] = E[M_n + \xi_{n+1} \mid \mathcal{F}_n] = E[M_n \mid \mathcal{F}_n] + E[\xi_{n+1} \mid \mathcal{F}_n]$$

by linearity. But $M_n \in \mathcal{F}_n$, so $E[M_n \mid \mathcal{F}_n] = M_n$, and ξ_{n+1} is independent of $\mathcal{F}_n$, so $E[\xi_{n+1} \mid \mathcal{F}_n] = E[\xi_{n+1}] = 0$.

If $E[\xi_i] \le 0$ we get a supermartingale, and if $E[\xi_i] \ge 0$ we get a submartingale.

Example 15.3. Independence of increments is not necessary for a martingale. It's okay if $M_1, \dots, M_n$ affect the distribution of the increment $M_{n+1} - M_n$ somehow, as long as even given all this information, it still has (conditional) expectation zero.

For example, let $\xi_1, \xi_2, \dots$ be iid fair coin flips (taking values ± 1), and $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$. Let's consider the following gambling strategy: bet 1 dollar on the first coin flip (so you win \$1 or lose \$1 with probability 1/2). If you win, keep your dollar and quit playing. If you lose, bet \$2 on the next flip. Keep this going; if you win, quit playing, and if you lose, double your bet. We can see that when you eventually win, your winnings cancel all your losses so far and leave you with a profit of \$1. Then your profit at time n can be written recursively as $M_0 = 0$ and

$$M_{n+1} = \begin{cases} 1, & M_n = 1 \\ M_n + 2^n \xi_{n+1}, & \text{otherwise.} \end{cases}$$

I claim M_n is a martingale with respect to $\mathcal{F}_n$. It is clearly adapted (use induction if you like), and we have

$$\begin{aligned} E[M_{n+1} \mid \mathcal{F}_n] &= E[1_{M_n=1} + 1_{M_n\neq 1}(M_n + 2^n\xi_{n+1}) \mid \mathcal{F}_n] \\ &= 1_{M_n=1} + 1_{M_n\neq 1}(M_n + 2^n E[\xi_{n+1} \mid \mathcal{F}_n]) \\ &= 1_{M_n=1} + 1_{M_n\neq 1}(M_n + 2^n E[\xi_{n+1}]) \\ &= 1_{M_n=1} + 1_{M_n\neq 1} M_n \\ &= M_n. \end{aligned}$$

In the second line we used linearity and Proposition 14.11 repeatedly; in the third line we used Proposition 14.12 since ξ_{n+1} is independent of $\mathcal{F}_n$.

This "doubling strategy" is a bit weird. It results in a martingale, which should be a fair game; however, as soon as any coin flip comes up heads, you win back all your losses and have a profit of \$1, and with probability 1, this will eventually happen. So this "fair" game is actually a guaranteed win.

Worse yet, we could repeat this with a biased coin that only comes up heads with probability $0 < p < 1/2$. Then this is an unfavorable game, and the argument above shows it is a supermartingale. But again it is a guaranteed win!

There are two catches though. First, you may have to wait arbitrarily long for a heads, so you'd better have unlimited time to play, because if you quit early the whole thing falls apart. Second, you could go very far into debt while waiting for the first heads, so you'd also better have unlimited credit. We'll show a little later that if either of these things is absent, you cannot use any such strategy to turn a fair or unfavorable game in your favor.

Example 15.4. The property that $E[M_n]$ is constant is necessary in order to be a martingale, but not sufficient. Let ξ be a fair coin flip, and let $M_1 = \xi$, $M_2 = \xi + \xi$. So if ξ is heads, you win a dollar at time 1, and another dollar at time 2; if ξ is tails you lose both times. Clearly $E[M_1] = E[M_2] = 0$, so this game is fair in a certain sense. But it isn't a martingale (with respect to its natural filtration, say) because $E[M_2 \mid M_1] = M_2 \neq M_1$.

You can interpret the "fairness" required to be a martingale as follows: no matter what has happened up to time n, the game from then on is still fair. Someone who has been watching the game without playing would be willing to jump in at any time, no matter what they have seen so far. This fails in this example; if the coin comes up tails at time 1, then given this information, the play at time 2 is a guaranteed loss, and an onlooker wouldn't want to enter the game at that point.

Here are a couple of simple facts:

Proposition 15.5. *If M_n, M'_n are martingales then so is $aM_n + bM'_n$. (The martingales are a vector space.)*

Proposition 15.6. *If M_n is a martingale and φ is convex, then $\varphi(M_n)$ is a submartingale.*

Proof. By the conditional Jensen inequality, $E[\varphi(M_{n+1}) \mid \mathcal{F}_n] \geq \varphi(E[M_{n+1} \mid \mathcal{F}_n]) = \varphi(M_n)$. □

Proposition 15.7. *If M_n is a submartingale and φ is convex* and nondecreasing *then $\varphi(M_n)$ is a submartingale.*

Proof. Just as above, $E[\varphi(M_{n+1}) \mid \mathcal{F}_n] \geq \varphi(E[M_{n+1} \mid \mathcal{F}_n]) \geq \varphi(M_n)$ where the second inequality is because $E[M_{n+1} \mid \mathcal{F}_n] \geq M_n$ and φ is nondecreasing. □

Here is a more elaborate, but extremely powerful, way to get new martingales from old.

Definition 15.8. A process $\{H_n\}_{n\geq 1}$ is said to be **predictable** if $H_n \in \mathcal{F}_{n-1}$ for each n. That is, from what you know at time $n-1$, you can determine exactly what H_n will be.

Definition 15.9. Suppose H_n is predictable and M_n is a martingale. Define a new process $(H \cdot M)_n$ by

$$(H \cdot M)_n = \sum_{i=1}^{n} H_i(M_i - M_{i-1}).$$

$(H \cdot M)_n$ is called the **discrete stochastic integral** or **martingale transform** of H_n with respect to M_n.

Perhaps the best way to think of this is as an investment strategy. Suppose M_n is the price of a stock at the close of day n. Our strategy will be: at the start of day n, buy H_n shares of stock (at the previous closing price M_{n-1}) and then sell it at the end of the day (at the price M_n). We can make the decision as to how much stock to buy using any information gathered up to day $n-1$ (including the closing price M_{n-1}) but of course we cannot know what day n's closing price will be.

Proposition 15.10. *If each H_n is bounded, then $(H \cdot M)_n$ is a martingale.*

So if the stock price is fair, then (on average) you can't make money trading it.

Proof. It's easy to see it is adapted, since $(H \cdot M)_n$ is defined completely in terms of $H_1, \ldots, H_n$ and $M_0, \ldots, M_n$, all of which are $\mathcal{F}_n$-measurable. We need H_n to be bounded in order to be sure that $(H \cdot M)_n$ is integrable. Then we note that

$$E[(H \cdot M)_{n+1} - (H \cdot M)_n \mid \mathcal{F}_n] = E[H_{n+1}(M_{n+1} - M_n) \mid \mathcal{F}_n] = H_{n+1}E[M_{n+1} - M_n \mid \mathcal{F}_n] = 0$$

since $H_{n+1} \in \mathcal{F}_n$. □

If M_n is a supermartingale, then $(H \cdot M)_n$ is also a supermartingale given the additional assumption that $H \geq 0$. (If the stock is tending to lose money, so will any strategy based on it, provided that the strategy is only allowed to hold positive shares of stock and isn't allowed to sell short.) The proof is just the same as above. The analogous statement for submartingales is also true.

It's worth observing that $(H \cdot M)$ is bilinear in H and M.

Here's a simple type of strategy: let τ be a stopping time, and set $H_n = 1_{\tau \leq n-1}$. This is clearly predictable. It corresponds to the strategy "buy one share, and hold it until time τ, then sell it." (On the event $\tau = \infty$ we just hold the stock forever.) Indeed, we have $(H \cdot M)_n = M_{n\wedge\tau}$, so our previous proposition shows:

Proposition 15.11. *If τ is a stopping time and M_n is a smartingale then so is $M_{n\wedge\tau}$.*

In particular, for martingales, for every n we have $E[M_{n\wedge\tau}] = E[M_0]$. As a corollary, if τ is a *bounded* stopping time, say $\tau \leq C$ a.s., then $E[M_\tau] = E[M_0]$. This says that something like our "doubling strategy" above, which makes us guaranteed free money, cannot work if we cannot wait arbitrarily long.

Theorem 15.12 (Doob decomposition)**.** *Yipu presents.*

If $\tau < \infty$ a.s., then M_τ is a random variable; what can we say about it, and in particular its expectation? We expect to have something like $E[M_\tau] = E[M_0]$ since this is what happens when τ is deterministic. If we can pass to the limit then this works:

Theorem 15.13 (Optional stopping theorem)**.** *If M_n is a submartingale, τ is a stopping time, $\tau < \infty$ almost surely, and $\{M_{n\wedge\tau}\}_{n\geq 0}$ is uniformly integrable, then $E[M_\tau] \geq E[M_0]$. For supermartingales the inequality reverses; for martingales we get equality.*

Proof. The proof is just the Vitali convergence theorem. If $\tau < \infty$ almost surely, then $M_{n\wedge\tau} \to M_\tau$ almost surely (indeed, for every ω with $\tau(\omega) < \infty$, we have $M_{n\wedge\tau}(\omega) = M_\tau(\omega)$ for all $n \geq \tau(\omega)$). So if $\{M_{n\wedge\tau}\}$ is ui, Vitali says $E[M_{n\wedge\tau}] \to E[M_\tau]$. But $E[M_{n\wedge\tau}] \geq E[M_0]$ for all n because $M_{n\wedge\tau}$ is a submartingale. □

Uniform integrability is the most general condition for this to work. But any condition that lets you pass to the limit in $E[M_{n\wedge\tau}]$ will also do, for instance, dominated convergence. The "crystal ball condition" $\sup E|X|^p < \infty$ is also useful.

We cannot do without uniform integrability. If we consider our doubling martingale, and let $\tau = \inf\{n : M_n = 1\}$, then $\tau < \infty$ almost surely, but $M_\tau = 1$, and so $1 = E[M_\tau] \neq 0 = E[M_0]$. Our doubling martingale was not ui.

This also goes to show that something like our doubling martingale cannot work if we don't have unlimited credit: if we had a finite amount of credit, the martingale would be bounded, hence ui, and every stopping strategy would have zero expectation.

Here are a couple of classic applications of optional stopping.

Example 15.14. Consider the simple random walk martingale: ξ_i are iid fair coin flips, $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$, and $M_n = \xi_1 + \cdots + \xi_n$. So we bet \$1 on successive fair coin flips and M_n records our net profit. A classic question is the **gambler's ruin problem**: suppose we decide that we will quit when we have amassed either b dollars of profit (victory) or a dollars of debt (ruin). What are the probabilities of victory and ruin?

Let $\tau = \inf\{n : M_n \in \{-a, b\}\}$ be the time at which we quit playing. Let's first argue that we will eventually quit, i.e. $\tau < \infty$ a.s.. For a crude argument, let $r = a + b$ and let $A_n = \{\xi_{nr+1} = \cdots = \xi nr + r = -1\}$ be the event that all the r coin flips from $nr + 1$ to $nr + r$ were tails. On this event, the game must be over by time $nr + r$, since if we had fewer than b dollars at time $nr + 1$, we will have less than $b - r = -a$ by time $nr + r$. Now $P(A_n) = 2^{-r} > 0$ and all the A_n are independent, so $P(\bigcup A_n) = 1$, i.e. almost surely such a run will eventually happen, and this will end the game if it hadn't ended already. (Actually such a run will happen infinitely often, by Second Borel–Cantelli.) Thus $\tau < \infty$ a.s. In particular M_τ is well defined, and we are asking for the probabilities of $\{M_\tau = -a\}$ and $\{M_\tau = b\}$.

Next, we have $-a \leq M_{n\wedge\tau} \leq b$ for all n, so in particular $\{M_{n\wedge\tau}\}$ is ui. So by the optional stopping theorem, $E[M_\tau] = E[M_0] = 0$. However, since M_τ must be either $-a$ or b, we have

$$0 = E[M_\tau] = -aP(M_\tau = -a) + bP(M_\tau = b).$$

Since $P(M_\tau = -a) + P(M_\tau = b) = 1$, we can solve this to find that $P(M_\tau = -a) = \frac{b}{a+b}$ (and $P(M_\tau = b) = \frac{a}{a+b}$). So ruin occurs with probability $\frac{b}{a+b}$. Note that this is increasing in b and decreasing in a, which makes sense.

Since we showed that the probability of hitting $-a$ before b is $\frac{b}{a+b}$, in particular the probability that we *ever* hit a is at least this large. But as $b \to \infty$ this tends to 1. This shows that, almost surely, the random walk will eventually hit any arbitrary negative value $-a$, and we can say the same for positive values. So taking a countable intersection, almost surely, simple random walk eventually hits every value. In fact it is not hard to show that simple random walk hits every value infintely many times; we say it is **recurrent**.

Lemma 15.15 (Upcrossing inequality, out of order)**.** *Let M_n be a submartingale, $a < b$, and U_k the number of upcrossings of $[a, b]$ that M_n makes by time k. Then*

$$(b - a)EU_n \leq E(M_n - a)^+ - E(M_0 - a)^+.$$

Proof. Nathan presents. □

Example 15.16. We can also use optional stopping to analyze an asymmetric random walk. Again let ξ_i be iid, but now with $P(\xi_i = 1) = p$, $P(\xi_i = -1) = 1 - p$, for some $p \neq 1/2$. Let $X_n = \xi_1 + \cdots + \xi_n$ as before. This is no longer a martingale, but we can turn it into one. The idea is to look at an **exponential martingale**: we will find a number θ such that $M_n := \theta^{X_n}$ is a martingale, and then use optional stopping on M_n.

To determine what value to use for θ, we compute

$$E[M_{n+1} - M_n \mid \mathcal{F}_n] = E[M_n(1 - \theta^{\xi_{n+1}}) \mid \mathcal{F}_n] = M_n(1 - E[\theta^{\xi_{n+1}}])$$

since $M_n \in \mathcal{F}_n$ and $\xi_{n+1} \perp\!\!\!\perp \mathcal{F}_n$. So we need to choose θ so that $E[\theta^{\xi_{n+1}}] = 1$. But we have

$$E[\theta^{\xi_{n+1}}] = p\theta + (1 - p)\theta^{-1}$$

so a little algebra shows we should take $\theta = \frac{1-p}{p}$. (The other solution is $\theta = 1$ but this results in the constant martingale $M_n = 1$ which won't be useful for anything.)

Now as before let $\tau = \inf\{n : X_n \in \{-a, b\}$. By a similar argument as before we have $\tau < \infty$ almost surely, and we have $\theta^{-a} \leq M_{n\wedge\tau} \leq \theta^b$, so again $M_{n\wedge\tau}$ is ui. Optional stopping now shows

$$1 = E[M_0] = E[M_\tau] = \theta^{-a}P(X_\tau = -a) + \theta^b P(X_\tau = b)$$

and we obtain that the probability of ruin is

$$P(X_\tau = -a) = \frac{1 - \theta^b}{\theta^{-a} - \theta^b}.$$

Lemma 15.17. *If $\tau < \infty$ a.s. and $\{M_n\}$ is a ui smartinagle then so is $\{M_{n\wedge\tau}\}$.*

Proof. (Skip this?) Let's start with the martingale case because it is simpler. Since $\{M_n\}$ is ui, we have $C := \sup_n E|M_n| < \infty$. Also, $|x|$ is a convex function so $|M_n|$ is a submartingale. Now $\tau \wedge k$ is a stopping time that is bounded by k, so by Diwakar's presentation we have $E|M_{\tau\wedge k}| \leq E|M_k| \leq C$. As $k \to \infty$, $M_{\tau\wedge k} \to M_\tau$ a.s., so by Fatou's lemma $E|M_\tau| \leq C$; in particular M_τ is integrable and so the singleton $\{M_\tau\}$ is ui.

We already know that $M_{n\wedge\tau}$ is a martingale. To show it is ui, we note that for any given ω, $M_{n\wedge\tau}(\omega)$ is either $M_n(\omega)$ or $M_\tau(\omega)$, so it is not hard to see that (rather crudely)

$$|M_{n\wedge\tau}|1_{\{|M_{n\wedge\tau}|\leq K\}} \leq |M_n|1_{\{|M_n|\geq K\}} + |M_\tau|1_{\{|M_\tau|\geq K\}}.$$

Thus

$$E[|M_{n\wedge\tau}|1_{\{|M_{n\wedge\tau}|\leq K\}}] \leq E[|M_n|1_{\{|M_n|\geq K\}}] + E[|M_\tau|1_{\{|M_\tau|\geq K\}}].$$

By uniform integrability, for any ϵ we can choose K so large that both terms on the right are less than $\epsilon/2$.

If M_n is instead a submartingale, we can use the fact that M_n^+ is a submartingale (since x^+ is an increasing convex function), whence $E[M^+_{\tau\wedge k}] \leq E[M_k^+] \leq C$. Next

$$E[M^-_{\tau\wedge k}] = E[M^+_{\tau\wedge k}] - E[M_{\tau\wedge k}] \leq E[M_k^+] - E[M_0] \leq 2C$$

where we used the fact that $M_{\tau\wedge n}$ is a submartingale and hence $E[M_{\tau\wedge k}] \geq E[M_0]$. In the end we have $E[|M_{\tau\wedge k}|] \leq 3C$ and can use Fatou as before to get $E|M_\tau| < \infty$, and proceed. □

Actually the assumption $\tau < \infty$ here is unnecessary because the martingale convergence theorem, below, implies that if $\{M_n\}$ is ui then it converges almost surely, so M_τ is well-defined even on the event $\{\tau = \infty\}$.

Theorem 15.18 (Martingale convergence theorem)**.** *Suppose M_n is a submartingale with $\sup_n EM_n^+ < \infty$. Then M_n converges almost surely to some random variable M_∞, and $E|M_\infty| < \infty$.*

Proof. Let $C = \sup_n EM_n^+ < \infty$.

Fix $a < b$ and let U_n be the number of upcrossings as in the previous lemma. Since $(X_n - a)^+ \le X_n^+$, the upcrossing lemma tells us that $EU_n \le C/(b-a)$. If U is the total number of upcrossings (for all time), clearly $U_n \uparrow U$, and so by MCT or Fatou we have $EU \le C/(b-a)$ as well; in particular $U < \infty$. Thus if $A_{a,b}$ is the event that M_n makes infinitely many upcrossings of $[a,b]$, we have $P(A_{a,b}) = 0$.

Now suppose $M_n(\omega)$ fails to converge, so that $\liminf M_n(\omega) < \limsup M_n(\omega)$. We can then choose rational a, b (depending on ω) such that $\liminf M_n(\omega) < a < b < \limsup M_n(\omega)$. Then $M_n(\omega)$ has to be less than a infinitely often, and greater than b infinitely often, so it makes infinitely many upcrossings of $[a,b]$; that is, $\omega \in A_{a,b}$. Thus we have

$$\{M_n \text{ does not converge}\} \subset \bigcup_{a,b\in\mathbb{Q}} A_{a,b}.$$

The right side is a countable union of probability-zero events, so we have that M_n converges almost surely.

Call the limit M; we have to show $E|M| < \infty$. Since $M_n^+ \to M^+$ almost surely, and $EM_n^+ \le C$, by Fatou's lemma we get $EM^+ \le C$. On the other hand, $E[M_n^-] = E[M_n^+] - E[M_n] \le C - E[M_0]$ since M_n is a submartingale and has $E[M_n] \ge E[M_0]$. Thus $E|M| = E[M^+] + E[M^-] \le 2C - E[M_0] < \infty$. In particular, M is finite almost surely. □

Corollary 15.19. *If M_n is a smartingale with $\sup_n E|M_n| < \infty$ then M_n converges almost surely. In particular this happens if $\{M_n\}$ is ui, and in this case it also converges in L^1.*

Proof. For a submartingale, we note that $M_n^+ \le |M_n|$ and use the previous theorem. For a supermartingale, consider the submartingale $-M_n$. In the ui case, we know that ui sequences are L^1-bounded □

Corollary 15.20. *If M_n is a positive supermartingale (or bounded below) it converges almost surely.*

Proof. $-M_n$ is a submartingale and $(-M_n)^+ = M_n- = 0$. □

Example 15.21. The martingale convergence theorem gives us a quick proof of the recurrence of simple random walk. Let M_n be simple random walk which is a martingale, pick an integer $a > 0$, and let $\tau = \inf\{n : M_n = a\}$. Then $M_{n\wedge\tau}$ is a martingale with $M_{n\wedge\tau} \le a$, so the martingale convergence theorem applies and $M_{n\wedge\tau}$ converges almost surely. But on the event $\{\tau = \infty\}$, $M_{n\wedge\tau} = M_n$ diverges, because M_n moves by ± 1 at every time step. Hence we must have $P(\tau = \infty) = 0$, i.e. M_n almost surely hits a. For $a < 0$, consider $-M_n$ instead.

Example 15.22. For asymmetric random walk, say with $p > 1/2$, then for any $a > 0$, $S_{n\wedge\tau_a}$ is a submartingale bounded above. As before, this means $P(\tau_a = \infty) = 0$. But we can do better: choosing $\theta = (1-p)/p < 1$ as in our example above, so that $M_n := \theta^{S_n}$ is a positive martingale, we have that M_n converges almost surely to a finite limit. This limit cannot take a nonzero value (because S_n itself cannot converge) so the only possibility is that $M_n \to 0$ a.s. That means that $S_n \to +\infty$ almost surely; the asymmetric random walk drifts to $+\infty$. Of course, we also knew this in several other ways.

Example 15.23 (Galton–Watson branching process)**.** This is another nice example of a process that we can analyze with martingale techniques. We are studying a population of some organism. In each generation, each individual in the population gives birth to a random number of offspring, and then dies. We are interested in the long-term behavior of the population: in particular, will it become extinct?

Let X_n be the number of individuals in the population at generation n; we'll take $X_0 = 1$. Let $\xi_{n,i}$ be the number of offspring produced by the ith individual in generation n; we'll assume that $\{\xi_{n,i} : n, i \geq 1\}$ are iid nonnegative integer-valued random variables. (Note that for any given n, only finitely many of the $\xi_{n,i}$ will be used, but the number depends on the previous generation, so we don't know in advance how many we'll need. So we do assume we have infinitely many $\xi_{n,i}$ available in our model, but in any given outcome most of them will go unused.) Then we can define

$$X_n = \sum_{i=1}^{X_{n-1}} \xi_{n,i}.$$

This sum may look funny because of the random limit, but on a pointwise level it makes sense. Let E be the event of extinction, i.e. $E = \bigcup_n \{X_n = 0\}$. We would like to know $P(E)$. Certainly it will depend on the distribution of the $\xi_{n,i}$ (the "offspring distribution"); specifically on their mean. So let $\mu = E[\xi_{n,i}]$ (assume the expectation exists and is finite). The case $\mu = 0$ results in immediate extinction (since there are no offspring at all) so assume $\mu > 0$. Set $\mathcal{F}_n = \sigma(\xi_{k,i} : k \leq n)$.

Proposition 15.24. *$M_n := X_n/\mu^n$ is a martingale.*

Proof. Clearly M_n is adapted. Next, we compute $E[X_n \mid \mathcal{F}_{n-1}]$. Intuitively, given the information $\mathcal{F}_{n-1}$ we know the number of individuals X_{n-1} at time $n-1$, but we have no information about how many offspring each one will have (since the $\xi_{n,i}$ are independent of $\mathcal{F}_{n-1}$) so the best guess we can make is $E[\xi_{n,i}] = \mu$. Thus our best guess at X_n should be μX_{n-1}. To make this precise, we can do the following:

$$\begin{aligned}
E[X_n \mid \mathcal{F}_{n-1}] &= E\left[\sum_{i=1}^{\infty} \xi_{n,i} 1_{\{X_{n-1} \geq i\}}\right] \\
&= \sum_{i=1}^{\infty} E[\xi_{n,i} 1_{\{X_{n-1} \geq i\}} \mid \mathcal{F}_{n-1}] && \text{(cMCT)} \\
&= \sum_{i=1}^{\infty} E[\xi_{n,i}] 1_{\{X_{n-1} \geq i\}} && \text{(since } X_{n-1} \in \mathcal{F}_{n-1} \text{ and } \xi_{n,i} \perp\!\!\!\perp \mathcal{F}_{n-1}) \\
&= \mu X_{n-1}.
\end{aligned}$$

Thus, $E[M_n \mid \mathcal{F}_{n-1}] = \mu^{-n} E[X_n \mid \mathcal{F}_{n-1}] = \mu^{-(n-1)} X_{n-1} = M_{n-1}$. □

In particular, being a nonnegative martingale, M_n converges almost surely to some M_∞.

Proposition 15.25. *If $\mu < 1$ then $P(E) = 1$, i.e. extinction is almost certain.*

This makes sense because when $\mu < 1$, the average individual is not producing enough offspring to replace itself.

Proof. The fact that M_n is a martingale means $E[M_n] = E[M_0] = 1$ for all n, i.e. $E[X_n] = \mu^n \to 0$. But by Markov, $P(X_n \geq 1) \leq E[X_n] \to 0$. Since $\{X_n = 0\} \subset E$ for every n, we have $P(E) \geq 1 - \mu^n$ for every n, so $P(E) = 1$. □

If $\mu = 1$ the situation is more subtle because the reproduction rate is critical. Of course one possibility is that $\xi_{n,i} = 1$ almost surely, i.e. every individual always has exactly one child, and in this case we have $X_n = 1$ for all n, and the population survives. Sadly, this is the only case where extinction is avoided.

If $\mu = 1$ but $\xi_{n,i}$ is not identically 1, then we must have $p_0 := P(\xi_{n,i} = 0) > 0$. Now in this case $\{X_n\}$ is a nonnegative martingale and hence converges almost surely. Since it is integer valued, it must be eventually constant. So consider $A_{n,k} = \{X_n = X_{n+1} = \cdots = k\}$. In order for $A_{n,k}$ to happen, we must avoid the events $E_{m,k} = \{\xi_{m,1} = \cdots = \xi_{m,k} = 0\}$, which is the event that in generation m, the first k individuals die childless. However, the events $E_{m,k}, m \geq n$ are independent, and $P(E_{m,k}) = p_0^k > 0$, so by (a trivial version of) Borel–Cantelli we have $P(\bigcup_{m \geq n} E_{m,k}) = 1$, and therefore $P(A_{n,k}) = 0$. Taking a countable union, $P(\bigcup_{n \geq 0, k \geq 1} A_{n,k}) = 0$. This says that X_n cannot be eventually constant at any positive value. Since by almost sure convergence X_n must be eventually constant at some value, it must be zero, which is extinction. (Note that we have $X_n \to 0$ but $E[X_n] = 1$, so the convergence is certainly not in L^1 and this is yet another example of a non-ui martingale.)

It can be shown that if $\mu > 1$ then there is a positive probability that extinction is avoided. See Durrett.

15.1 L^1 convergence

Here is a very simple type of martingale, which is actually very general as we shall see: Let X be an integrable random variable, and set $M_n = E[X \mid \mathcal{F}_n]$. The idea is that M_n is the best approximation of X that we can get given the information available at time n. This is clearly a martingale thanks to the tower property.

Theorem 15.26. *If X is integrable, then the set of random variables $\{E[X \mid \mathcal{G}] : \mathcal{G} \subset \mathcal{F}$ is a σ-field$\}$ is ui.*

Proof. Evan presents. □

Corollary 15.27. *$M_n := E[X \mid \mathcal{F}_n]$ is a ui martingale. In particular, it converges almost surely and in L^1 to some M_∞.*

Theorem 15.28. *$M_\infty = E[X \mid \mathcal{F}_\infty]$.*

Proof. Lemuel presents. □

So in the limit, these successive approximations approach the best approximation you could hope for.

Corollary 15.29 (Lévy zero-one law)**.** *If $A \in \mathcal{F}_\infty$ then $P(A \mid \mathcal{F}_n) \to 1_A$ a.s. and in L^1. (In particular, the limit of $P(A \mid \mathcal{F}_n)(\omega)$ is almost surely either 0 or 1, depending on whether $\omega \in A$.)*

Interestingly, we can use this to prove the Kolmogorov zero-one law. Let $\mathcal{F}_n = \sigma(\mathcal{G}_1, \ldots, \mathcal{G}_n)$ where the $\mathcal{G}_i$ are independent σ-fields, and we let $\mathcal{T} = \bigcap_n \sigma(\mathcal{G}_n, \mathcal{G}_{n+1}, \ldots)$ be the tail σ-field. If $A \in \mathcal{T} \subset \mathcal{F}_\infty$ then $A \perp\!\!\!\perp \mathcal{F}_n$ for each n, and so $P(A \mid \mathcal{F}_n) = P(A)$. Lévy says the left side converges almost surely to 1_A so we must have $1_A = P(A)$ a.s. This is only possible if $P(A)$ is 0 or 1.

Actually it turns out every ui martingale is of the form $M_n = E[X \mid \mathcal{F}_n]$.

Theorem 15.30. *Suppose M_n is a martingale, and $M_n \to M_\infty$ in L^1. Then $M_n = E[M_\infty \mid \mathcal{F}_n]$ almost surely.*

Proof. We use uniqueness to verify that the conditional expectation $E[M_\infty \mid \mathcal{F}_n]$ is in fact M_n. Clearly $E[M_\infty \mid \mathcal{F}_n] \in \mathcal{F}_n$. Now let $A \in \mathcal{F}_n$. For any $k > n$, since $M_n = E[M_k \mid \mathcal{F}_n]$, we have $E[M_n 1_A] = E[M_k 1_A]$. Now let $k \to \infty$; since $M_k \to M_\infty$ in L^1 we have

$$|E[M_k 1_A] - E[M_\infty 1_A]| \leq E[|M_k - M_\infty| 1_A] \leq E|M_k - M_\infty| \to 0.$$

So $E[M_k 1_A] \to E[M_\infty 1_A]$. Passing to the limit we have $E[M_n 1_A] = E[M_\infty 1_A]$. □

Corollary 15.31. *If M_n is a martingale, the following are equivalent:*

1. *$\{M_n\}$ is uniformly integrable;*
2. *M_n converges almost surely and in L^1;*
3. *M_n converges in L^1;*
4. *There exists an integrable random variable X with $M_n = E[X \mid \mathcal{F}_n]$.*

16 Markov chains

The tricky part about any stochastic process is the dependence between the random variables $X_1, X_2, \ldots$. In most models you want some relationship between X_n and X_{n+1}; they shouldn't be independent, but they also shouldn't be deterministically related, otherwise there is no randomness. If the dependence structure is too complicated, you get a model that you can't analyze.

Random walks are a nice simple example: if $X_n = \xi_1 + \cdots + \xi_n$, for iid ξ_i, then X_{n+1} is X_n with a litlle bit of extra randomness added (literally). But requiring that relationship to be additive is quite restrictive; indeed, we often want to look at processes in state spaces that have no notion of addition or group structure. The Markov chain is a model where the process evolves by making incremental random changes to its state, and it's a good compromise of being reasonable to analyze while still being quite a flexible model.

Fix a measurable space $(S, \mathcal{S})$ to be our state space. For most of our development, we will take S to be a finite or countable set, with $\mathcal{S} = 2^S$. This makes all the measure theory very simple: all subsets of S are measurable, and any measure on S can be represented by a probability mass function: $\mu(A) = \sum_{x \in A} p(x)$. Occasionally we might allow for more generality, but countable sets are what you should keep in mind.

The simple picture of a Markov chain is as a weighted directed graph. The vertex set is the state space S, and each directed edge (x, y) is labeled with a probability $p(x, y)$ in such a way that for each x, $\sum_y p(x, y) = 1$. (Any directed edge that is not present in the graph can be thought of as having weight 0.) The process starts at some vertex x_0 and evolves: if at some time it is at vertex x, its next move is to a randomly chosen neighbor of x, so that the probability of moving to y is $p(x, y)$. We have to say something about how these random moves are made; in particular their dependence on one another. Intuitively we think the move from x should be independent of all previous moves, but that is not quite right, since the previous moves determined the vertex x where we are now sitting, which in turn determines the possible vertices for the next move (and their respective probabilities). So this intuitive idea takes a little more work to describe formally.

We can describe it this way:

Definition 16.1. (Preliminary) Fix a filtered probability space. An adapted stochastic process $\{X_n\}_{n \geq 0}$ with state space $(S, \mathcal{S})$ is a (discrete-time, time-homogeneous) **Markov chain** if, for each n and each measurable $B \subset \mathcal{S}$, the conditional probability $P(X_{n+1} \in B \mid \mathcal{F}_n)$ only depends on B and X_n. So if we know X_n, the state of the process at time n, then we cannot improve our estimate of $P(X_{n+1} \in B)$ using further information from $\mathcal{F}_n$ (which includes older history about the process).

In other words, there should exist a function $p : S \times \mathcal{S} \to [0, 1]$ so that $P(X_{n+1} \in B \mid \mathcal{F}_n) = p(X_n, B)$. This function should have the properties that:

1. For each $B \in \mathcal{S}$, $x \mapsto p(x, B)$ is a measurable function on S.
2. For each $x \in S$, $B \mapsto p(x, B)$ is a probability measure on S.

p is called the **transition function** or **transition probability** of the chain $\{X_n\}$. In words, if the process is in state x at some time, then $p(x, B)$ is the probability that it will be in B at the next step.

When S is countable, any probability measure can be defined by a mass function. So we could instead think of $p : S \times S \to [0, 1]$ being jointly measurable and satisfying $\sum_{y \in S} p(x, y) = 1$ for every x, and let $p(x, B) = \sum_{y \in B} p(x, y)$. Then if the process is at state x, $p(x, y)$ is the probability that its next step will be to state y.

Corollary 16.2. $E[f(X_{n+1} \mid \mathcal{F}_n)] = \int_S f(y)p(X_n, dy)$.

Proof. Standard mantra. □

If p is allowed to depend on n as well, then $\{X_n\}$ is a **time-inhomogeneous Markov chain**: the probability of transitioning from x to y may change over time. These models are considerably more annoying to deal with so we will not.

If S is finite and we identify it with $\{1, \dots, N\}$, we can think of p as an $N \times N$ matrix whose ij'th entry is $p(i, j)$, the probability of transitioning from i to j.

There is another technicality to deal with. Normally a stochastic process in S would just be a sequence of S-valued random variables $X_0, X_1, \dots$. However, we would like to be able to consider arbitrary starting points $x_0 \in S$. One option would be to have a whole family of processes, indexed by S, like $\{X_0^x, X_1^x, \dots\}_{x \in S}$ where $X_0^x = x$. However a convention that turns out to work better, though it looks weird at the outset, is to use a single process but vary the probability measure P. Thus our starting point is a measurable space $(\Omega, \mathcal{F})$ (usually with a filtration $\{\mathcal{F}_n\}$), a sequence $X_0, X_1, \dots$ of measurable maps $X_n : \Omega \to S$, and a *family* of probability measures $\{P_x\}_{x \in S}$, having the property that $P_x(X_0 = x) = 1$. You can read P_x as "probability when started at x"; so $P_x(X_1 = y)$ is the probability that the process, when started at x, visits y at the next step. We will use E_x to denote expectation (or conditional expectation) with respect to the measure P_x.

If I write P or E without a subscript (which I will try and avoid), it normally means that the statement holds for every P_x. Thus $P(X_n \to x) = 1$ means "for every $x \in S$, $P_x(X_n \to z) = 1$". That is, no matter where the process starts, it converges to z almost surely. Likewise, I will try to qualify "almost sure" statements by writing "P_x-a.s." but if I just say "a.s." it means "P_x-a.s. for every $x \in S$".

An alternative approach is to use a single probability measure P with the property that $P(X_0 = x) > 0$ for every $x \in S$, and think of $P_x(A)$ as the conditional probability $P(A \mid X_0 = x)$. This works okay for countable S, but it breaks down when S is uncountable, since $P(X_0 = x)$ is necessarily zero for all but countably many $x \in S$, so you find yourself conditioning on events of probability zero.

So here's our full definition. It is different from Durrett's but we will see that they are equivalent.

Definition 16.3. Suppose:

1. $(S, \mathcal{S})$ is a measurable space;
2. $(\Omega, \mathcal{F})$ is a measurable space;
3. $\{\mathcal{F}_n\}_{n \geq 0}$ is a filtration on $(\Omega, \mathcal{F})$;
4. $\{P_x\}_{x \in S}$ is a family of probability measures on $(\Omega, \mathcal{F})$, such that for each $A \in \mathcal{F}$, the function $x \mapsto P_x(A)$ is measurable;
5. $\{X_n\}_{n \geq 0}$ is a sequence of measurable functions from Ω to S which is adapted ($X_n \in \mathcal{F}_n$);
6. For each $x \in S$, $P_x(X_0 = x) = 1$.

The process X_n is a (discrete-time, time-homogeneous) **Markov chain** with transition function $p : S \times \mathcal{S} \to [0,1]$ if for each n, each $x \in S$, and each measurable $B \subset S$, we have

$$P_x(X_{n+1} \in B \mid \mathcal{F}_n) = p(X_n, B), \quad P_x\text{-a.s.}$$

In this case, we can identify the transition function p as $p(x, B) = P_x(X_1 \in B)$.

Lemma 16.4. *If $f_0, \dots, f_k : S \to \mathbb{R}$ are bounded measurable functions, then*

$$E_x(f_0(X_n)f_1(X_{n+1}) \dots f_k(X_{n+k}) \mid \mathcal{F}_n) = \int \cdots \int f_k(y_k)p(y_{k-1}, dy_k)f_{k-1}(y_{k-1})p(y_{k-2}, dy_{k_1}) \dots f_1(y_1)p(X_n, dy_1)f_0(X_n).$$

Proof. The base case is trivial since it just reads $E_x[f_0(X_n) \mid \mathcal{F}_n] = f_0(X_n)$, which holds because $X_n \in \mathcal{F}_n$. Now suppose it holds for $k \geq 0$. We have by the tower property

$$\begin{aligned} E_x[f_0(X_n) \dots f_k(X_{n+k})f_{k+1}(X_{n+k+1}) \mid \mathcal{F}_n] &= E_x[f_0(X_n) \dots f_k(X_{n+k})E[f_{k+1}(X_{n+k+1}) \mid \mathcal{F}_{n+k}] \mid \mathcal{F}_n] \\ &= E_x[f_0(X_n) \dots f_k(X_{n+k}) \int f_{k+1}(y_{k+1})p(X_{n_k}, dy_{k+1}) \mid \mathcal{F}_n] \\ &= \int \cdots \int f_{k+1}(y_{k+1})p(y_k, dy_{k+1})f_k(y_k)p(y_{k-1}, dy_k) \dots f_1(y_1)p(X_n, dy_1)f_0(X_n) \end{aligned}$$

applying the induction hypothesis, replacing $f_k(y_k)$ by $f_k(y_k) \int f_{k+1}(y_{k+1})p(y_k, dy_{k+1})$ which is a bounded measurable function of y_k. □

Using this, we can prove a key property of a Markov process: that it "starts fresh" at every step. It doesn't remember its history, so if at time n you observe that the process is in some state x, from then on it behaves exactly like a new process which was started at x.

It is a bit challenging to state this precisely, in its full generality. Durrett does it by assuming that the underlying probability space is the infinite product space $S^{\mathbb{N}}$ and defining shift operators on this space. To me this feels weird, so I'm taking the following approach which makes the use of the sequence space more explicit.

Let $S^{\mathbb{N}}$ be the product of countably many copies of S, equipped with the infinite product σ-algebra $\mathcal{S}^{\mathbb{N}}$ generated by cylinder sets $B_0 \times \cdots \times B_n \times S \times \dots$. The elements of $S^{\mathbb{N}}$ are sequences $z = (z(0), z(1), \dots)$ of elements of S. (Here we are abusing notation to think of $\mathbb{N}$ as starting with 0.)

For each $x \in S$, we let μ_x be the probability measure on $S^{\mathbb{N}}$ which is the law of the Markov chain $\{X_n\}$ started at x; i.e. $\mu_x(B) = P_x((X_0, X_1, \dots) \in B)$. Observe that μ_x puts all its mass on those sequences whose 0th term is x. Note that for each measurable $B \subset S^{\mathbb{N}}$, the map $x \mapsto \mu_x(B)$ is measurable.

Theorem 16.5 (Markov property)**.** *For any measurable $B \subset S^{\mathbb{N}}$, any $x \in S$, and any $n \geq 0$, we have*

$$P_x((X_n, X_{n+1}, \dots) \in B \mid \mathcal{F}_n) = \mu_{X_n}(B), \quad P_x\text{-a.s.} \tag{13}$$

So if you have a question to ask about the behavior of the process after time n, and you know its location X_n at time n, then the answer is the same as for a brand new process whose starting point is X_n. More precisely, conditionally on $\mathcal{F}_n$, the law of $\{X_n, X_{n+1}, \dots\}$ is the same as that of $\{X_0, X_1, \dots\}$, under P_{X_n}.

Proof. Fix $x \in S$ and $n \geq 0$. Let $\mathcal{L}$ be the collection of all $B \subset S^{\mathbb{N}}$ such that (13). It's easy to verify that $\mathcal{L}$ is a λ-system. Let $\mathcal{P}$ be the set of all $B = B_0 \times \cdots \times B_k \times S \times \dots$ with $B_0, \dots, B_k \subset S$ measurable. This is a

π-system which generates $\mathcal{S}^{\mathbb{N}}$. By the previous lemma, setting $f_i = 1_{B_i}$, we have

$$P_x((X_n, X_{n+1}, \cdots \in B) \mid \mathcal{F}_n) = E_x[1_{B_0}(X_n) \dots 1_{B_k}(X_{n+k}) \mid \mathcal{F}_n] \tag{14}$$

$$= \int \cdots \int 1_{B_k}(y_k) p(y_{k-1}, dy_k) 1_{B_{k-1}}(y_{k-1}) p(y_{k-2}, dy_{k_1}) \dots 1_{B_1}(y_1) p(X_n, dy_1) 1_{B_0}(X_n). \tag{15}$$

On the other hand, if we use the previous lemma again with $n = 0$ and replacing x by an arbitrary $y \in S$, we have

$$P_y((X_0, X_1, \dots) \in B \mid \mathcal{F}_0) = \int \cdots \int 1_{B_k}(y_k) p(y_{k-1}, dy_k) 1_{B_{k-1}}(y_{k-1}) p(y_{k-2}, dy_{k_1}) \dots 1_{B_1}(y_1) p(X_0, dy_1) 1_{B_0}(X_0).$$

Taking E_y of both sides, we get

$$\mu_y(B) = P_y((X_0, X_1, \dots) \in B) \tag{16}$$

$$= \int \cdots \int 1_{B_k}(y_k) p(y_{k-1}, dy_k) 1_{B_{k-1}}(y_{k-1}) p(y_{k-2}, dy_{k_1}) \dots 1_{B_1}(y_1) p(y, dy_1) 1_{B_0}(y) \tag{17}$$

Substituting X_n for y, we see this is the right side of (15), and we have established (13). □

Corollary 16.6. *The law μ_x of $\{X_n\}$ is completely determined by the transition function p.*

Proof. In (17) we showed that $\mu_x(B)$ can be written in terms of x and p for every B in a π-system that generates $\mathcal{S}^{\mathbb{N}}$. □

Corollary 16.7 (Multistep transition probabilities)**.** *Define p^k recursively by $p^{k+1}(x, B) = \int_S p(y, B)\, p^k(x, dy)$. (In the countable case, we can say $p^{k+1}(x, z) = \sum_{y \in S} p(x, y) p^k(y, z)$; in the finite case, note this is just matrix multiplication.) Then $P(X_{n+k} \in B \mid \mathcal{F}_n) = p^k(X_n, B)$, and $E[f(X_{n+k} \mid \mathcal{F}_n] = \int f(y) p^k(X_n, dy)$. In particular (with $n = 0$ and taking expectations), $P_x(X_k \in B) = p^k(x, B)$.*

Corollary 16.8. *For any measurable $f : S^{\mathbb{N}} \to \mathbb{R}$, we have*

$$E_x[f(X_n, X_{n+1}, \dots) \mid \mathcal{F}_n] = \int_{S^{\mathbb{N}}} f d\mu_{X_n}$$

provided that either side exists.

Proof. Standard mantra. □

We've just showed (in fancy language) that the map from Markov chains (up to distribution) to transition functions is one-to-one. The next statement says it is also onto. So all you have to do is write down a transition function and you know that the corresponding Markov chain exists.

Theorem 16.9. *Given any transition function p on a standard Borel space $(S, \mathcal{S})$, there is a Markov chain $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \{P_x\}, \{X_n\})$ whose transition function is p.*

Proof. Let $\Omega = S^{\mathbb{N}}$ (here again $\mathbb{N} = \{0, 1, 2, \dots\}$) with its product σ-field, and let $X_n : S^{\mathbb{N}} \to S$ be the coordinate maps. Set $\mathcal{F}_n = \sigma(X_0, X_1, \dots, X_n)$. It remains to produce the measures $\{P_x\}$ and verify that they make $\{X_n\}$ into a Markov chain with transition function p.

Fix $x \in S$. For measurable $A \subset S^{n+1}$, define $\mu_n(A) = \int\cdots\int 1_A(x, x_1, \dots, x_n)p(x_{n-1}, dx_n) \dots p(x, dx_1)$ (where $\mu_0 = \delta_x$). This is a consistent family of measures (verify). Let P_x be the measure on $S^{\mathbb{N}}$ produced by the Kolmogorov extension theorem. We have $P_x(X_0 = x) = \mu_0(\{x\}) = 1$.

Now we have to check that $P_x(X_{n+1} \in B \mid \mathcal{F}_n) = p(X_n, B)$; we use the uniqueness of conditional expectation. $p(X_n, B)$ is clearly $\mathcal{F}_n$-measurable. Let $A \in \mathcal{F}_n$; since $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ we have $A = \{(X_0, \dots, X_n) \in C\}$ for some measurable $C \subset S^{n+1}$. By definition of P_x we have

$$\begin{aligned}
E_x[1_B(X_{n+1})1_A] &= E_x[1_{C\times B}(X_0, \dots, X_{n+1})] \\
&= \mu_{n+1}(C \times B) \\
&= \int\cdots\int 1_C(x, x_1, \dots, x_n)1_B(x_{n+1})p(x_n, dx_{n+1} \dots p(x, dx_1) \\
&= \int\cdots\int 1_C(x, x_1, \dots, x_n)p(x_n, B) \dots p(x, dx_1) \\
&= \int 1_C(x, \dots, x_n)p(x_n, B)\, d\mu_n \\
&= E_x[1_C(X_0, \dots, X_n)p(X_n, B)] \\
&= E_x[1_A p(X_n, B)].
\end{aligned}$$

Repeating this for every $x \in S$ we have a Markov chain with the desired properties.

To check that $x \mapsto P_x(A)$ is measurable, note that it holds for all cylinder sets A, and use π-λ. □

The following **strong Markov property** says that the Markov property also holds at stopping times. After time τ, the process behaves like a fresh process whose starting point was X_τ, and any other history information from before time τ was irrelevant.

Theorem 16.10 (Strong Markov property)**.** *Let τ be a stopping time. Then for every $x \in S$ and every $B \in \mathcal{S}^{\mathbb{N}}$ we have*

$$P_x(\tau < \infty, (X_\tau, X_{\tau+1}, \dots) \in B \mid \mathcal{F}_\tau) = 1_{\tau<\infty}\mu_{X_\tau}(B), \quad P_x\text{-a.s.} \tag{18}$$

The $\{\tau < \infty\}$ event is inserted to ensure that we are only writing X_τ for outcomes where it makes sense. Usually we apply it to stopping times with $\tau < \infty$ almost surely, in which case this is redundant.

Proof. It suffices to show that for any n, we have

$$P_x(\tau = n, (X_\tau, X_{\tau+1}, \dots) \in B \mid \mathcal{F}_\tau) = 1_{\{\tau=n\}}\mu_{X_\tau}(B)$$

since then we simply have to sum over all finite n. The right side is $\mathcal{F}_\tau$ measurable (since τ and X_τ both are). Let $A \in \mathcal{F}_\tau$, so that $A \cap \{\tau = n\} \in \mathcal{F}_n$. Then

$$\begin{aligned}
E_x[1_A 1_{\{\tau=n\}} 1_B(X_\tau, X_{\tau+1}, \dots)] &= E_x[1_A 1_{\{\tau=n\}} 1_B(X_n, X_{n+1}, \dots)] \\
&= E_x[E_x[1_A 1_{\{\tau=n\}} 1_B(X_n, X_{n+1}, \dots) \mid \mathcal{F}_n]] \\
&= E_x[1_A 1_{\{\tau=n\}} E_x[1_B(X_n, X_{n+1}, \dots) \mid \mathcal{F}_n]] \\
&= E_x[1_A 1_{\{\tau=n\}} \mu_{X_n}(B)] \\
&= E_x[1_A 1_{\{\tau=n\}} \mu_{X_\tau}(B)].
\end{aligned}$$

□

This proof was pretty easy; we just had to consider all the possible values for τ, of which there are countably many because we are working in discrete time. In continuous time this gets harder, and in fact there can be processes that satisfy the Markov property but not the strong Markov property. Strong Markov is really the most useful one, and hence in continuous time one typically only studies processes that satisfy strong Markov.

17 Markov chain examples

Example 17.1 (Simple random walk)**.** Take $S = \mathbb{Z}$ and $p(x, x+1) = p(x, x-1) = 1/2$, $p(x, y) = 0$ otherwise. (We will verify later that $\xi_n := X_n - X_{n-1}$ are iid, or possibly in homework.)

Example 17.2 (General random walk)**.** Take $S = \mathbb{R}$, μ any probability measure on $\mathbb{R}$, and $p(x, B) = \mu(B-x)$.

Example 17.3 (Branching process)**.** If μ is a probability measure on $\{0, 1, 2, \dots\}$ then the branching process with offspring distribution μ is a Markov chain with transition function $p(n, B) = \mu^{*n}(B)$. Less fancifully, $p(n, m) = P(\xi_1 + \cdots + \xi_n = m)$ where $\xi_i \sim \mu$ are iid.

Example 17.4 (Random walk on a graph)**.** $G = (V, E)$ is a countable, locally finite graph; $p(x, y) = 1/d(x)$. (So the next step from x is to a uniformly chosen neighbor of x.)

18 Transience and recurrence

Here's a nice application of the strong Markov property.

Theorem 18.1. *Let $x \in S$, $T_0 = 0$, and $T_{k+1} = \inf\{n > T_k : X_n = x\}$, so that T_k is the k'th time that the process reaches the state x. Then $P_x(T_k < \infty) = P_x(T_1 < \infty)^k$.*

Proof. Let B be the set of all $z \in S^{\mathbb{N}}$ such that $z(n) = x$ for some $n > 0$ (you can verify that this is measurable). Then $T_{k+1} < \infty$ iff $T_k < \infty$ and $(X_{T_k}, X_{T_1+1}, \dots) \in B$. By the strong Markov property, we have

$$\begin{aligned} P_x(T_{k+1} < \infty) &= E_x[P_x(T_k < \infty, (X_{T_k}, X_{T_k+1}, \dots) \in B \mid \mathcal{F}_{T_k})] \\ &= E_x[1_{\{T_k<\infty\}} \mu_{X_{T_k}}(B)] \\ &= \mu_x(B) P_x(T_k < \infty) \end{aligned}$$

since $X_{T_k} = x$. But $\mu_x(B) = P_x((X_0, X_1, \dots) \in B) = P_x(T_1 < \infty)$. So by induction we are done. □

In particular, if $P_x(T_1 < \infty) = 1$, i.e. if we start at x we are guaranteed to return at least once, then $T_k < \infty$ for all k (P_x-a.s.), so we are guaranteed to visit infinitely many times. In this case we say x is a **recurrent** state. On the other hand, if $P_x(T_1 < \infty) < 1$, i.e. there is a positive probability we will never return, then we have

$$P_x\left(\bigcap_k \{T_k < \infty\}\right) \le P_x(T_k < \infty) = P_x(T_1 < \infty)^k \to 0$$

so we are guaranteed *not* to visit infinitely many times. In this case we say x is **transient**. Note this is a sort of zero-one law: the probability of infinitely many returns to x is either 0 or 1.

Recurrent states are the ones we should study if we want to learn about long-term behavior of a Markov chain. In classifying states as transient or recurrent, some connectivity properties are important.

Notation 18.2. For $y \in S$, let $\tau_y = \inf\{n \geq 1 : X_n = y\}$. For $x, y \in S$, let $\rho_{xy} = P_x(\tau_y < \infty)$ be the probability that, when starting at x, we eventually reach y. (Note that ρ_{xx} is the probability of a return to x; the visit at time 0 doesn't count.) If $\rho_{xy} > 0$ we write $x \to y$; this means it is possible to get from x to y.

We show that $\to$ is the basically the connectivity relationship in the transition graph of the chain, i.e. it's the transitive closure of the graph.

Lemma 18.3. *If $p(x, y) > 0$ then $x \to y$.*

Proof. Obvious, since $p(x, y) = P_x(X_1 = y) = P_x(\tau_y = 1) \leq P_x(\tau_y < \infty)$. □

Lemma 18.4. *$\rho_{xz} \geq \rho_{xy}\rho_{yz}$. In particular, if $x \to y$ and $y \to z$ then $x \to z$.*

Proof. Homework? □

Lemma 18.5. *If S is countable and $x \to y$, there is a path $x = x_0, x_1, \ldots, x_m = y$ with $p(x_i, x_{i+1}) > 0$.*

Proof. If $\rho_{xy} = P_x(\tau_y < \infty)$ then there exists some m with $P_x(\tau_y \leq m) > 0$, i.e. it is possible to get from x to y in m steps. There are countably many sequences $x = x_0, \ldots, x_m = y$ so it must be that for one of them, we have $P_x(X_0 = x_0, \ldots, X_m = x_m) > 0$. Unwinding the notation in Lemma 16.4 shows that $P_x(X_0 = x_0, \ldots, X_m = x_m) = p(x_0, x_1) \ldots p(x_{m-1}, x_m)$, so all these factors must be positive. □

This can fail if S is uncountable. Consider a chain on state space $\{a\} \cup [0, 1] \cup \{b\}$ with $p(a, \cdot)$ uniform on $[0, 1]$, $p(x, \{b\}) = 1$ for $x \in [0, 1]$ or $x = b$. We have $p(a, x) = 0$ for every x, but $\rho(a, b) = 1$.

Recurrence is contagious.

Theorem 18.6. *If x is recurrent and $x \to y$ then $\rho_{xy} = \rho_{yx} = 1$ and y is recurrent.*

Proof. Let A be the event that $\tau_y < \infty$ and the chain visits x at some time after τ_y. On the one hand, since the chain must visit x infinitely many times if it starts there, we must have $P_x(A) = P_x(\tau_y < \infty) = \rho_{xy}$. On the other hand, we can use the strong Markov property: if $B \subset S^{\mathbb{N}}$ is the set of sequences that contain at least one x, then $A = \{\tau_y < \infty\} \cap \{(X_{\tau_y}, X_{\tau_y+1}, \ldots) \in B\}$. So strong Markov says

$$\begin{aligned} P_x(A) &= E_x[P_x(\tau_y < \infty, (X_{\tau_y}, X_{\tau_y+1}, \ldots) \in B \mid \mathcal{F}_{\tau_y})] \\ &= E_x[1_{\tau_y<\infty}\mu_{X_{\tau_y}}(B)] \\ &= E_x[1_{\tau_y<\infty}\mu_y(B)] \\ &= \mu_y(B)E_x[1_{\tau_y<\infty}] \\ &= P_y(\tau_x < \infty)P_x(\tau_y < \infty) = \rho_{yx}\rho_{xy}. \end{aligned}$$

Thus we have shown $\rho_{xy} = \rho_{yx}\rho_{xy}$. Since $\rho_{xy} > 0$ we must have $\rho_{yx} = 1$.

Let $T_k = \inf\{n > T_{k-1} : X_n = x\}$ be the time of the kth return to x, as above (with $T_0 = 0$); by recurrence we have $T_k < \infty$ for all k, P_x-a.s. Let $A_k = \{\exists n \in (T_k, T_{k+1}) : X_n = y\}$ be the event that the chain visits y between the kth and $k + 1$th visits to x. Then if $B \subset S^{\mathbb{N}}$ is the set of all sequences that start with $s(0) = x$, and then contain a y before the next x, we have $A_k = \{(X_{T_k}, X_{T_{k+1}}, \ldots) \in B\}$. By strong Markov,

$$P_x(A_k \mid \mathcal{F}_{T_k}) = \mu_{X_{T_k}}(B) = \mu_x(B) = P_x(\tau_y < \tau_x).$$

The right side is deterministic, so we have shown that A_k is independent of $\mathcal{F}_{T_k}$ (under P_x). Since clearly $A_1, \ldots, A_{k-1} \in \mathcal{F}_{T_k}$, we have shown that the events A_i are independent. Moreover, as they all have probability $P_x(\tau_y < \tau_x) \geq P_x(\tau_y < \infty) = \rho_{xy} > 0$, Borel-Cantelli tells us that $P_x(A_k \text{ i.o.}) = 1$, i.e. if we start at x, we almost surely make infinitely many visits to y. In particular, we almost surely make at least one, so $\rho_{xy} = 1$.

Finally, we have $\rho_{yy} \geq \rho_{yx}\rho xy = 1$ so y is recurrent. □

Definition 18.7. A chain is **irreducible** if for every x, y we have $x \to y$. That is, from any state we can get to any other state.

For a countable chain, this happens iff the transition digraph is connected.

Corollary 18.8. *If $\{X_n\}$ is irreducible then either every state is recurrent, or every state is transient.*

For finite irreducible chains only one of these is possible.

Proposition 18.9. *If S is finite then there exists a recurrent state.*

Proof. (Sketch) Let N_y be the total number of visits to y. Suppose to the contrary every state is transient, so for each y we have $P_y(N_y = \infty) = 0$. Fix $x \in S$, and use the strong Markov property to show that for every $y \in S$, we have $P_x(N_y = \infty) = 0$. Use a pigeonhole argument to get a contradiction. □

For an example of an (infinite-state) irreducible chain with every state transient, consider asymmetric simple random walk. Another classic example is simple random walk on $\mathbb{Z}^d$ with $d \geq 3$, though we have not proved this yet.

19 Stationary distributions

We already have a mechanism for starting a chain at any given state x thanks to our family of probability measures P_x. We could also start the chain at a state chosen randomly according to any given distribution.

Definition 19.1. Let ν be a probability measure on S, and define a probability measure P_ν on Ω by

$$P_\nu(A) = \int_S P_x(A)\nu(dx).$$

Proposition 19.2.

1. *P_ν is indeed a probability measure;*
2. *Under P_ν, $X_0 \sim \nu$;*
3. *Under P_ν, X_n is still a Markov chain with transition function p, and the strong Markov property still holds.*
4. *If S is countable, then $P_\nu(X_n = y) = \sum_x p^n(x, y)\nu(x)$. If we think of p as a matrix, we should think of ν as a row vector; then the distribution of X_n is νp.*

Definition 19.3. A probability measure π on S is a **stationary** measure for X_n if, under P_π, we have $X_n \sim \pi$ for every n. (In the countable case, this means that π is a left eigenvector of p.)

Proposition 19.4. *It suffices to verify the above with $n = 1$; if $X_1 \sim \pi$ under P_π then π is stationary.*

Proof. We show, by induction on n, that $X_n \sim \pi$ for every n. By assumption this holds for $n = 1$. Let us write this another way: we have

$$\pi(B) = P_\pi(X_1 \in B) = E_\pi[P_\pi(X_1 \in B \mid \mathcal{F}_0)] = E_\pi[p(X_0, B)] = \int p(x, B)\pi(dx).$$

Now if it holds for n we have

$$\begin{aligned} P_\pi(X_{n+1} \in B) &= E_\pi[P_\pi(X_{n+1} \in B \mid \mathcal{F}_n)] \\ &= E_\pi[p(X_n, B)] \\ &= \int p(x, B)\pi(dx) = \pi(B). \end{aligned}$$

□

Example 19.5. Let X_n be random walk on a finite graph (G, E). Then $\pi(x) = \frac{1}{2|E|}d(x)$ is a stationary probability measure.

Proof. It is clearly a probability measure since $\sum_x d(x) = 2|E|$ (every edge is counted twice, one for each end of it). Then we have

$$\sum_x \pi(x)p(x, y) = \sum_{x\sim y} \frac{1}{2|E|}d(x) \times \frac{1}{d(x)} = \frac{1}{2|E|}d(y)$$

since there are $d(y)$ neighbors of x. □

Example 19.6. If X_n is random walk on a finite regular graph, then the uniform measure on G is stationary.

Example 19.7. Simple random walk has no stationary distribution. If it did, we would have $\pi(x) = \frac{1}{2}(\pi(x-1)+\pi(x+1))$. Rearranging, $\pi(x+1)-\pi(x) = \pi(x)-\pi(x-1)$. By induction, $\pi(x+1)-\pi(x) = \pi(1)-\pi(0) := C$ for every x, i.e. $\pi(x) = \pi(0) + Cx$. If $C \neq 0$ this cannot be a positive measure, since $\pi(x)$ is negative for appropriate x. If $C = 0$ then it is a uniform measure on $\mathbb{Z}$ which cannot be a probability measure.

Example 19.8. Asymmetric reflecting random walk. You'll compute a stationary distribution in your homework.

Proposition 19.9. *If X_n has a stationary distribution π, and $\pi(x) > 0$, then x is recurrent. In particular if X_n has a stationary distribution then it has at least one recurrent state.*

Proof. On the one hand, we have

$$P_\pi(N_x = \infty) = P_\pi(\limsup_n \{X_n = x\}) \geq \limsup_n P_\pi(X_n = x) = \pi(x) > 0$$

as you showed in an early homework. On the other hand, let B be the set of all sequences containing infinitely many x. Then by the strong Markov property

$$\begin{aligned} P_\pi(N_x = \infty) &= P_\pi(\tau_x < \infty, (X_{\tau_x}, X_{\tau_x+2}, \dots) \in B) \\ &= P_\pi(\tau_x < \infty)\mu_x(B) \\ &= P_\pi(\tau_x < \infty)P_x(N_x = \infty). \end{aligned}$$

So we must have $P_x(N_x = \infty) > 0$, which means x is recurrent. □

Note symmetric simple random walk is a Markov chain where every state is recurrent but no stationary distribution exists. So recurrence is necessary but not sufficient. In fact the right condition is:

Definition 19.10. A state x in a Markov chain is **positive recurrent** if $E_x\tau_x < \infty$. That is, not only do you return to x, but the average time to do so is finite. If x is recurrent but not positive recurrent, we say it is **null recurrent**

For simple random walk, every state is null recurrent, since we showed using martingales that for any x, y, $E_x\tau_y = \infty$.

We collect some related results, but won't give the proofs here. Consult Durrett.

Proposition 19.11. *If x is positive recurrent and $x \to y$ then y is positive recurrent.*

Corollary 19.12. *If X_n is irreducible and has one positive recurrent state, then every state is positive recurrent.*

Proposition 19.13. *If X_n has a positive recurrent state, then it has at least one stationary distribution.*

Proposition 19.14. *If π is a stationary distribution and $\pi(x) > 0$ then x is positive recurrent.*

So positive recurrence is a necessary and sufficient condition for the existence of a stationary distribution. We could explore this further but won't; most of the time, if a stationary distribution exists, it is obvious or easy to compute directly what it is.

We are heading for a result that says that (under appropriate conditions) a Markov chain converges weakly to its stationary distribution π. That is, if you start the chain in some arbitrary state, let it run for a long time, and then look at where it is, the state you see looks a lot like a sample drawn from the stationary distribution π.

Let's figure out what some of these "appropriate conditions" are, by looking at some examples where this fails to hold.

One obvious issue is that if there is more than one stationary distribution, which one we converge to might depend on the starting point.

Example 19.15. Consider a chain on two states a, b, with $p(a, a) = p(b, b) = 1$ (i.e. we never move). Then δ_a and δ_b are both stationary. (In fact, every probability measure on $\{a, b\}$ is stationary; the transition matrix is the identity matrix.) Under P_a, we have $X_n \sim \delta_a$ for every n; under P_b we have $X_n \sim \delta_b$.

Evidently, the problem is that we have several subsets of S that don't communicate with one another. In some sense we should break the state space up and consider it as two separate chains.

As we shall see, irreducibility is a sufficient condition to rule this out, as we shall see. In an irreducible chain, if a stationary distribution exists, it is unique.

Another obstruction to convergence is "periodicity".

Example 19.16. Again take a chain with two states a, b, with $p(a, b) = p(b, a) = 1$ (so we flip-flop back and forth). It is not hard to see that the unique stationary distribution is the uniform distribution $\pi(a) = \pi(b) = 1/2$. However, X_n does not converge weakly to π if started at a; its distribution alternates between δ_a and δ_b.

This kind of behavior rarely happens except in contrived examples. (But note that it also happens in random walk on $\mathbb{Z}$; if we start at 0, say, then we can only visit odd-numbered states at odd-numbered times, and even-numbered states at even-numbered times. However, random walk on $\mathbb{Z}$ doesn't have a stationary distribution anyway.) We could study it further but instead we'll just discuss how to rule it out.

Definition 19.17. A state $x \in S$ is **aperiodic** if there exists a number r_x such that for all $n \geq r_x$, we have $p^n(x, x) > 0$. (So after waiting long enough, there are no obstructions to returning to x at any given time.) The chain X_n is **aperiodic** if every state is aperiodic.

Notice that the above example fails this property: we have $p^n(a,a) = 0$ for every odd n.

Here are some useful properties for verifying aperiodicity. I don't know if I'll go into the proofs.

Proposition 19.18. *If $p(x,x) > 0$ then x is aperiodic.*

Proof. Obvious. You could sit at x for n steps with probability $p(x,x)^n$, so $p^n(x,x) \geq p(x,x)^n > 0$. □

Proof. If there exist numbers $n_1, \ldots, n_m$ whose greatest common divisor is 1 and with $p^{n_i}(x,x) > 0$ for all i, then x is aperiodic. □

Proof. Note that we have $p^n(x,x) > 0$ for all n of the form $n = a_1 n_1 + \cdots + a_m n_m$ with $a_i \geq 0$, since $p^n(x,x) \geq (p^{n_1}(x,x))^{a_1} \ldots (p^{n_m}(x,x))^{a_m} > 0$, i.e. we could return to x in n_1 steps and repeat this a_1 times, etc. Now combine this with the elementary number theory fact that if $n_1, \ldots, n_m$ are relatively prime then any sufficiently large integer can be written in this form. (The Euclidean algorithm says any integer can be written in this form if negative coefficients are allowed.) □

Proposition 19.19. *If x is aperiodic and $y \to x \to y$ then y is aperiodic. In particular, in an irreducible chain, if any state is aperiodic then the chain is aperiodic.*

Proof. Since $y \to x \to y$ we have $p^s(y,x) > 0$ and $p^t(x,y) > 0$ for some s, t. We also have $p^n(x,x) > 0$ for all $n \geq r_x$. So for all $n \geq s + t + r_x$ we have $p^n(y,y) \geq p^s(y,x) p^{n-s-t}(x,x) p^t(x,y) > 0$. □

Example 19.20. Random walk on an odd cycle is aperiodic. Random walk on an even cycle is not. Reflecting random walk is aperiodic (since $p(0,0) > 0$).

So here is the main convergence theorem for countable Markov chains.

Theorem 19.21. *Let X_n be a Markov chain on a countable state space S. Suppose that X_n is irreducible, aperiodic, and has a stationary distribution π. Then for every $x \in S$, we have $P_x(X_n = z) \to \pi(z)$ for all z (in fact, uniformly in z). In particular, under any P_x, we have $X_n \to \pi$ weakly (in fact, in total variation).*

Proof. The idea of the proof is a technique known as "coupling". We will run two independent copies of the chain, one (call it X_n) started at a fixed state x, and the other (Y_n) started at the stationary distribution π. Then we will run them until they meet; the hypotheses will guarantee that this happens. After this time, since they have both been transitioning by the same rules from the same point (the place where they met), they must have the same distribution.

Note first that the theorem can be expressed solely in terms of the transition function p: the conclusion is that $p^n(x,z) \to \pi(z)$. So the specific random variables involved are unimportant; we can run the proof using any Markov chain with this transition function.

Let (X_n, Y_n) be two independent copies of the original chain; so (X_n, Y_n) is a Markov chain on state space $S \times S$ whose transition function is $\tilde{p}((x,y),(x',y')) = p(x,x')p(y,y')$. We can check that X_n alone is a Markov chain whose transition function is the original p; we have for any (x,y)

$$\begin{aligned} P_{(x,y)}(X_{n+1} = x' \mid \mathcal{F}_n) &= \sum_{y' \in S} P_{(x,y)}((X_{n+1}, Y_{n+1}) = (x', y')) \\ &= \sum_{y' \in S} \tilde{p}((X_n, Y_n), (x', y')) \\ &= \sum_{y' \in S} p(X_n, x') p(Y_n, y') \\ &= p(X_n, x') \sum_{y' \in S} p(Y_n, y') \end{aligned}$$

but the sum equals 1. The same argument applies to Y_n.

Next, note that $\pi \times \pi$ is a stationary distribution for (X_n, Y_n); we have

$$\sum_{x,y \in S} \pi(x)\pi(y)\tilde{p}((x,y),(x',y') = \left(\sum_{x \in S} \pi(x)p(x,x')\right)\left(\sum_{y \in S} \pi(y)p(y,y')\right) = \pi(x')\pi(y').$$

In particular (X_n, Y_n) has a recurrent state.

Now I claim that (X_n, Y_n) is irreducible. Fix x, y, x', y'; we will show that $\tilde{p}^m((x,y),(x',y')) > 0$ for some m. By irreducibility, there exist s, t such that $p^s(x,x') > 0$ and $p^t(y,y') > 0$. By aperiodicity, there exists an N so large that $p^n(x,x) > 0$ and $p^n(y,y) > 0$ for all $n \geq N$. Now take $m = N + s + t$; we have

$$\tilde{p}^m((x,y),(x',y')) = p^m(x,x')p^m(y,y') \geq (p^{N+t}(x,x)p^s(x,x'))(p^{N+s}(y,y)p^t(y,y')) > 0.$$

In particular, since (X_n, Y_n) has a recurrent state, every state (x, y) is recurrent.

Let $\tau = \inf\{n : X_n = Y_n\}$ be the first time at which X_n, Y_n meet (in other words τ is the hitting time of the diagonal). By irreducibility and recurrence, from any starting point, (X_n, Y_n) visits every state infinitely often. In particular it visits every state on the diagonal of $S \times S$, so we have $\tau < \infty$ almost surely.

Now fix an $x \in S$; we will start the chain (X_n, Y_n) with initial distribution $\delta_x \times \pi$ (so $X_0 = x$ and $Y_0 \sim \pi$). (From now on in this proof, P means $P_{\delta_x \times \pi}$ to save typing.) Fix some $k \leq n$ and look at the event that meeting occured at time k and $X_n = z$; we have

$$\begin{aligned}
P(X_n = z, \tau = k) &= E[P(X_n = z, \tau = k \mid \mathcal{F}_k)] \\
&= E[1_{\{\tau=k\}} P(X_n = z \mid \mathcal{F}_k)] && \text{since } \{\tau = k\} \in \mathcal{F}_k \\
&= E[1_{\{\tau=k\}} p^{n-k}(X_k, z)] \\
&= E[1_{\{\tau=k\}} p^{n-k}(Y_k, z)] \\
&= P(Y_n = z, E[1_{\{\tau=k\}} p^{n-k}(Y_k, z)] \\
&= P(Y_n = z, \tau = k).
\end{aligned}$$

Summing over $k \leq n$, we have

$$P(X_n = z, \tau \leq n) = P(Y_n = z, \tau \leq n). \tag{19}$$

Rearranging,

$$P(X_n = z) - P(X_n = z, \tau > n) = P(Y_n = z) - P(Y_n = z, \tau > n). \tag{20}$$

But $P(Y_n = z) = \pi(z)$ for all n, since Y_n starts in the stationary distribution π and hence keeps that distribution. Rearranging some more,

$$|P(X_n = z) - \pi(z)| = |P(X_n = z, \tau > n) - P(Y_n = z, \tau > n)| \leq P(X_n = z, \tau > n) + P(Y_n = z, \tau > n) \leq 2P(\tau > n). \tag{21}$$

Since $\tau < \infty$ almost surely, $P(\tau > n) \to 0$. □

Remark 19.22. Note that if we can estimate $P(\tau > n)$ for a particular chain then we will have a statement about the rate of convergence.

Remark 19.23. There was nothing magical about choosing X_n, Y_n to be *independent* copies of the chain; it was just a construction that made it convenient to verify that the meeting time τ was finite. We could consider any other coupling, i.e. any adapted process (X_n, Y_n) on $S \times S$ such that X_n and Y_n are individually Markov chains on S with transition function p; we just have to be able to show that they meet almost surely. We don't even need (X_n, Y_n) to be a Markov chain. It may be that some other coupling could give us a better bound on $P(\tau > n)$ and thus a better bound on the rate of convergence.

Corollary 19.24. *If X_n is an irreducible Markov chain on a countable state space then it has at most one stationary distribution.*

Proof. Suppose it has two; call them π, π'. Suppose first that X_n is aperiodic. Fix any $x \in S$; by the above theorem, we have $P_x(X_n = y) \to \pi(y)$ for all y. But by the same logic, we also have $P_x(X_n = y) \to \pi'(y)$. Hence $\pi = \pi'$.

To dispose of the periodicity assumption, let p be the transition function of X_n, and define a new chain $\tilde{X}_n$ whose transition function is defined by

$$\tilde{p}(x, y) = \begin{cases} \frac{1}{2} + \frac{1}{2}p(x, x), & x = y \\ \frac{1}{2}p(x, y), & x \neq y. \end{cases}$$

So $\tilde{X}_n$ flips a coin at each step; if it is heads it doesn't move at that step, and if it is tails it moves according to p. We've produced a "lazy" version of the chain. $\tilde{X}_n$ is still irreducible, is clearly aperiodic since $\tilde{p}(x, x) > 0$ for every x, and it is easy to check that π, π' are stationary distributions for $\tilde{X}_n$. So by the previous case, $\pi = \pi'$. □

Remark 19.25. This proof itself was also lazy because we took advantage of a big theorem, and used a sneaky trick to avoid the aperiodicity hypothesis. For a more honest proof, see Durrett's Theorem 6.5.7.

Remark 19.26. The convergence theorem is the basis for the so-called Markov Chain Monte Carlo (MCMC) method of sampling from a probability distribution. Suppose we have a probability measure μ on S and we want to generate a random variable ξ with distribution μ. One approach is to come up with a Markov chain X_n whose stationary distribution is μ, start it at some arbitrarily chosen state x_0, and run it for some large number of steps N, then take $\xi = X_N$. The convergence theorem says that the distribution of ξ is approximately μ. And if we have information about the rate of convergence, we can work out how large we need to take N (the so-called "mixing time") to get the distribution of ξ within any given ϵ of μ.

For example, when you shuffle a deck of cards, you are running a Markov chain on the finite group S_{52} (whose transition function depends on your specific shuffling technique). For a very simple example, suppose a "shuffle" consists of swapping the top card with a randomly chosen card. This is a random walk on S_{52} using the generators $\{id, (1\,2), (1\,3), \ldots, (1\,52)\}$, or in other words a random walk on the Cayley graph, in which every vertex has degree 52. Thus the stationary distribution is uniform. So if you repeat this "shuffle" enough times, you approximately choose a uniform permutation of the deck; every permutation has approximately equal probability. This also works with a more elaborate shuffling procedure, such as a riffle or "dovetail" shuffle, which mixes faster; Bayer and Diaconis famously showed that most of the mixing happens within 7 shuffles, i.e. X_7 is already quite close to uniform.

It may seem dumb to use a Markov chain algorithm to sample from the uniform distribution on a finite set such as S_{52}. After all, there are a zillion algorithms to generate (pseudo)-random integers of any desired size, so it's easy to sample uniformly from $\{1, 2, \ldots, 52!\}$. But the problem is that it is not so easy to find an explicit and easily computable bijection from $\{1, 2, \ldots, 52!\}$ to S_{52} (given an integer, how do you quickly find the permutation to which it corresponds?).